Legal Knowledge and Information Systems

JURIX 2019: ***The Thirty-second Annual Conference***

Editors:
Michał Araszkiewicz
Víctor Rodríguez-Doncel

JURIX 2019

Legal Knowledge and Information Systems

JURIX 2019:
The Thirty-second Annual Conference

Editors:
Michał Araszkiewicz
Víctor Rodríguez-Doncel

In recent years, the application of machine learning tools to legally relevant tasks has become much more prevalent, and the growing influence of AI in the legal sphere has prompted the profession to take more of an interest in the explainability, trustworthiness, and responsibility of intelligent systems.

This book presents the proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX 2019), held in Madrid, Spain, from 11 to 13 December 2019. Traditionally focused on legal knowledge representation and engineering, computational models of legal reasoning, and analyses of legal data, more recently the conference has also encompassed the use of machine learning tools. A total of 81 submissions were received for the conference, of which 14 were selected as full papers and 17 as short papers. A further 3 submissions were accepted as demo presentations, resulting in a total acceptance rate of 41.98%, with a competitive 25.5% acceptance rate for full papers. The 34 papers presented here cover a broad range of topics, from computational models of legal argumentation, case-based reasoning, legal ontologies, and evidential reasoning, through classification of different types of text in legal documents and comparing similarities, to the relevance of judicial decisions to issues of governmental transparency.

The book will be of interest to all those whose work involves the use of knowledge and information systems in the legal sphere.

JURIX 2019

ISBN 978-1-64368-048-4

ISBN 978-1-64368-048-4 (print)
ISBN 978-1-64368-049-1 (online)
ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

LEGAL KNOWLEDGE AND INFORMATION SYSTEMS

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Volume 322

Recently published in this series

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Legal Knowledge and Information Systems

JURIX 2019: The Thirty-second Annual Conference

Edited by

Michał Araszkiewicz

Jagiellonian University, Kraków, Poland

and

Víctor Rodríguez-Doncel

Universidad Politécnica de Madrid, Spain

IOS Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-64368-048-4 (print)
ISBN 978-1-64368-049-1 (online)
Library of Congress Control Number: 2019956261
doi: 10.3233/FAIA322

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

PRINTED IN THE NETHERLANDS

Preface

We are delighted to announce the proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX 2019). The JURIX annual conference, organized under the auspices of the Dutch Foundation for Legal Knowledge-Based Systems (http://www.jurix.nl), has been established as an internationally renowned forum for the exchange of ideas concerning theoretical models and practical applications developed in the broadly construed sphere of artificial intelligence (AI) and law research. Traditionally, this field has been concerned with legal knowledge representation and engineering, computational models of legal reasoning, and analyses of legal data. However, recent years have witnessed the application of machine learning tools to legally relevant tasks rising to prominence.

The constantly growing influence of AI on different spheres of social life has prompted the community's emerging interest in the explainability, trustworthiness, and responsibility of intelligent systems—and not in vain, as a high-level expert group the European Commission convened this year published the *Ethics Guidelines for Trustworthy AI*. It declared that the very first attribute of trustworthy AI was "lawfulness." The research presented at JURIX conferences is an excellent example of interdisciplinary research integrating the methods and approaches from different branches of jurisprudence and computer science.

The 2019 edition of JURIX, which runs from 11 to 13 December, is hosted by the Ontological Engineering Group at the Artificial Intelligence Department of the Technical University of Madrid (Universidad Politécnica de Madrid). For this edition, we have received 81 papers, from which 14 were selected as full papers (10 pages in the proceedings) and 17 as 6-page short papers. Moreover, three submissions have been accepted as demo presentations. These figures result in a total acceptance rate of 41.98% and a competitive 25.5% acceptance rate for full papers. The accepted papers cover a broad array of topics, from computational models of legal argumentation, case-based reasoning, legal ontologies, and evidential reasoning, through classification of different types of text in legal documents and comparing similarities and the relevance of judicial decisions, to issues of governmental transparency.

Two invited speakers have honored JURIX 2019 by kindly agreeing to deliver two keynote lectures: Danièle Bourcier and Francesca Toni. Daniéle Bourcier has been responsible for pioneering research in the field of law, computers, and linguistics—currently, she is a director of research emeritus at Centre Nationale de la Recherche Scientifique (CNRS) and leads the Law and Governance Technologies Department at the Centre for Administrative Science Research (CERSA) at the University of Paris II. She is actively involved in the AI and law community, currently serving as a member of the Executive Committee of the International Association of Artificial Intelligence and Law. Francesca Toni is one of the most significant representatives of the computational argumentation research community. She is Professor of Computational Logic in the Department of Computing at Imperial College London, a member of the AI research theme, and the leader of the Computational Logic and Argumentation research group (CLArg). Francesca Toni has contributed extensively to different topics in logic,

agents-based systems, and argumentation, recently focusing her attention *inter alia* on the application of argumentation models to generate explanations.

Traditionally, the main JURIX conference is accompanied by co-located events comprising workshops and tutorials. This year's edition welcomes seven workshops: the CEILI Workshop on Legal Data Analysis; GDPR Compliance—Theories, Techniques, Tools; IberLegal: NLP for Legal Domain in Languages of the Iberian Peninsula (Spanish, Catalan, Galician, Basque, and Portuguese); LegRegSW JURIX 2019 – A Legislation and Regulation Semantic Web; MIREL 2019 – Mining and Reasoning with Legal Texts; TeReCom – The 3rd Workshop on Technologies for Regulatory Compliance; XAILA 2019 – The EXplainable AI in Law Workshop; and Defeasible Logic for Normative Reasoning (a tutorial). The continuation of well-established events and the organization of entirely new ones provide a great added value to the JURIX conference, enhancing its thematic and methodological diversity and attracting members of the broader community. Since 2013, JURIX has also offered researchers entering the field as Ph.D. students the opportunity to present their work during the Doctoral Consortium session, and this edition is no exception. Finally, for the first time, this edition of JURIX offers the Industry Session—a special event enabling business representatives to present their products to the academy to foster further discussions concerning state-of-the-art developments in legal tech.

Organizing this edition of the conference would not have been possible without the support of many people and institutions. Special thanks are due to the local organizing team chaired by Víctor Rodríguez-Doncel and Elena Montiel Ponsoda (https://jurix2019.oeg-upm.net), and the enthusiasm of UPM's Vice Chancellor for Research and the outstanding AI researcher, Asunción Gómez-Pérez. We would like to thank the workshops' and tutorials' organizers for their excellent proposals and for the effort involved in organizing the events. We owe our gratitude to Monica Palmirani, who kindly assumed the function of the Doctoral Consortium Chair. We are particularly grateful to the 91 members of the Program Committee for their excellent work in the rigorous review process and for their participation in the discussions concerning borderline papers. Finally, we would like to thank the former and current JURIX executive committee and steering committee members not only for their support and advice but also generally for taking care of all the JURIX initiatives.

Michał Araszkiewicz, JURIX 2019 Program Chair
Víctor Rodríguez-Doncel, JURIX 2019 Organization Chair

Conference Organisation

Program Commitee Chair
Michał Araszkiewicz, Jagiellonian University in Kraków

Local Chairs
Víctor Rodríguez-Doncel, Universidad Politécnica de Madrid
Elena Montiel Ponsoda, Universidad Politécnica de Madrid

Doctoral Consortium Chair
Monica Palmirani, University of Bologna

Local Organisation Committee
Ana Ibarrola de Andrés
José Ángel Ramos Gargantilla
Patricia Martín Chozas
María Navas Loro
Virginia de Pablo Llorente
Juan Utande
Sanju Tiwari

Program Committee
Thomas Ågotnes, University of Bergen
Tommaso Agnoloni, ITTIG-CNR
Francisco Andrade, University of Minho
Grigoris Antoniou, University of Huddersfield
Michał Araszkiewicz, Jagiellonian University
Kevin Ashley, University of Pittsburgh
Katie Atkinson, University of Liverpool
Matteo Baldoni, Università di Torino
Cesare Bartolini, University of Luxembourg
Trevor Bench-Capon, University of Liverpool
Floris Bex, Utrecht University
Alexander Boer, University of Amsterdam
Michael Bommarito, Bommarito Consulting, LLC
Danièle Bourcier, University of Paris 2/CNRS
Karl Branting, The MITRE Corporation
Elena Cabrio, Université Côte d'Azur, CNRS, Inria
Pompeu Casanovas, Universitat Autònoma de Barcelona
Marcello Ceci, University College Cork
Federico Cerutti, Cardiff University
Giuseppe Contissa, University of Bologna
Federico Costantini, Università degli Studi di Udine
Matteo Cristani, University of Verona
Claudia d'Amato, University of Bari

Giovanni Sileno, University of Amsterdam
Barry Smith, SUNY Buffalo
Clara Smith, UNLP and UCALP
Sarah Sutherland, Canadian Legal Information Institute
Leon van der Torre, University of Luxembourg
Tom Van Engers, University of Amsterdam
Marc van Opijnen, KOOP
Anton Vedder, KU Leuven
Bart Verheij, University of Groningen
Serena Villata, CNRS
Fabio Vitali, University of Bologna
Vern R. Walker, Maurice A. Deane School of Law at Hofstra University
Doug Walton, University of Windsor
Radboud Winkels, University of Amsterdam
Adam Wyner, Swansea University
Hajime Yoshino, Meiji Gakuin University
John Zeleznikow, Victoria University
Haozhen Zhao, Ankura
Tomasz Zurek, Maria Curie-Sklodowska University

JURIX steering commitee
Katie Atkinson, University of Liverpool
Pompeu Casanovas, Universitat Autonomá de Barcelona
Erich Schweighofer, University of Vienna
Serena Villata, CNRS

JURIX executive committee
Tom van Engers, University of Amsterdam, president
Bart Verheij, University of Groningen, vice-president/secretary
Floris Bex, Utrecht University/Tilburg University, treasurer

This page intentionally left blank

Contents

Short Papers

This page intentionally left blank

Full Papers

This page intentionally left blank

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190301

Identification of Rhetorical Roles of Sentences in Indian Legal Judgments

Paheli BHATTACHARYA [a,1,2], Shounak PAUL [a,1], Kripabandhu GHOSH [b], Saptarshi GHOSH [a] and Adam WYNER [c]

[a] *Indian Institute of Technology Kharagpur, India*
[b] *Tata Research Development and Design Centre (TRDDC) Pune, India*
[c] *Swansea University, United Kingdom*

Abstract. Automatically understanding the rhetorical roles of sentences in a legal case judgement is an important problem to solve, since it can help in several downstream tasks like summarization of legal judgments, legal search, and so on. The task is challenging since legal case documents are usually not well-structured, and these rhetorical roles may be subjective (as evident from variation of opinions between legal experts). In this paper, we address this task for judgments from the Supreme Court of India. We label sentences in 50 documents using multiple human annotators, and perform an extensive analysis of the human-assigned labels. We also attempt automatic identification of the rhetorical roles of sentences. While prior approaches towards this task used Conditional Random Fields over manually handcrafted features, we explore the use of deep neural models which do not require hand-crafting of features. Experiments show that neural models perform much better in this task than baseline methods which use handcrafted features.

Keywords. Semantic Segmentation, Rhetorical Roles, Legal Case Documents, Deep Learning, BiLSTM

1. Introduction

Rhetorical role labelling of sentences in a legal document refers to understanding what semantic function a sentence is associated with, such as facts of the case, arguments of the parties, the final judgement of the court, and so on. Identifying the rhetorical roles of sentences in a legal case document can help in a variety of downstream tasks like semantic search [1], summarization [2,3], case law analysis [4], and so on. However, legal case documents are usually not well structured [5,6], and various themes often interleave with each other. For instance, the reason behind the judgment (Ratio of the decision) often interleaves with Precedents and Statutes. Hence it sometimes becomes difficult even for human experts to understand the intricate differences between the rhetorical roles. Hence, *automating* the identification of these rhetorical roles is a challenging task.

For supervised machine learning of the roles, it is important to develop a high quality gold standard corpus, capturing the rhetorical roles of sentences as accurately as possible. Different approaches for the task have constructed their own set of annotated doc-

[1] Equal contribution by the first and second authors.
[2] Corresponding Author: Paheli Bhattacharya; Email: paheli.cse.iitkgp@gmail.com

uments [1,2,4], but do not report an extensive analysis on the annotation process. Apart from Inter-Annotator Agreement (IAA) scores, it is useful to understand issues such as the amount of subjectivity associated to the labels. In this paper, we perform a systematic annotation study and an extensive inter-annotator study. We show that even legal experts find it difficult to distinguish some specific pairs of labels, thus showing that some subjectivity is inherent in these labels.

Prior attempts to automate the identification of rhetorical roles of sentences in legal documents [2–4] rely on hand-crafted features (see Section 2 for details) such as linguistic cue phrases indicative of a particular rhetorical role [2, 3, 7], the sequential arrangement of labels [2], and so on. Some of these features, e.g., indicator cue phrases, are *largely dependent on legal-expert knowledge* which is expensive to obtain. Also, the hand-crafted features developed in the prior works are often specific to one or a few domains/categories (e.g., Cyber crime and Trade secrets in [4]). It has not been explored whether one can devise a set of features that works for documents across domains.

Recently developed deep learning, neural network models do not require hand-engineering features, but are able to automatically learn the features, given sufficient amounts of training data. Additionally, such models perform better in tasks like classification than methods using hand-crafted features.

In this paper, we explore two neural network models to automatically identify the rhetorical roles of sentences in legal documents – (i) a Hierarchical BiLSTM model, and (ii) a Hierarchical BiLSTM-CRF model. Similar models have been used in the medical domain [8], but to our knowledge, this work is the first to use them in the legal domain. We use these models for supervised classification across *seven rhetorical labels* (classes) and over documents from *five different legal domains*. The Hierarchical BiLSTM-CRF model achieves a very good performance (Macro F-score in the range $[0.8 - 0.9]$), outperforming baseline methods that use hand-crafted features. We also analyse the rhetorical roles predicted by our model, and find that the subjectivity between certain pairs of labels (e.g., Ratio vs. Precedent) that is present among the human annotators is also reflected in the predictions by the algorithm.

This is the first paper on identifying rhetorical roles of sentences in legal documents that brings together (i) an extensive annotation study, and (ii) deep learning models for automating the task. [3]

2. Related Work

In this section we discuss prior work about annotation, automatic rhetorical labelling, and applications of deep learning in the legal domain.

Automatic labelling of the rhetorical role of sentences relies heavily on manual annotation. While papers that aim to automate the task of semantic labelling also perform an annotation analysis [4,5], other works focus on the process of annotation – developing a manual/set of rules for annotation, inter-annotator studies, curation of a gold standard corpus, and so on. TEMIS, a corpus of 504 sentences, that were annotated both syntactically and semantically, was developed in [9]. An in-depth annotation study and curation of a gold standard corpus for the task of sentence labelling can be found in [10], where

[3] The dataset and implementations of the proposed neural model are available at *https://github.com/Law-AI/semantic-segmentation*.

assessor agreement was low for labels like Facts and Reasoning Outcomes. Towards automating the annotation task, [11] discusses an initial methodology using NLP tools on 47 criminal cases drawn from the California Supreme Court and State Court of Appeals.

There have been several prior attempts towards automatically identifying rhetorical roles of sentences in legal documents. Initial experiments for understanding the rhetorical/thematic roles in court case documents/judgements/case laws were developed as a part of achieving the broader goal of summarizing these documents [2, 3, 12]. For instance, Saravanan et al. [2] used Conditional Random Fields (CRF) [13] for the task on 7 rhetorical roles. Segmenting a document into functional (Introduction, Background, Analysis and Footnotes) and issue-specific parts (Analysis and Conclusion) was looked into by [4] on U.S. court documents using CRF with handcrafted features. A method for identification of factual and non-factual sentences was developed in [1] using fastText classifier. In another line of work, Walker et al [14] compared use of rule-based scripts (that require much lesser amount of training data) with Machine Learning approaches for the rhetorical role identification task.

Almost all prior attempts towards automatic identification of rhetorical roles in the legal domain have used handcrafted features. In contrast, this paper uses Deep Learning models for this task, where no handcrafted features are needed. Deep Learning (DL) methods are increasingly being applied in the legal domain, e.g., classification of factual and non-factual sentences in a legal document [1], crime classification [15, 16], summarization [6, 17] and other tasks. But, to our knowledge, DL methods have not yet been applied to the task of automatically identifying rhetorical roles of sentences in legal documents.

3. Dataset

In this paper, we consider legal judgments from the Supreme Court of India, crawled from the website of Thomson Reuters Westlaw India (`http://www.westlawindia.com`)[4]. We crawled 53,210 documents in total. Westlaw assigns each document a legal domain, such as 'Criminal', 'Constitutional', etc. We calculated the frequency of these domains, chose the top 5 domains and randomly sampled 50 documents from these 5 domains in proportion to their frequencies. Thus we have the following set of 50 documents from 5 domains – (i) Criminal – 16 documents (ii) Land and property – 10 documents (iii) Constitutional– 9 documents (iv) Labour and Industrial – 8 documents (v) Intellectual Property Rights – 7 documents. All experiments reported in this paper are performed on these 50 case documents.

4. Annotation Details

In this section we shall describe our annotation study, covering the rhetorical roles / semantic labels we consider in this work, the annotation procedure, and finally, analysis of inter-annotator agreement.

[4]We use only the publicly available full text judgement. All other proprietary information had been removed before performing the experiments.

4.1. Annotation Labels / Rhetorical Roles

Our annotators were three senior Law students from the Rajiv Gandhi School of Intellectual Property Law, India (*http://www.iitkgp.ac.in/department/IP*). Based on discussions with the annotators, we consider the following seven (7) rhetorical roles in our work.

1. **Facts (abbreviated as FAC)**: This refers to the chronology of events that led to filing the case, and how the case evolved over time in the legal system (e.g., First Information Report at a police station, filing an appeal to the Magistrate, etc.)
2. **Ruling by Lower Court (RLC)**: Since we are considering Supreme Court case documents, there were some judgements given by the lower courts (Trial Court, High Court) based on which the present appeal was made (to the Supreme Court). The verdict of the lower Court and the ratio behind the judgement by the lower Court was annotated with this label.
3. **Argument (ARG)**: The Court's discussion on the law that is applicable to the set of proven facts by weighing the arguments of the contending parties.
4. **Statute (STA)**: Established laws, which can come from a mixture of sources – Acts , Sections, Articles, Rules, Order, Notices, Notifications, Quotations directly from the bare act, and so on.
5. **Precedent (PRE)**: Prior case documents. Instructions similar to statute citations.
6. **Ratio of the decision (Ratio)**: Application of the law along with reasoning/rationale on the points argued in the case; Reason given for the application of any legal principle to the legal issue.
7. **Ruling by Present Court (RPC)**: Ultimate decision / conclusion of the Court following from the natural / logical outcome of the rationale

4.2. Annotation Process

The annotataors used GATE Teamware tool [18] to annotate the documents, following the methodology of [5, 10]. An annotation manual was developed in discussion with the annotators, containing descriptions and example sentences for each rhetorical role, along with other instructions (e.g., a label should be assigned to a full sentence and not a part of it, a sentence should have only one label, etc.). Initially, each annotator was asked to annotate 5 documents independently, i.e., without consulting each other. Then we had a joint discussion with all the annotators to resolve any issues, and refined the manual if necessary. This process was followed iteratively for annotation of the 50 documents.

4.3. Analysis of Inter-Annotator Agreement and Curation of Gold Standard

We compute the Inter-annotator Agreement (IAA) for the annotation task as follows.

IAA measure: As noted in [10], aggregated pairwise Precision, Recall and F-measure are more suitable measures for IAA than measures like Kappa. Following the same line, we compute these pairwise IAA measures using GATE's Annotation Diff tool.[5] Since we have three annotators (A_1, A_2 and A_3), we compute three sets of pairwise IAA (A_1, A_2), (A_2, A_3), (A_1, A_3), and then take the average of the three sets. We briefly define the metrics below.

GATE maintains three counts based on the extent to which two annotators' labels match. The three counts are as follows – **(1) Correct:** If for a sentence, the two annotators

[5] `https://gate.ac.uk/sale/tao/splitch10.html`

Table 1. Average inter-annotator agreement of the 3 annotators in terms of F-score as measured by GATE tool

Labels →	ARG	FAC	PRE	Ratio	RLC	RPC	STA
Strict	0.692	0.716	0.654	0.677	0.74	0.654	0.857
Lenient	0.953	0.934	0.878	0.908	0.925	0.968	0.967
Average	0.823	0.817	0.814	0.821	0.819	0.798	0.898

mark exactly the same span of text (covering all the words and punctuation marks) with the same label, then this is considered a Correct match. **(2) Partial:** If for a sentence, the two annotators mark the same label, but a different span of text (e.g., leaving few words or punctuation marks), then this sentence is considered a partial match. **(3) Missing and Spurious:** If for a sentence, the two annotators mark different labels, they are called *missing* or *spurious* (both terms used interchangeably). Based on the above definitions, Precision, Recall and F-score are calculated as follows:

$Precision = (Correct + 0.5 \times Partial)/(Correct + Spurious + Partial)$

$Recall = (Correct + 0.5 \times Partial)/(Correct + Missing + Partial)$

$Fscore = ((\beta^2 + 1) \times Precision \times Recall)/((\beta^2 \times Precision) + Recall)$

where β is the weighting of Precision vs. Recall. We use the default value of 1, meaning that both are weighed equally. For each of the Precision, Recall and F-score measures, GATE computes three variants as follows – **(1) Strict measure:** considers all partial matches as *incorrect* (spurious), **(2) Lenient measure:** considers all partial matches as *correct*, and **(3) Average measure:** average of the strict and lenient measures.

Analysis of F-scores: We primarily report the F-scores, since they combine both the Precision and Recall scores. The F-score IAA values computed by using GATE's Annotation Diff tool are presented in Table 1. As is expected, the strict scores are low and the lenient scores are quite high. This is due to differences in how different annotators use the graphical interface of the GATE tool. For instance, one of the annotators may have mistakenly excluded the full-stop (end of sentence marker) in a sentence while marking the label, while the other annotator included the full-stop.[6] The lenient method does not take into account these errors while the strict measure does.

Table 1 reports the IAA (F-score values) for each rhetorical role individually. In terms of strict scores, we observe Statute, Ruling by Lower Court and Facts have a high agreement whereas the scores are lower for Precedent and Ruling by Present Court. But in terms of lenient scores, all labels show high IAA of over 0.85. These IAA scores are comparable with what has been reported in similar prior studies [10].

Analysis of sentence-level agreement: To understand in more detail where the annotators tend to disagree, we perform a *sentence-level agreement study*. We construct an *agreement matrix C* (whose rows and columns are the labels) for two annotators A_x and A_y. An entry $C[i][j]$ of this matrix denotes the number of sentences which Annotator A_x labeled as L_i, but Annotator A_y labeled the *same* sentences as label L_j. Table 2 shows this agreement matrix for the annotator pair (A_2, A_3) who have the *lowest IAA* (as reported by GATE). Similar tables for the annotator pairs (A_1, A_2) and (A_1, A_3) are given in the Supplementary Information accompanying this paper.[7]

[6] Though clear instructions were given to include the end of sentence marker in the label, the annotators committed this mistake while marking some of the sentences.

[7] Supplementary Information: `http://cse.iitkgp.ac.in/~saptarshi/docs/Bhattacharya-et-al-JURIX19-SuppleInfo.pdf`

Table 2. Table showing the the sentence level agreement between the two annotators (A_2, A_3) who have the lowest IAA (0.79, as measured by GATE)

$A_2 \downarrow A_3 \rightarrow$	FAC	ARG	PRE	STA	Ratio	RLC	RPC
FAC	**2154**	5	0	3	40	8	0
ARG	17	**822**	16	1	0	0	0
PRE	0	11	**1425**	0	47	0	0
STA	0	0	0	**635**	12	2	0
Ratio	4	13	4	5	**3499**	1	0
RLC	47	1	0	0	25	**294**	0
RPC	6	0	0	0	21	0	**262**

The high values in the diagonal elements indicate that the annotators have a high overall agreement in general. Among the non-diagonal elements, we see relatively high values (signifying some disagreement or subjectivity) for some label-pairs. For instance, there is subjectivity among the label pairs (PRE, Ratio), (FAC, Ratio), (RLC, Ratio) and (RPC, Ratio), since the reason behind the final judgement (Ratio) depends on the facts (FAC), as well as judgements in prior cases (PRE) and the Ruling in the lower courts (RLC). There is also a tendency of annotators to differ between the labels (ARG, FAC) because framing the arguments relies on the facts of the case.

Analysis of agreement across domains: Since we have documents from five domains of law, we checked the average IAA F-score for the labels across each domain. We found that inter-annotator agreement is uniform across different domains. Detailed results can be found in the Supplementary Information.

Curation of the gold standard: The gold standard dataset was curated as follows: For a particular sentence, we took a *majority voting* of the labels given by the 3 annotators. There was a clear majority verdict regarding the label (rhetorical role) of each sentence. We use this annotated dataset in our experiments to automate the task of assigning semantic roles to sentences. Further statistics of the dataset are given in the Supplementary Information.

5. Methods for automatically identifying rhetorical roles

Now we describe our efforts towards automating the task of identifying rhetorical roles of sentences in a legal document. We treat this problem as a 7-class sequence labeling problem, where supervised Machine Learning models are used to predict one label (rhetorical role) for every sentence in a document.

Pre-processing the documents: Each document was split into sentences using the SpaCy tool (`https://spacy.io/`). Splitting a legal document into sentences is challenging due to frequent presence of abbreviations [19]. We observed SpaCy to do a reasonably good splitting (accuracy close to 90%), which agrees with observations in prior works [1]. There were 9,380 sentences in total in these 50 documents, as identified by SpaCy. Each such sentence was considered a unit for which one label (out of the seven rhetorical roles) is to be predicted.

Baseline: CRF with handcrafted features: As stated in Section 2, this is the approach adopted in most prior works. Each document is treated as a sequence of sentences. Some dependencies exist in the corresponding sequence of labels; e.g., RLC usually follow FAC, RPC is always the end label, etc. Conditional Random Fields (CRFs) [13] can be

used to model such sequences, since they consider both *emission scores* (probability of a label given the sentence) and *transition scores* (probability of a label given the previous label) while generating the label sequence.

To implement the baseline approaches [2,4], we represent each sentence as a vector of all features stated in these works – parts-of-speech tags (used in [4]), layout features (used in both [2, 4]), presence of cue phrases (used in [2]), and occurrence of named entities like Supreme Court, High Court in the sentence (used in [2]). The CRF works on these vectors to predict the labels (rhetorical roles).

We consider three baseline approaches: (1) CRF using the features of [2]; (2) CRF using the features of [4]; and (3) CRF using a combination of features from both [2,4].

Neural model 1: Hierarchical BiLSTM Classifier: We use a hierarchical BiLSTM (Bi-directional Long Short Term Memory) architecture [20] to automatically extract features for identifying the rhetorical roles. This requires us to feed the sequence of sentence embeddings to the BiLSTM, which returns a sequence of feature vectors. The BiLSTM model needs some initialization of the sentence embeddings, with which learning can start. We try two variations of sentence embeddings: (1) We construct sentence embeddings from *randomly initialized* word embeddings using another BiLSTM; and, (2) We used a large set of documents from the same domain to construct *pre-trained sentence embeddings*. Specifically, we used *sent2vec* [21] to construct the sentence embeddings from the set of 53K court case documents that we had collected (see Section 3), *excluding the 50 documents considered for this task*.

Neural model 2: Hierarchical BiLSTM CRF Classifier: The probability scores generated by the above model do not take into account label dependencies, and thus can be regarded as simple *emission scores*. To enrich the model further, we deploy a CRF on top of the Hierarchical BiLSTM architecture. This CRF is fed with the feature vectors generated by the top-level BiLSTM. As described above, we try both variations of sentence embeddings – randomly initialized embeddings, and pre-trained embeddings trained over a large set of legal documents.

6. Results and Analysis

We now compare the performance of the models (stated in the previous section) on the set of 50 manually-annotated documents (described in Section 4.3).

6.1. Experimental setup and evaluation metrics

We perform 5-fold cross validation with the 50 documents, which is a standard way of evaluating Machine Learning models. In each fold, we have 40 documents for training the model, and the other 10 documents for testing the performance of the model. The performance measures reported are averaged over all five folds.

Evaluation metrics: For a particular sentence, the label (rhetorical role) predicted by a model is considered to be correct, if it matches with the label assigned by the majority opinion of the human annotators (see Section 4.3). We use standard metrics for evaluating the performance of algorithms – macro-averaged Precision, Recall and F-score. For *macro-averaged metrics*, we compute these metrics for each class separately, and then take their average (to prevent any bias towards the high-frequency classes).

Table 3. Macro Precision, Recall and F-score of the baseline methods and neural network-based methods. Best performances highlighted in boldface.

Category	Method	Variations	Precision	Recall	F-score
Baselines (CRF with handcrafted features)	Features from [2]	-	0.4138	0.3308	0.4054
	Features from [4]	-	0.4580	0.4196	0.3250
	Features from [4] and [2]	-	0.5070	0.4358	0.4352
Neural models	Hier-BiLSTM	Pretrained emb	0.8168	0.7852	0.7968
		Random initialization	0.5358	0.5254	0.5236
	Hier-BiLSTM-CRF	Pretrained emb	**0.8396**	**0.8098**	**0.8208**
		Random initialization	0.6528	0.5524	0.5784

Table 4. F-score of the Hier-BiLSTM-CRF model, for the different labels, and for each domain of law. The last column indicates the average F-score for each domain. The last row indicates the average F-score for each of the seven labels (rhetorical roles).

	FAC	ARG	Ratio	STA	PRE	RPC	RLC	*Macro Average (across categories)*
Constitutional	0.903	0.659	0.909	0.832	0.904	0.857	0.85	0.845
Labour & Industrial Law	0.776	0.505	0.929	0.423	0.728	0.783	0.681	0.689
Criminal	0.836	0.567	0.945	0.689	0.891	0.917	0.865	0.816
Land & Property	0.847	0.624	0.908	0.841	0.845	0.98	0.778	0.832
Intellectual Property	0.832	0.607	0.927	0.824	0.901	0.964	0.886	0.849
Macro Average (across labels)	0.8388	0.5924	0.9236	0.7218	0.8538	0.9002	0.812	–

6.2. Comparing performances of different models

The comparative results are presented in Table 3. Clearly the neural models perform much better than the baselines, which shows that the latent features learnt by the neural models are better than the hand-crafted features used in prior works [2,4].

Effect of pre-trained embeddings: Using pretrained embeddings (learned over a large legal corpus) shows a high improvement in performance for both the neural models, as compared to using random initializations for the embeddings. Since deep neural models require lot of data to learn efficiently, it is especially beneficial to use pretrained embeddings learned over large domain-specific data.

Effect of combining CRF with neural model: Hier-BiLSTM-CRF performs only a little better than Hier-BiLSTM (both with pretrained embeddings). This is because legal documents consist of large sequences (average of 200 sentences per document), and we have few such documents; thus the CRF is unable to learn the transition scores well. Hence, there is not much additional benefit in combining CRF with the neural model.

6.3. Detailed analysis of the best performing model (Hier-BiLSTM-CRF)

Table 4 shows the F-score values of the best performing model (Hier-BiLSTM-CRF) for each of the seven labels and the five domains.

Performance on specific labels: From the last row of Table 4, we find that the model performs the best in predicting the Ratio and Ruling by Present Court (RPC). Ratio has the highest fraction of sentences in the corpus (38.63%), and this large amount of training data enabled this label to be predicted well. Ruling by the Present Court, though having less sentences (2.79% of the dataset), always has a fixed position – towards the end of a document. Hence this label could also be identified well.

Table 5. Label agreement matrix for labels assigned by (i) the best performing Hier-BiLSTM-CRF model, and (ii) majority opinion of the human annotators.

Human ↓ Model →	FAC	ARG	Ratio	STA	PRE	RPC	RLC
FAC	**1986**	109	43	28	35	0	18
ARG	265	**455**	49	22	52	0	2
Ratio	129	51	**3334**	33	72	3	2
STA	57	23	47	**461**	55	0	3
PRE	16	46	64	11	**1330**	1	0
RPC	0	0	9	1	3	**231**	18
RLC	33	5	7	0	7	8	**256**

The model performs satisfactorily for all other labels, except 'Arguments' (F-score of 0.5924). The 'Argument' sentences get interleaved with other labels. Additionally, only 9% of the total number of sentences in our corpus contribute to this label. Hence the neural model did not perform well in identifying these sentences.

Performance across Domains: The last column of Table 4 shows how generalizable the model is across the 5 different domains. The model gives consistent performance (F-score in $[0.82-0.86]$) across all the domains, except for 'Labour & Industrial law'. This performance is consistent with the inter-annotator agreement scores, where the IAA was low for the domain 'Labour & Industrial law' (see Supplementary Information).

Comparing inter-annotator agreement and annotator-model agreement: We now compare the agreement between the human annotators (IAA), and agreement between the model and the annotators. We create an agreement matrix (Table 5), where the rows represent the human-assigned labels (majority opinion of the annotators), and the columns represent the labels assigned by the model. The *diagonal elements* show the number of sentences for which the model-assigned label matches with the human-assigned label. The *non-diagonal* elements $C[i][j]$ shows the number of sentences where the human-assigned label i does not match the model-assigned label j.

We focus on the non-diagonal elements that have relatively high values. For instance, the model seems to have frequent confusion between the labels Arguments (ARG) and Facts (FAC), and between the labels Ratio, Fact and Precedence (PRE). Comparing Table 5 with the IAA agreement matrix (Table 2), we find that these label-pairs are exactly the ones where the IAA values were also low, i.e., there is sufficient confusion around these label-pairs even among the human annotators. This observation suggests that these rhetorical roles are largely subjective. Hence it is natural that the model will also face some difficulty in identifying these subjective rhetorical roles.

7. Conclusion and Future Work

We show that deep learning models can much better identify rhetorical roles of sentences in legal documents, compared to methods using hand-crafted features. We also perform an extensive annotation study, and analyse the agreement between different human annotators, as well as the agreement of the model with the annotators.

The principal advantage of neural models is that no hand-crafting of features is needed, hence expensive legal expertise is not essential. However, this property also poses difficulties in understanding why exactly a sentence is more likely to be assigned to one rhetorical role than the others. Thus, neural models trade-off explainability/transparency with the cost of hand-crafting features. Deep Learning models can be

used for tasks like identifying rhetorical roles of sentences, if it can be assumed that achieving good performance is more important than transparency.

In future, we plan to check how deep learning models generalize across different jurisdictions, by experimenting on legal documents of other countries.

Acknowledgments. The authors thank the law students who annotated the sentences. The research is partially supported by SERB, Government of India, through the project 'NYAYA: A Legal Assistance System for Legal Experts and the Common Man in India'. P. Bhattacharya is supported by a Fellowship from Tata Consultancy Services.

References

[1] I. Nejadghoii, R. Bougueng and S. Witherspoon, A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases, in: *Proc. JURIX*, 2017.
[2] M. Saravanan, B. Ravindran and S. Raman, Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization, in: *Proc. International Joint Conference on Natural Language Processing: Volume-I*, 2008.
[3] A. Farzindar and G. Lapalme, Letsum, an Automatic Legal Text Summarizing System, 2004.
[4] J. Savelka and K.D. Ashley, Segmenting U.S. Court Decisions into Functional and Issue Specific Parts, in: *Proc. JURIX*, 2018.
[5] O. Shulayeva, A. Siddharthan and A.Z. Wyner, Recognizing cited facts and principles in legal judgements, *Artificial Intelligence and Law* **25**(1) (2017), 107–126. doi:10.1007/s10506-017-9197-6.
[6] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh and S. Ghosh, A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments, in: *Proc. ECIR*, 2019.
[7] M. Saravanan, Ontology-Based Retrieval and Automatic Summarization of Legal Judgments.
[8] D. Jin and P. Szolovits, Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts, in: *Proc. EMNLP*, 2018.
[9] G. Venturi, Design and development of TEMIS: a syntactically and semantically annotated corpus of italian legislative texts, in: *Proc. Workshop on Semantic Processing of Legal Texts (SPLeT)*, 2012.
[10] A.Z. Wyner, W. Peters and D. Katz, A Case Study on Legal Case Annotation, in: *Proc. JURIX*, 2013.
[11] A. Wyner, Towards annotating and extracting textual legal case elements, *CEUR Workshop Proceedings* **605** (2010), 9–18.
[12] B. Hachey and C. Grover, Extractive summarisation of legal texts, *Artificial Intelligence and Law* **14**(4) (2006), 305–345.
[13] J. Lafferty, A. McCallum and F.C.N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).
[14] V.R. Walker, K. Pillaipakkamnatt, A.M. Davidson, M. Linares and D.J. Pesce, Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning, in: *Proc. Workshop on Automated Semantic Analysis of Information in Legal Texts (with ICAIL)*, 2019.
[15] P. Wang, Z. Yang, S. Niu, Y. Zhang, L. Zhang and S. Niu, Modeling dynamic pairwise attention for crime classification over legal articles, in: *Proc. ACM SIGIR*, 2018.
[16] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang and J. Guo, Hierarchical Matching Network for Crime Classification, in: *Proc. ACM SIGIR*, 2019.
[17] C.-L. Liu and K.-C. Chen, Extracting the Gist of Chinese Judgments of the Supreme Court, in: *Proc. ICAIL*, 2019.
[18] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, in: *Proc. ACL*, 2002.
[19] G. Sanchez, Sentence Boundary Detection in Legal Text, in: *Proc. Natural Legal Language Processing Workshop*, 2019.
[20] A. Graves, S. Fernández and J. Schmidhuber, Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition, in: *Proc. Int'l Conf. on Artificial Neural Networks (ICANN)*, 2005.
[21] M. Pagliardini, P. Gupta and M. Jaggi, Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features, in: *Proc. NAACL*, 2018.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190302

Improving the Processing of Question Answer Based Legal Documents

Saurabh CHAKRAVARTY [1], Maanav MEHROTRA,
Raja Venkata Satya Phanindra CHAVA, Han LIU, Matthew KRIVANSKY and Edward A. FOX
Virginia Tech, Blacksburg, VA 24061 USA

Abstract.
In the legal domain, documents of various types are created in connection with a case. Some are transcripts prepared by court reporters, based on notes taken during the proceedings of a trial or deposition. For example, deposition transcripts capture the conversations between attorneys and deponents. These documents are mostly in the form of question-answer (QA) pairs. Summarizing the information contained in these documents is a challenge for attorneys and paralegals because of their length and form. Having automated methods to convert a QA pair into a canonical form could aid with the extraction of insights from depositions. These insights could be in the form of a short summary, a list of key facts, a set of answers to specific questions, or a similar result from text processing of these documents. In this paper, we describe methods using NLP and Deep Learning techniques to transform such QA pairs into a canonical form. The resulting transformed documents can be used for summarization and other downstream tasks.

Keywords. NLP, QA Normalization, Chunking, Deep learning, Legal Deposition

1. Introduction

Documents such as legal depositions comprise conversations between a set of two or more people, with the goal of identifying observations and the facts of a case. These conversations are in the form of discrete question-answer (QA) pairs. Like other general conversations, these documents are noisy, only loosely following grammatical rules.

Humans, because of their prior learning and experience, readily understand such documents since the number of types of questions and answers is limited. These types provide strong semantic clues that aid comprehension. Accordingly, we seek to leverage the QA types found, to aid textual analysis.

Classifying each QA pair type can ease the processing of the text, which in turn can facilitate downstream tasks like question answering, information retrieval, summarization, and knowledge graph generation. This is because special rules can be applied to each QA type, allowing transformations that are oriented to supporting existing NLP tools. This can facilitate text parsing techniques like constituency, syntax, and depen-

[1]Corresponding Author: Saurabh Chakravarty, Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA; Email:saurabc@vt.edu

dency parsing, and also enable us to break the text into different chunks based on part-of-speech (POS) tags using techniques like Chunking and Chinking.

Dialog Acts (DAs) [1,2] can represent the communicative intention behind a speaker's utterance in a conversation. Identifying the type of DA of each question and answer in a conversation [3] thus is a key first step in automatically determining intent and meaning.

Unfortunately, automatically transforming sentences based on DAs isn't straightforward. But, a possible solution is to transform the most prevalent combinations.

For a given type of QA pair, with its pair of types of question and answer DAs, we want to convert the QA pair into a canonical form. Table 1 shows an example question and answer, each with its respective dialog act, along with the desired canonical form.

Table 1. A QA pair with its canonical form.

Type	**Text**	**Dialog Act**
Question	Were you able to do physical exercises before the accident?	bin
Answer	Yes. I used to play tennis before. Now I cannot stand for more than 5 minutes.	y-d
Canonical Form	I was able to do physical exercises before the accident. I used to play tennis before. Now I cannot stand for more than 5 minutes.	-

As part of our work in Dialog Act (DA) classification [3], we observed common patterns associated with deposition QA pairs according to the different question and answer dialog acts. For each such common pattern, we can use traditional NLP parsing techniques like Chunking and Chinking [4] and create custom transformation rules to transform the text into a canonical form. Section 4.3.1 describes Chunking and Chinking in more detail. Section 4.3.2 describes an alternative approach for transformation into a canonical form, using Deep Learning.

The core contributions of this work are as follows.

1. An annotated dataset of QA pairs along with their Dialog Acts and canonical forms.
2. A collection of analysis and transformation methods using traditional NLP techniques like Chunking and Chinking.
3. A collection of Deep Learning based pre-trained sentence transformation models that can transform a QA pair into a canonical form.

2. Related Work[2]

Our earlier work [3] describes our ontology of Dialog Acts for legal depositions. This work also used two datasets to identify the various types of DA present in the deposition questions and answers. Deep Learning based classification methods were used to identify the DA associated with each of the question and the answer portion of a QA pair. For the current study, we re-purposed the DA classification methods in [3].

Once the types of DA for a QA pair have been identified, we want to transform the text into a canonical form. We have not been able to find any work that proposed a

[2]While we have completed an extensive literature review, limitations imposed on this submission have forced only mentioning a few of the many related works.

solution to this kind of problem. Traditional NLP based parsing techniques like Chunking and Chinking [4] can parse the constituents of a sentence based on the part-of-speech (POS) tags. These methods have been implemented in NLP libraries like NLTK [5] and spaCy [6] and have very good performance. Though the efficacy of these libraries is generally task based, an empirical analysis of the results helps make the best choice [7]. For our work we used the NLTK library for performing Chunking and Chinking.

Transforming a QA pair into a canonical form also can be formulated as a machine translation problem. Though we have the same source and target languages, the input and output differ in form. Works like [8] employ an encoder-decoder based approach to translate text from one language to other.

Work in COPYNET [9] added the idea of copying from the source input in sequence-to-sequence models. Pointer Generator Network (PGN) [10] is an abstractive summary generation system that used the same idea as COPYNET, but added more optimizations on how the summary is generated. It addressed two challenges: avoiding the generation of inaccurate text in the summaries, and controlling the repetition of text. During the training process, the system learns whether to generate or copy from the input sentence, and also to minimize the repetition while maximizing the probability of the generated sequence. We used the PGN architecture to transform a QA pair into a canonical form.

3. Datasets

For our work, we used DA combinations from datasets that each were a collection of depositions. We also curated the ground truth for our experiments for these datasets. The following sections describe these in more detail.

3.1. Dataset Description

We used depositions from a proprietary as well as a public dataset. The details for these datasets are as follows.

- **Mayfair Dataset** - This was a proprietary dataset that was provided to us by Mayfair Group LLC. This collection is comprised of 350 depositions. We randomly selected 10 depositions from this collection. Table 2 shows the distribution of the top 10 question-answer DA combinations across the Mayfair dataset.
- **Tobacco Dataset** - This dataset comes from the 14 million Truth Tobacco Industry Documents that are publicly accessible [11]. Over 2,000 of these are deposition transcripts. We randomly selected 8 depositions from this collection. Table 3 shows the distribution of the top 10 question-answer DA combinations across the tobacco dataset.

3.2. Dataset Annotation

One of the authors, along with volunteers selected by Mayfair, annotated the ground truth for the datasets. This involved annotating each QA pair with a simple sentence or other suitable canonical form of the QA pair. In our experiments we made use of about 4000 and 3300 annotated pairs for the Mayfair and tobacco datasets respectively.

Table 2. Distribution of the Top 10 DA combinations for the proprietary Mayfair dataset.

Question DA	Answer DA	# of samples	% of Total
wh	sno	517	13.00
bin	y	326	8.20
bin-d	y	322	8.10
bin	sno	277	6.96
bin	n	270	6.79
bin-d	sno	177	4.45
sno	sno	159	4.00
ack	sno	142	3.57
wh-d	sno	121	3.04
bin	y-d	99	2.49

Table 3. Distribution of the Top 10 DA combinations for the tobacco dataset [11].

Question DA	Answer DA	# of samples	% of Total
bin-d	sno	454	13.58
bin	sno	441	13.19
wh	sno	297	8.88
bin-d	y	235	7.02
bin	y	183	5.47
bin	n	143	4.27
sno	sno	143	4.27
bin	y-d	118	3.52
bin	dno	95	2.84
bin-d	y-d	92	2.75

4. Methods

4.1. Dialog Acts:

For our task of transformation, classifying the Dialog Acts (DAs) [1,2] would aid in isolating and grouping QA pairs of similar type. Custom rules can be developed for each DA type to process a conversation QA pair and transform it into a suitable form for subsequent analysis. Using methods to classify the DAs in a conversation thus would help us delegate the transformation task to the right transformer method. We have used the ontology and the methods in [3] to classify the DAs in our dataset.

4.2. Pre-processing:

The text in the QA pairs contained noise which needed to be removed to perform the transformation step in an efficient way. Table 4 shows some sample questions with the noise that we needed to remove via pre-processing.

For some DAs, the question and answer text also consisted of a well formed sentence in the beginning and the end, as shown in Table 5. We used text-processing techniques along with regular expression based rules to separate the declarative part from the question and the answer.

Table 4. Questions, with the noisy text in bold.

You also mentioned earlier that he busted his lips**; is that correct?**
Okay. So you mentioned you had a son**; correct?**
I see. So, did you think it was the bartender?

Table 5. Questions and answers that include a well formed sentence. Declarative parts shown in bold.

Text	Dialog Act
And the damage that you showed earlier in the diagram, you said that damage was accidental?	bin-d
And a fracture that runs through the whole arm joint is a pretty severe fracture. When was the examination done?	wh-d
Yes. **We sent out this to that operating company.**	y-d
No. **I did not read any depositions or I think the second part is kind of general, but I haven't read any depositions.**	n-d

4.3. Transformation

We used two different methods to transform the QA pair into a canonical form. The following sections describe the methods in more details.

4.3.1. Transformation via Chunking and Chinking

Chunking refers to the process of extracting chunks from a sentence based on certain POS tag rules. These rules are represented using simple regular expressions. Chinking refers to the process of defining what is not to be included in a chunk. A Chunking process creates chunks and Chinking breaks up those chunks into more granular parts using rules. Referring to the example present in Table 1, we started with the question text and created a simple sentence parse tree as shown in Figure 1. Then we broke it up into a chunk based on a preposition rule of *"<.*>?<PRP><.*>?."* This rule specifies that any preposition that has any POS tag before and after it should be extracted as a chunk. In this case, it extracted "Were" and "able" that were before and after the preposition word. Figure 2 shows the chunk formed as part of the Chunking process.

Figure 1. Sentence root.

Figure 2. Extracting a chunk based on a rule.

For transformation to a canonical form, we needed to transform the identified chunk into a first person description. This description will be from the perspective of the depo-

nent. The transformed sentence in this case would be *"I was able to do physical exercises before the accident"*. We swapped the position of "were" and "you" in the chunk tree and transformed "you" to "I" and "were" to "was". For each of these simple transformations of a QA pair word to a canonical form word, we created a dictionary entry to keep track of that transformation. The dictionary was expanded to account for different transformations that were required for other words that needed to be transformed. We iteratively improved our transformation based on the results we observed from the data. We developed specific methods for each combination of a question and answer DA.

4.3.2. Transformation via Deep Learning.

The Deep Learning based transformation was implemented with a prototype we devised to evaluate the feasibility of using Deep Learning based methods. There are no known works that have addressed our exact problem, so we investigated how Deep Learning based models would perform for this task. We used the OpenNMT Toolkit [12] to train sentence transformers for the different combinations of DA.

Deep Learning models are dependent on a large number of training examples; this is more pronounced for sequence-to-sequence models where there are a large number of parameters in play. Since the amount of training data we could obtain was limited, we focused our collection of training data on a particular set of the combinations of DA. In particular, we only developed Deep Learning based methods for the combinations of [bin, y], [bin, n], [bin, y-d], and [bin, n-d].

4.4. Evaluation Methods.

Evaluation of text processing and transformation is much more difficult than for simple classification since the results are often subjective. For our preliminary evaluation studies, we started by using ROUGE-1/2 scores and sentence similarity for evaluation. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [13] can be used to compare generated sentences with the canonical forms annotated by human actors. We used the ROUGE-1 and ROUGE-2 scores, which measure how much generated sentences overlap with the uni-gram and bi-gram representation of the annotated canonical forms.

Another evaluation metric we used is sentence similarity. Transforming a pair of sentences to their vector space representations and measuring their cosine-similarity can be used to measure sentence similarity. For that transformation, we used InferSent [14] to generate sentence embedding vectors; it is a based on fastText[15] word embeddings.

4.5. Experimentation

For our experiments, we considered the top 11 DA classes for the proprietary dataset as given in Table 2. The top 11 DA combinations represented more than 65% and 60% of the total data for the proprietary and tobacco datasets, respectively. This was a good set to target for our work. The DA combinations that we left out each represented less than 3% of the data. We plan to develop methods for these DA combinations in future work.

We developed transformation methods involving Chunking methods for 10 of the 11 DA classes. Regarding the ["bin-d", "sno"] DA combination, we found the question text to be problematic for transformation via our methods. The DAs for most of the questions were incorrectly classified as "bin-d", whereas it had a mix of "bin-d", "wh-d" and "sno".

For this reason, we omitted occurrences of the [“bin-d”, “sno”] DA combination from our experiments. We will address this omission in future work.

Table 6 describes the transformation methods used for our experiments.

Table 6. The transformation methods used for the experiments.

Method	Description
Just Answer	In this method we used the answer as is.
Question and Answer	In this method we used the combination of the concatenated question and answer text.
Chunking based Transformation	In this method we performed DA classification followed by Chunking based transformation.
Deep Learning based Transformation	In this method we performed DA classification followed by Deep Learning based transformation.

5. Results

5.1. Experiment Results and Analysis

The following discussion is of studies with the Mayfair dataset. Table 7 shows the results of the transformation experiments for the four different methods. We calculated the ROUGE-1(R-1)/2(R-2) and the similarity (Sim) scores between the ground truth and the generated sentence. We averaged the scores across all of the samples, for each DA combination. The following sections discuss the results in more detail for each method.

Table 7. Evaluation Results. Best results are highlighted in bold.

Qstn DA	Ans DA	Just Answer			Q+A			Chunking		
		R-1	R-2	Sim	R-1	R-2	Sim	R-1	R-2	Sim
wh	sno	0.73	0.67	0.79	0.77	0.66	**0.87**	**0.78**	**0.70**	0.86
bin	y	0.09	0.03	0.29	0.75	0.56	0.84	**0.85**	**0.70**	**0.90**
bin-d	y	0.016	0.002	0.11	0.81	0.70	0.90	**0.9**	**0.81**	**0.93**
bin	sno	0.67	0.63	0.76	0.83	0.75	0.90	**0.84**	**0.79**	**0.91**
bin	n	0.08	0.04	0.36	0.72	0.54	0.81	**0.83**	**0.70**	**0.91**
sno	sno	0.67	0.62	0.73	**0.9**	**0.85**	**0.95**	0.85	0.80	0.92
ack	sno	**0.98**	**0.98**	**0.99**	0.94	0.93	0.94	**0.98**	**0.98**	**0.99**
wh-d	sno	**0.82**	**0.78**	**0.87**	0.64	0.57	0.78	**0.82**	**0.78**	**0.87**
bin	y-d	0.55	0.47	0.68	**0.78**	**0.65**	**0.87**	0.75	**0.65**	0.82
bin	n-d	0.45	0.31	0.74	**0.59**	**0.43**	**0.80**	0.54	0.39	0.78

5.1.1. Use answer (results for top 5 combinations):

- *wh |sno* - The transformer performance was quite reasonable for both the ROUGE scores and similarity. For the best scores, we observed that the answer is descriptive and has a good overlap with the ground truth. For the worst scores, we observed that the answer was short and lacked the context that was present in the question
- *bin |y* - The transformer performance was poor for this case. This happens because the answer DA is “y” and in such cases the answer is in the form of “yes” or “yeah”, which does not contain enough context to match well with the ground truth.

- *bin-d |y* - The transformer performance was poor for this case, as with the previous one. The scores are also of similar nature and for the same reasons.
- *bin |sno* - The transformer performance was quite reasonable. The reasoning for this is similar to what was dicussed in wh |sno pair.
- *bin |n* - The transformer performance was poor for this case, similar to the bin |y combination. The scores are also of similar nature and for the same reasons.

5.1.2. Use question and answer (results for each DA combination):

- *wh |sno* - The transformer performance was very good for both the ROUGE scores and similarity. For the best scores, we observed that the answer is descriptive and has a high chance of having a good overlap with the ground truth. For the worst scores, we observed that the generated text contained the text from both question and the answer, whereas the ground truth was a good paraphrase of the same.
- *bin |y* - The transformer performance was very good for both the ROUGE scores and similarity. This happens because the answer DA is "y" and in such cases the answer is in the form of "yes" or "yeah", but the question contains enough context to have a good overlap with the ground truth.
- *bin-d |y* - The transformer performance was very good for this case, similar to the previous one. The scores are a little better, but for the same reason that the question and answer together in one sentence is bound to have good overlap and similarity with the ground truth.
- *bin |sno* - The transformer performance was very good for this case, similar to the previous one.
- *bin |n* - The transformer performance was reasonably good for this case. We observed higher scores for simple questions and long answer combinations. This because a combination of the two provides enough context. For the worst scores, we observed a high similarity score but a poor ROUGE-2 score. The generated sentence had a very poor bi-gram overlap with the ground truth.

5.1.3. Transformation via Chunking (results for each DA combination):

- *wh |sno* - The transformer performance was very good for both the ROUGE scores and similarity. For the other methods there were very rare or no occurrences of perfect ROUGE-2 scores. This underlines that the Chunking based methods had a good paraphrasing ability that matched the annotated ground truth. For the worst scores, we observed that the generated text was a good paraphrase of the question and answer, but it was not of the exact form as the ground truth.
- *bin |y* - The transformer performance was very good for this case, similar to the *wh |sno* case.
- *bin-d |y* - The transformer performance was very good for this case, similar to the previous one.
- *bin |sno* - The transformer performance was very good for this case. There were many instances of perfect ROUGE-2 scores. For the worst scores we observed that the Chunking based transformers were not able to break the QA pair using the predefined grammar rules and hence emitted the answers for these cases.
- *bin |n* - The transformer performance was reasonably good for this case. It was the best among all the methods used for the ROUGE scores and similarity. There were

many instances of perfect ROUGE-2 scores. For the worst scores, the Chunking based method had challenges with the grammar and some generated bi-grams had an incorrect form.

5.1.4. Transformation using Deep Learning:

We broke the dataset into a 70-20-10 proportion for training, validation, and test. Separate models were trained using the annotated data which was run for all 4 DA combinations. The results as shown in Table 8. The modest results could be attributed to the fact that we had very little training data to train with. The results do indicate a potential to improve with more training data. We plan to address this in our future work.

Table 8. Deep Learning results.

Question DA	Answer DA	ROUGE-1	ROUGE-2	Sentence Similarity
bin	y	0.6	0.38	0.73
bin	n	0.71	0.54	0.83
bin	y-d	0.48	0.26	0.74
bin	n-d	0.44	0.24	0.67

6. Conclusion and Future Work

We developed methods to transform a QA pair in a legal deposition to a canonical form. We used traditional NLP based techniques like Chunking and Chinking, along with methods based on Deep Learning. We found that the transformation methods based on Chunking had the best ROUGE-2 scores in 8 of the 10 DA combinations and had the best semantic similarity scores in 6 out of the 10 DA combinations. For most of the other comparisons, NLP techniques were competitive with the other best results.

To confirm the findings reported above for the Mayfair dataset, we ran additional experiments on the tobacco dataset. The results indicated equally good transformation performance in 8 of the 10 DA classes for the Chunking based methods. This indicates generality of the transformation methods across datasets.

As per our knowledge, this is the first work of its kind that transforms a QA pair into a canonical form. Given the encouraging results, we plan to improve it further and scale up the experiments with a larger corpora and additional evaluations.

We plan to improve the DA classification by adding a pre-processing step so that it can break a long question into a series of statements and questions. This would allow the classifier to be applied to shorter texts, which should result in increased DA accuracy.

We also plan to generate word embeddings for the legal domain, especially for depositions. We can use the BERT [16] system to train on a large deposition corpora and learn embeddings that are specific to legal depositions.

For the Deep Learning based transformers, we plan to train with more data and more DA combinations to improve transformation efficacy. Using grammatical correctness as a constraint for the generation of transformed text should improve results further.

We plan to refine our evaluation methods by using human actors to subjectively evaluate the quality of the transformed sentences using criteria like readability, context and polarity retention, and grammatical correctness.

Acknowledgments. This work was made possible by Virginia Tech's Digital Library Research Laboratory (DLRL). Data in the form of legal depositions was provided by Mayfair Group LLC, which also managed obtaining annotations. In accordance with Virginia Tech policies and procedures and our ethical obligations as researchers, we are reporting that Dr. Edward Fox has an equity interest in Mayfair Group LLC, whose data was used in this research. Dr. Fox has disclosed those interests fully to Virginia Tech, and has in place an approved plan for managing any potential conflicts arising from this relationship.

References

[1] D. Jurafsky, E. Shriberg, B. Fox and T. Curl, Lexical, prosodic, and syntactic cues for dialog acts, in: *Stede, Manfred, Leo Warner, and Eduard Hovy (eds.): Discourse Relations and Discourse Markers. Proceedings of the workshop, 15 August, Montreal, Quebec, Canada: COLING-ACL '98.*, New Brunswick, NJ: Association for Computational Linguistics, 1998, pp. 114–120.

[2] J. Williams, A belief tracking challenge task for spoken dialog systems, in: *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, 2012, pp. 23–24.

[3] S. Chakravarty, R.V.S.P. Chava and E.A. Fox, Dialog Acts Classification for Question-Answer Corpora, in: *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), June 21, 2019, Montreal, QC, Canada*, 2019.

[4] F. Zhai, S. Potdar, B. Xiang and B. Zhou, Neural models for sequence chunking, in: *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] E. Loper and S. Bird, NLTK: the natural language toolkit, *arXiv preprint cs/0205028* (2002).

[6] spaCyCommunity, *spaCy: Industrial-Strength Natural Language Processing in Python*, 2016, `https://spacy.io`.

[7] F.N.A. Al Omran and C. Treude, Choosing an NLP library for analyzing software documentation: a systematic literature review and a series of experiments, in: *Proceedings of the 14th International Conference on Mining Software Repositories*, IEEE Press, 2017, pp. 187–197.

[8] K. Cho, B. van Merrienboer, D. Bahdanau and Y. Bengio, On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, in: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 103–111.

[9] J. Gu, Z. Lu, H. Li and V.O. Li, Incorporating copying mechanism in sequence-to-sequence learning, *arXiv preprint arXiv:1603.06393* (2016).

[10] A. See, P.J. Liu and C.D. Manning, Get to the point: Summarization with pointer-generator networks, *arXiv preprint arXiv:1704.04368* (2017).

[11] U. Library and Center for Knowledge Management, *Truth Tobacco Industry Documents*, 2002, `https://www.industrydocuments.ucsf.edu/tobacco`.

[12] G. Klein, Y. Kim, Y. Deng, J. Senellart and A.M. Rush, Opennmt: Open-source toolkit for neural machine translation, *arXiv preprint arXiv:1701.02810* (2017).

[13] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.

[14] A. Conneau, D. Kiela, H. Schwenk, L. Barrault and A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 670–680.

[15] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, Advances in Pre-Training Distributed Word Representations, in: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[16] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR* **abs/1810.04805** (2018).

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190303

Weakly Supervised One-Shot Classification Using Recurrent Neural Networks with Attention: Application to Claim Acceptance Detection

Charles CONDEVAUX [a,1], Sébastien HARISPE [b], Stéphane MUSSARD [a] and Guillaume ZAMBRANO [a]
[a] *CHROME, Univ. Nîmes, France*
[b] *LGI2P, IMT Mines Alès, Univ. Montpellier, Alès, France*

Abstract. Determining if a claim is accepted given judge arguments is an important non-trivial task in court decisions analyses. Application of recent efficient machine learning techniques may however be inappropriate for tackling this problem since, in the Legal domain, labelled datasets are most often small, scarce and expensive. This paper presents a deep learning model and a methodology for solving such complex classification tasks with only few labelled examples. We show in particular that mixing one-shot learning with recurrent neural networks and an attention mechanism enables obtaining efficient models while preserving some form of interpretability and limiting potential overfit. Results obtained on several types of claims in French court decisions, using different vectorization processes, are presented.

Keywords. Classification, legal analysis, one-shot learning, deep learning.

1. Introduction

Text classification has long been identified as an important topic for Computer Science applications in the Legal domain [1]. A large diversity of applications can indeed be framed around classifying legal *entities* that are, or can be represented as, sequences of words, e.g. cases, decisions, claims, contracts. Interest for text classification has for instance be motivated by the recurrent need expressed by lawyers to find the most relevant cases according to specific contexts of interest [2] –text classification can indeed be used to structure case corpora by populating a predefined organization of cases that will further be used to improve case retrieval. Such classification techniques can also be used for filtering cases, and deciding whether it is relevant or not for a law firm to accept or reject a new case [3]. Recent applications for analyzing the impact of legal change through case classification have also been proposed [4]. Other examples of applications in the legal domain are: legal norms classification [5], detection of the semantic type of

[1]Corresponding Author: charles.condevaux@unimes.fr. Granted by Région Occitanie: project PREMAT-TAJ.

legal sentences [6,7], detection of clause vagueness [8], or prediction of supreme court rule and the law area to which a case belongs to [9].

This paper presents our work on the definition of a deep learning model and a methodology for solving complex text classification tasks with only few labelled examples. The selected application is a general court decision classification setting applied on a French corpus of court decisions – court decision classification is an important non-trivial task in court decisions analyses. Court decisions have been labelled based on the acceptance of claims. In that context, we study in particular how mixing one-shot learning with recurrent neural networks and attention mechanisms in order to obtain efficient models.

The paper is organized as follows. Section 2 reviews some related works. Section 3 presents the model. Section 4 presents the datasets and the word vectorizations used in our experiment. Finally, Section 5 discusses the results.

2. Related works

Rule-based approaches and Machine Learning (ML) approaches are generally distinguished in text classification [5]. Rule-based classification systems rely on predefined domain expert rules, e.g. if the decision contains the utterance "*Article 700*" then label it with class `Damage`. The broad literature related to these approaches cannot be reduced to this simple example. Nevertheless, even if effective and relevant in specific cases[2], rule-based approaches *de facto* suffer from the need to express rules and to manage rule interactions for ensuring good performance. ML approaches may be used to overcome this limitation by implicitly inferring the decision rules of interest to drive efficient classification.[3] From a labelled dataset composed of numerous classification examples, learning algorithms are used for building predictive models. These approaches are today often preferred and have proven successful in numerous application contexts.

A large literature in Machine Learning, Natural Language Processing (NLP) and Computational Linguistics studies (text) classification. Among the most popular models widely used for text classification since the past two decades, we can cite: Naive Bayes Classifier, Logistic Regression, Random Forest, Support Vector Machine (SVM) and Multilayer Perceptron. These models require defining a vector representation of a text that will later be considered as model input. Specific and often *ad hoc* features of interest are sometimes used for building these representations – e.g. a boolean feature could be "*Does the text contains an utterance of 'Article 700*?' ". To overcome the limitation of manually defining features, information related to the words composing the vocabulary used in the text corpora is often used as features. From simple bag-of-words and vector space models from the late 60s, to more refined weighting scheme modelling word relevance for classification, e.g. TF-IDF, these approaches have led to the definition of efficient models able to automatically solve interesting problems framed within text classification [10]. For instance, SVM have been used to perform legal norm classification with an accuracy of more that 90% for more than 13 different classes [5]. Using the same model, with TF-IDF vector representations, accuracy rates up to 94% have also

[2]Interpretability and the fact that these approaches do not rely on datasets may be interesting advantages.

[3]We focus on supervised ML, the traditional setting considered for text classification.

been obtained in a sentence classification task [11]. Despite these encouraging successes, more complex problems related to text classification are still out of reach.

The recent developments in Deep Learning have led to a fruitful diversity of radically new efficient neural network-based classification models, among which specific developments are of particular interest for text classification. Recurrent neural networks, such as Long Short-Term Memory (LSTM), are very useful for processing sequential data (e.g. such as texts, sequences of words) [12]. Embedding techniques have been developed and refined for encoding entities of interest, such as words, sentences, or texts, in law dimension spaces (e.g., BERT, ELMo, FastText) – these representations can next be used for classification or other ML pipelines [13,7]; note that these representations do not need to be defined explicitly through feature definition, as done in traditional Machine Learning approaches.

Attention mechanisms are also developed for better identifying and incorporating important information during the decision process. Technical aspects related to these approaches and techniques will later be introduced. They have led to very interesting performance improvements in various popular challenges offered to text classification [14,15,16]. Nevertheless, due to their intrinsic properties, deep learning models require large (labelled) datasets to be trained. This is an important issue for their use in the legal domain since it is most often difficult to mobilize experts in this domain, generally leading to data scarcity with only expensive and small labelled corpora available [4]. This limitation contributes to explaining the reduced amount of works on the use of Deep Learning for text classification in the legal domain. Active researches in ML focus on reducing the need of labelled data using (i) approaches to reuse models trained in related contexts (e.g., transfer learning, fine-tuning), (ii) by exploiting unlabeled data (e.g. via embeddings), or (iii) by exploiting as much as possible the information expressed in labelled data (e.g, one-shot learning, siamese neural networks).

Applying advanced deep learning techniques on small datasets is indeed possible given the right setup while avoiding overfitting. A strategy experimented in this paper is to implement one-shot learning aiming at solving classification tasks only using few examples [17]. This approach is today mainly used in computer vision [18] with memory-augmented networks [19] but can be adapted to NLP [20], or even to estimate word embeddings [21]. Instead of learning to directly map an input to an output class, the one-shot approach implemented using siamese networks aims at estimating a similarity function between pairs of observations [22]. This problem can be reduced to a binary classification task by setting a given label if both inputs are *similar* (i.e. share the same original label). Using such a discriminant approach, a model can be learned from a single example per class. However, since this task is non trivial in NLP due to the sequential aspect of language, entire datasets are generally used instead.

Applications and development such advanced deep learning techniques have to be encouraged in the legal domain in order to fully benefit from recent advances in Machine Learning. This paper presents how they can be used to classify judge decisions.

3. One-shot learning using a siamese recurrent network with attention

The proposed model implements one-shot learning using a siamese recurrent network with an attention mechanism to fine tune sentence representations. These embeddings are next reused jointly with selected features to solve a classification task.

3.1. General architecture

The general architecture of our model is presented in Figure 1.

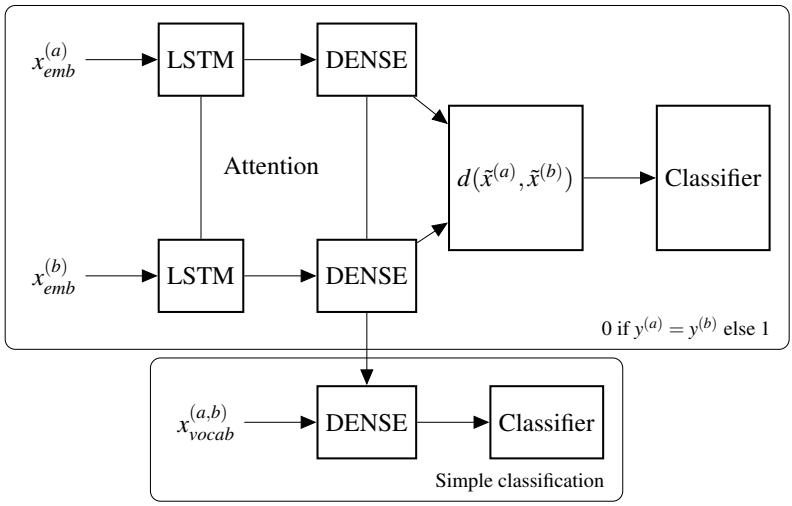


Figure 1. Proposed siamese network architecture

A siamese network composed of two symmetric sub-networks sharing the same weights but taking different inputs (sentences) is considered (see Section 4.2 for word representations). As we need to process sequences of words, the architecture relies on a bidirectional Long-Short Term Memory (LSTM) cell which takes pre-vectorized sentences as inputs ($x^{(a)}_{emb}$ and $x^{(b)}_{emb}$). In order for the network to focus on specific areas, an attention mechanism is added on top of the recurrent layer: to each part of the sentence is assigned a specific weight which denotes how important a word is, relative to other ones. These weights are then used to compute a weighted sum over the output of the LSTM, yielding a fixed 2-dimensional tensor encoding the sentence – this is defined in the literature has a many-to-one attention mechanism. Depending on the dataset and the setup, we use two different variants of this mechanism: the concatened version and the general Luong dot product attention [15] which rely on slightly different sub-networks for computation. A fully connected layer (dense) is also applied on the 2-dimensional tensor; this projection is later reused for the second classification task (bottom of Figure 1).

One-shot learning requires to compute a distance function between two samples that go through the siamese network. As we are using a binary cross entropy loss, the output is squashed using a dense layer with a sigmoid activation. Two distance functions $(d(\tilde{x}^{(a)}, \tilde{x}^{(b)})$ yield better performances given specific setups (details presented in Section 4). Those functions are applied feature-wise: the first one is based on the absolute difference between the two projections and the second one on a modified cosine distance. If the weighted sum of distances is minimal, the sigmoid outputs a 0 and both samples will share the same label.

As datasets used for this task are very small (less than 100 labelled sentences), the network is compact with small layer sizes and dropout (0.25) to prevent overfitting. The LSTM has a (16 x 2) hidden size and the attention added on top is composed of 16 units.[4]

3.2. Classification

The top part of the network can be used independently for prediction. Given a new sample, it can be compared to known and precomputed samples from the training set. The same label as the closest or top closest sentences can then be associated. Using a second classifier however is often more reliable, stable and provides better performances.

The output of the dense layer preceding the distance function can be seen as a sentence embedding with a fixed size. This representation is transferred and concatenated with a selected set of discriminant words for this task. As classification is done directly on the original dataset (no couples involved), the number of features must be small to prevent overfitting. Word selection has been done comparing frequencies between classes. We employ a simple absolute distance metric which shows the best performances in our case. As we are in a binary case (the judge accepts or rejects a claim), frequencies have been defined as follows: with f_c^w the frequency of word w in sentences from category c, the discriminant words are those maximizing the difference $|f_i^w - f_j^w|$.

4. Datasets and words representations

In this section we describe the 5 datasets used for the classification task and the experiment setup. The preprocessing pipeline and the way specialized words embeddings are trained are also presented.

4.1. Datasets

Five datasets are chosen to cover different types of claims. They have been manually annotated by lawyers who labelled the part where the judge gives its argument for the specific claim and the result associated (accept or reject). This result is not straightforward to infer as French legal language has very specific vocabulary and expressions which are mostly unknown and extremely ambiguous for nonexpert people. The datasets are balanced and relative to name change requests (600.NOM, 74 observations), unpaid debts (600.DEC, 96 observations), lawyers' liability (500.RES, 400.RES, 100 observations each) and damage and interest claims for serious injuries (300.DOM, 98 observations).

4.2. Representation

Representing words and sentences is a challenging task and has a strong impact on classifiers performance. We compare different vectorization approaches from the simple TF-IDF to state-of-the-art models like BERT [23].

We built specialized word embeddings ranging from 32 up to 128 dimensions. All of them have been trained on a large corpus composed of 670 millions tokens from (French) court decisions and written laws. As French legal texts are very sensitive to case and

[4]This number is doubled when the data augmentation strategy later introduced is applied.

punctuation (e.g. semicolons are important separators), these symbols have been kept. We estimate 3 different word embeddings: FastText [24], ELMo [25] and Flair [26]; BERT has however not been trained on our corpus as it requires massive computation power – the Bert-base multilingual cased pretrained model has been used instead.[5] On the one hand, all of them can handle out of vocabulary words and have specific characteristics: FastText and BERT use n-grams, Flair focuses on characters, while ELMo considers words and characters at the same time. On the other hand, FastText is static while all others are contextual, meaning they can change the word representation with respect to a specific context for disambiguation purpose. This is done by using recurrent layers or multi-head attention from the transformer architecture [16]. Training requires a few hours for FastText while ELMo and Flair need days to converge. As legal texts tend to have similar structures, a niche vocabulary and redundant expressions, very compact models can achieve low perplexity (< 20 for ELMo with 64 base dimensions) – this shows that French legal language is predictable.

4.3. Data augmentation

Data augmentation is a common way to deal with small datasets. Its main purpose is to artificially enrich and increase the number of observations (words) lying in the texts of the dataset. This is a challenging task in NLP as modifying one single word can drastically change the meaning of a sentence, which implies bias in the prediction over tiny samples. We investigate different approaches to see whether this technique can be done on legal language. First, words are randomly replaced by their synonyms using a thesaurus. This creates poor quality sentences as standard synonyms are not suitable for juridical specific vocabularies. Second, a random noise has been added on vectorized sentences with and without random word permutations. This does not yield any improvement, even leading to worse generalization capacities. Last, a translation tool is employed by going through several translations until returning to French language. This yields interesting new sentences *relevant* to the original ones. Augmenting data this way significantly improves the performance of the classifiers (see Section 5) allowing the models to be trained deeper while avoiding overfitting.

5. Experiment and results

We compare different approaches to find how fine tuning word embeddings and vocabulary selection can improve performances on small datasets (Figure 2). We start by investigating standard algorithms coupled with simple vectorization processes: the first one is based on TF-IDF while the second one relies on a selected vocabulary and a naive sentence embedding based on the average word representation (Table 1). We then show how fine-tuning using one-shot learning (Table 2) and data augmentation (Table 3) yield significant improvements. All results are averaged with 10-fold cross validation.

[5] https://github.com/google-research/bert

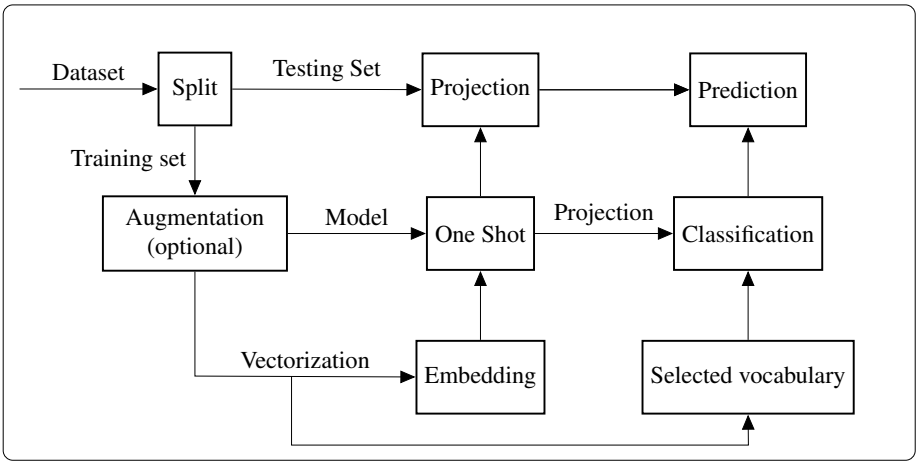


Figure 2. Train-test one-shot learning

5.1. Standard algorithms

Table 1. Comparing classification performances with different inputs

	SVM			Random Forest			Logistic		
	P	R	F	P	R	F	P	R	F
600.NOM*	0.798	0.740	0.748	0.782	0.822	**0.786**	0.743	0.740	0.728
600.NOM**	0.820	0.820	0.781	0.867	0.780	**0.813**	0.752	0.748	0.739
600.DEC*	0.669	1.000	0.788	0.747	0.872	**0.795**	0.6578	0.96	0.767
600.DEC**	1.000	0.842	0.908	0.931	1.000	**0.959**	0.889	0.934	0.902
500.RES*	0.591	0.583	0.568	0.651	0.731	0.619	0.674	0.700	**0.645**
500.RES**	0.716	0.712	**0.659**	0.623	0.733	0.636	0.639	0.546	0.570
400.RES*	0.943	0.710	0.795	0.826	0.890	**0.837**	0.810	0.882	0.828
400.RES**	0.783	0.848	0.789	0.924	0.876	**0.893**	0.709	0.820	0.736
300.DOM*	0.827	0.923	0.834	0.847	0.888	**0.854**	0.847	0.903	0.825
300.DOM**	0.963	0.916	0.931	1.000	0.925	**0.952**	1.000	0.857	0.910

* TF-IDF vectorization — Precision (P), Recall (R) and F-measure (F)

** Selected vocabulary + mean embedding

The random forest better performs on each demand category except for lawyers' liability (500.RES) for which SVM and logistic classifiers provide better F-measures. Words representations are averaged to provide a sentence embedding which is concatenated with the selected vocabulary. This yields significant gains compared with TF-IDF which

is obviously overfitting, for instance a F-measure gap of 0.164 is recorded on unpaid debts (600.DEC). TF-IDF shows poor performances for two main reasons: vocabulary is large and fixed, leading to a sparse representation; it is unable to handle variations consistently (e.g plural) unlike word embeddings. Selecting a subset of discriminant words often achieves similar performances with far less parameters and computation.

5.2. One-shot siamese recurrent network

The results of the one-shot siamese recurrent network are presented in Table 2. In this case, the mean embedding is replaced by the one-shot strategy which acts as a powerful sentence embedding model (fine-tuned weighted sum). The classifier outperforms the random forest on each demand category with the aid of ELMo. BERT shows lower performances as it has not been trained on a large legal corpus. Models are trained over embeddings with 32 dimensions, concatenated attention, and the ℓ_1 distance function.

Table 2. Comparing embeddings on one-shot classification task

	FastText			ELMo			Flair			BERT		
	P	R	F	P	R	F	P	R	F	P	R	F
600.NOM	0.842	0.810	0.818	0.867	0.830	**0.846**	0.842	0.850	0.843	0.755	0.923	0.789
600.DEC	0.945	1.000	0.967	0.975	1.000	0.986	0.986	1.000	**0.992**	0.986	0.983	0.983
500.RES	0.756	0.690	0.701	0.774	0.714	**0.734**	0.805	0.665	0.709	0.610	0.863	0.686
400.RES	0.868	0.916	0.882	0.907	0.921	**0.908**	0.828	0.901	0.854	0.921	0.843	0.872
300.DOM	1.000	1.000	**1.000**	1.000	0.983	0.990	0.983	0.983	0.982	0.963	0.983	0.971

These overall improvements come from the fact that we now rely on a model able to take advantage of the sequential aspect and long-short term dependencies (using LSTM and attention). This is fundamental as French legal language tends to be extremely ambiguous with double negatives, references, implicit reasoning...

Table 3. Best overall models with and without augmentation

	With augmentation			Without augmentation			Overall gains	
	P	R	F	P	R	F	ΔF*	ΔF**
600.NOM	0.870	0.960	**0.880**	0.867	0.830	0.846	+0.034	+0.094
600.DEC	0.986	1.000	**0.992**	0.986	1.000	**0.992**	+0.000	+0.197
500.RES	0.794	0.903	**0.817**	0.774	0.714	0.734	+0.083	+0.172
400.RES	0.918	0.969	**0.940**	0.907	0.921	0.908	+0.032	+0.107
300.DOM	1.000	1.000	**1.000**	1.000	1.000	**1.000**	+0.000	+0.146
Average							+0.030	+0.142

* F-measure difference with and without augmentation

** F-measure difference with augmentation and naive TF-IDF

Finally, Table 3 presents the contribution of the text augmentation. As we have access to more examples, we can deepen our model architecture by increasing layer sizes (doubled) and using larger word embeddings (64 dimensions). Coupled with Luong attention and a cosine distance function, generalization is better given the extra flexibility yields by more parameters. Further increasing embeddings size does not provide additional gain.

5.3. Attention for interpretability

Vocabulary selection provides a way to extract discriminant words but fails to take into account less frequent expressions or variations (e.g plural). Attention is a soft selection mechanism linking each input to a specific score given the context. As we feed words, we can find out which part of the sentence has high weights and where the network is focusing. The output of attention is a weighted sum over the temporal dimension, this leads to a more accurate and fine grained sentence embedding compared with a simple word average (see Table 1 and 2). Adding this mechanism also helps dealing with long term dependencies as it is insensitive to sequence length, even LSTM cells can suffer and forget large information parts from long sequences (> 30 words).

6. Conclusion

The one-shot siamese recurrent network proposed in this paper outperforms traditional algorithms of the literature for the purpose of predicting decisions outcome given highly ambiguous judge arguments. The results obtained with attention mechanisms as well as data augmentation seem to be promising; they illustrate how the Legal domain could benefit from advanced deep learning techniques suited for contexts in which only small labelled datasets are available. This work also opens the way on the employ of recent network architectures in jurimetrics such as adversarial networks, which provide some good potential to find discriminant words and expressions.

References

[1] T. Gonçalves and P. Quaresma, Is Linguistic Information Relevant for the Classification of Legal Texts?, in: *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, ACM, New York, NY, USA, 2005, pp. 168–176. ISBN ISBN 1-59593-081-7.

[2] S. Brüninghaus and K.D. Ashley, Toward Adding Knowledge to Learning Algorithms for Indexing Legal Cases, in: *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, ICAIL '99, ACM, New York, NY, USA, 1999, pp. 9–17. ISBN ISBN 1-58113-165-8.

[3] R. Bevan, A. Torrisi, D. Bollegala, K. Atkinson and F. Coenen, Efficient and Effective Case Reject-Accept Filtering: A Study Using Machine Learning, in: *Proc. of the 31st International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2018, pp. 171–175.

[4] R. Slingerland, A. Boer and R. Winkels, Analysing the Impact of Legal Change Through Case Classification, in: *Proc. of the 31st International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2018, pp. 121–1230.

[5] B. Waltl, J. Muhr, I. Glaser, E.S. Georg Bonczek and F. Matthes, Classifying Legal Norms with Active Machine Learning, in: *Proc. of the 30st International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2017, pp. 11–20.

[6] E. de Maat, K. Krabben and R. Winkels, Machine learning versus knowledge based classification of legal texts, in: *Proc. of the 23th International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2010, pp. 87–96.

[7] I. Glaser, E. Scepankova and F. Matthes, Classifying Semantic Types of Legal Sentences: Portability of Machine Learning Models, in: *Proc. of the 31st International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2018, pp. 121–1230.

[8] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Palka, G. Sartor and P. Torroni, Automated Processing of Privacy Policies Under the EU General Data Protection Regulation, in: *Proc. of the 31st International Conference on Legal Knoweledge and Information Systems (JURIX)*, 2018, pp. 51–60.

[9] Octavia-Maria, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu and J. van Genabith, Exploring the Use of Text Classification in the Legal Domain, in: *Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL)*, London, United Kingdom, 2017.

[10] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: *European conference on machine learning*, Springer, 1998, pp. 137–142.

[11] E. de Maat and R. Winkels, A Next Step Towards Automated Modelling of Sources of Law, in: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, ACM, New York, NY, USA, 2009, pp. 31–39. ISBN ISBN 978-1-60558-597-0.

[12] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *nature* **521**(7553) (2015), 436.

[13] I. Angelidis, I. Chalkidis and M. Koubarakis, Named Entity Recognition, Linking and Generation for Greek Legislation, in: *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018.*, 2018, pp. 1–10.

[14] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, *ArXiv* **1409** (2014).

[15] M. Luong, H. Pham and C.D. Manning, Effective Approaches to Attention-based Neural Machine Translation, *CoRR* **abs/1508.04025** (2015). http://arxiv.org/abs/1508.04025.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser and I. Polosukhin, Attention is All you Need, in: *Advances in Neural Information Processing Systems 30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, Curran Associates, Inc., 2017, pp. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[17] L. Fei-Fei, R. Fergus and P. Perona, One-Shot Learning of Object Categories, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(4) (2006), 594–611.

[18] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra and T.P. Lillicrap, One-shot Learning with Memory-Augmented Neural Networks, *CoRR* **abs/1605.06065** (2016). http://arxiv.org/abs/1605.06065.

[19] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra and T. Lillicrap, Meta-learning with Memory-augmented Neural Networks, in: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 2016, pp. 1842–1850. http://dl.acm.org/citation.cfm?id=3045390.3045585.

[20] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang and B. Zhou, Diverse Few-Shot Text Classification with Multiple Metrics, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1206–1215. https://www.aclweb.org/anthology/N18-1109.

[21] A.K. Lampinen and J.L. McClelland, One-shot and few-shot learning of word embeddings, *CoRR* **abs/1710.10280** (2017). http://arxiv.org/abs/1710.10280.

[22] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger and R. Shah, Signature Verification using a "Siamese" Time Delay Neural Network, in: *Advances in Neural Information Processing Systems 6*, J.D. Cowan, G. Tesauro and J. Alspector, eds, Morgan-Kaufmann, 1994, pp. 737–744. http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf.

[23] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR* **abs/1810.04805** (2018). http://arxiv.org/abs/1810.04805.

[24] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching Word Vectors with Subword Information, *CoRR* **abs/1607.04606** (2016). http://arxiv.org/abs/1607.04606.

[25] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, *CoRR* **abs/1802.05365** (2018). http://arxiv.org/abs/1802.05365.

[26] A. Akbik, D. Blythe and R. Vollgraf, Contextual String Embeddings for Sequence Labeling, in: *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1638–1649. https://www.aclweb.org/anthology/C18-1139.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190304

Governmental Transparency in the Era of Artificial Intelligence

Tom M. van ENGERS [a] and Dennis M. de VRIES [b]
[a] *Leibniz Center for Law, University of Amsterdam*
[b] *Informatics Institute, University of Amsterdam*

Abstract. In the last years governments started to adapt new types of Artificial Intelligence (AI), particularly sub-symbolic data-driven AI, after having used more traditional types of AI since the mid-eighties of past century. The models generated by such sub-symbolic AI technologies, such as machine learning and deep learning are generally hard to understand, even by AI-experts. In many use contexts it is essential though that organisations that apply AI in their decision-making processes produce decisions that are explainable, transparent and comply with the rules set by law. This study is focused on the current developments of AI within governments and it aims to provide citizens with a good motivation of (partly) automated decisions. For this study a framework to assess the quality of explanations of legal decisions by public administrations was developed. It was found that communication with the citizen can be improved by providing a more interactive way to explain those decisions. Citizens could be offered more insights into the specific components of the decision made, the calculations applied and sources of law that contain the rules underlying the decision-making process.

Keywords. Artificial Intelligence, XAI, Transparency, Explanations, Government

1. Introduction

All over the world, governments have started to adopt new Artificial Intelligence (AI) technologies to improve the efficiency and effectiveness of public administration. Automated Decision-making Systems (ADS), for example, can help governmental agencies with various tasks such as deciding on tax assessment and student finance. In these application domains, the citizens' stakes are high. Therefore, it is of great importance that those (partly) automated decision systems are transparent on their reasoning mechanisms and carefully explain their decisions. This study focus on the improvement of explanations of governmental agencies' communications regarding (partly) automated decisions.

Over the years, a substantial number of studies have been published on opening the black boxes of artificial intelligence [1,2]. Only a few studies suggest procedures on how decisions made by artificial intelligence should be explained in a proper manner [3,4]. Besides some studies of prestigious consulting firms, academic research on how to improve explainability and transparency of automated decision-making systems in governments is laying back [5,6,4].

1.1. Governments adopting Artificial Intelligence

AI-based systems have been used by governments for decision-making purposes since the mid-eighties of the 20th century. Most of these systems were and still are rule-based systems, with the 'rules' elicited from (legal) experts [7,8]. The primary purpose for the government to invest in AI systems is to provide better services and improve the effectiveness and efficiency of public administration [9], e.g. in application domains such as optimising traffic flows [10], tax assessment [11], assessing visa applications [12] and crime prevention [13].

Many of those AI systems are used for decision-support and use a rule-based reasoning mechanism using determined rules to come to a specific decision [14,15,16]. With the increase of computer power, sheer unlimited data availability, and the boost of the internet, new AI-technologies have emerged and become popular. Particularly data-driven, sub-symbolic AI technologies, that are known under various names, such as machine learning, deep learning, and neural nets, became popular again in the 21st century [17]. Since the end of the nineteen-nineties, e.g. machine vision methods were used for various pattern recognition task including that of handwritten addresses from envelopes [18]. Contrary to symbolic AI that is typically connected to deductive approaches, sub-symbolic AI is typically connected to inductive approaches. This focuses on learning systematic patterns from the data, and then apply those learned patterns on new input determining the appropriate output [19]. The use of AI in fields where the stakes are high, however come with some worries.

1.2. Challenges of AI

Ever since the introduction of AI-technologies people have feared the lack of human touch and empathy, the lack of transparency and unfairness when smart AI-components replace the human in the loop [13].

The COMPAS system developed for predicting the likelihood of recidivism of criminals for example became infamous for its bias against Afro-Americans [20]. Such bias against a specific group within society could easily lead to more segregation and then decreasing opportunities for that specific group and as a result produce a self-fulfilling prophecy [21]. In order to be able to trust organisations in taking (legally) justified decisions, these decisions when produced by AI applications need to be explained and argued for in such a way that the persons subjected to those decisions at least understand what the decision is based upon.

The main challenge that is addressed in this study, is providing insight into the reasoning mechanisms of AI-algorithms for citizens. This is needed in order to check their correctness, fairness, normative compliance and sensitivity to potential biases in their judgements.

1.3. A Renewed Interest in Explainable AI

Data-driven AI-technologies that 'learn' from data, are vulnerable for bias and the models induced from the data are generally hard to understand even for experts. This is even getting worse if the AI-algorithms keep 'learning', i.e. adapting their models, while being used. Because of the lack of transparency it is hard to 'trust' those AI algorithms hence

the recent demand 'Responsible AI', a term that includes explainable AI (XAI) and fairness, but is somewhat ambiguous as responsibility could refer both to the AI-technology itself as well to the developers and organisations exploiting these technologies. Holding AI responsible for anything, i.e. attributing some kind of personality to it, would bring us back to dark ages, so let's keep the human stakeholders responsible, like we do with all other artefacts! The call for XAI has become louder after a few scandals, and it is needless to say that specifically governmental agencies that deploy AI to support their tasks have to meet the traditional government requirements for explainability, transparency, accountability and auditability [6].

In order to try to protect some essential social fundamental values, The Dutch Council of State (advisory body to the government) published a report on the influence of new technologies on constitutional relations [22]. The Council advises the government to pay closer attention to the motivation of their automated decisions. They demand that it should be clear which decision-rules (algorithms) and data the governmental authority used for a specific decision. Furthermore, it should be made clear which data is taken from other governmental authorities. Explainability in Europe further pushed by the General Data Protection Regulation (GDPR) [23] that is applicable since May 25th 2018 in all European member states. The GDPR includes Article 22 on *'Automated individual decision-making, including profiling'* forcing organisations to be transparent about the decision-making process of their algorithms.

2. Literature Review

As stated in the previous section the need for explainable AI is not an entirely new topic; it has been addressed in many reports and academic papers and is discussed at plenty of conferences such as those of ACM's CHI community [24,25,26]. The increased popularity of sub-symbolic AI has just put the topic back on the agenda again.

2.1. Why Explanations Matter

One key part of XAI is the explanation itself, The Oxford English Dictionary defines EXPLANATION as: 1) *'A statement or account that makes something clear'* and 2) *'A reason or justification given for an action or belief'* [27]. Therefore, an explanation mainly aims to answer the *how* and *why* questions, which can be useful to clarify or justify the behaviour of an AI agent respectively [28]. Within our daily lives, explanations are used by humans to share information and in order to better understand each other. Therefore, explanations lead to better acceptations about specific statements [29]. Over the years, studies from various disciplines suggest that providing explanations on the mechanisms of AI systems improve the acceptance of the user in regards to the decisions, conclusions and recommendations of those systems [30,28,31,24,32]. As a result, systems that provide better explanations on their reasoning will improve the acceptance by citizens in the outcome of those systems. Other studies suggest that explanations from AI systems help to acquire or maintain trust from the user in the accuracy of those systems [33,18,34,26,35].

2.2. Explaining Good Explanation

Research into explanations has a long history. Early examples of research in this subject include topics such as logic, causality and human discourse [36,37]. Related work can be found in various areas such as philosophy and psychology. Based on earlier studies, an evaluative framework that enables to evaluate the quality of XAI was developed. In literature, several criteria have been described that can be used to determine the satisfaction of an explanation. The framework presented in this paper includes those criteria that are most frequently mentioned and extensively discussed in the field of cognitive sciences and AI literature. Below we'll present six primary quality criteria for explanations and references to preliminary research on these criteria:

The first quality criterion for explanation is called **EXTERNAL COHERENCE** [38]. Some researchers suggest that the likelihood of acceptance of a decision increases when the explanation is consistent with one's former beliefs [39]. This means that explanations should be compatible with what the reader already knows in the specific context at hand [40].

The second quality criterion is **INTERNAL COHERENCE**. This concept points out the sense of how good the several elements of an explanation fit together [40]. There should be a logical relation between propositions to improve the completeness of the explanation and improve the perceived understanding [41,38].

The third quality criterion is **SIMPLICITY**. Two studies tested the theory of Thagard on Explanatory Coherence [38] and found that people preferred explanations that invoke fewer causes [42,43].

The fourth quality criterion is **ARTICULATION**. One particular study presents several linguistic markers that examine clear articulation of a letter [40]. One of the three elements is the number of words used in the explanation. Another one is the average word length of the statement. The median word frequency of the text can also be used as an indicator [40].

The fifth quality criterion is **CONTRASTIVENESS**. This criterion expresses the clarity of the arguments that explain why event P happened rather than event Q [39,44]. This specific factor also emphasises questions such as what would happen when a particular condition in the process is changed [45].

Finally, some research mentions that the user's satisfaction with an explanation might increase when the possibility for **INTERACTION** between the explainer and explainee is provided [46]. What is needed for an explanation also depends on what the explainee already knows and specifically; still wants to know [47]. This criterion proposes new opportunities in the field of Human-Computer Interaction (HCI) [39]. By providing interactive dialogue, the satisfaction of the user might increase.

Several criteria for the layout of a letter (used fonts, use of color, etc.) might influence the receiver as well. This research was scoped to mainly focuses on the structure of a letter and therefore the layout criteria have been left out.

The evaluative framework described here will be used to analyse a specific ADS-generated governmental decision later. Thereafter, the framework will be used to create an alternative presentation format for that decision with the main goal to enhance the citizen's satisfaction and acceptance of the decision.

3. Case Study on Student Loans in the Netherlands

Ideally, this study would focus on an AI application that is representative of approaches that raised the issue of explainability, in other words deep-learning or similar sub-symbolic technologies. However, The Council of State said that with the less complex technologies, problems still emerge. The absence of sub-symbolic tools in administrative practice means that a decision was made to look into popular, rule-based, symbolic AI.

The case selected is the application that is used to decide on student loans, deployed by the Education Executive Agency (referred to as DUO in Dutch), an administrative agency that falls under the responsibility of the Dutch Ministry of Education. The ADS for deciding on student loans uses symbolic AI. More specifically, it is a rule-based system that contains different rules that are evaluated when deciding on the entitlement of students to financial support.

3.1. Designing a Conceptual Disposal

After analysing the original letter from DUO, an improved version of the presentation format was developed using the principles described in the framework from section 2.2. This conceptual online letter was set up with the main goal of providing better insight into the reasoning mechanisms of the algorithm, the data used to make the decision, and the presentation of the decision in a clearer way. The six criteria for explanation, as defined earlier, were used to improve the letter in the following ways. First, the letter contains a section that informs the receiver about the change in address that affects the student's monthly loan (external coherence criterion). The order of messages, one per section, was reorganised to give a better relation between the various parts of the letter (internal coherence criterion). Different from the original letter, the conceptual letter explains the reasoning that led to the decision. As in the original letter, only one cause (change in address) was presented to explain the change in the loan to the student (simplicity criterion). The number of words in the letter was reduced for the conceptual disposal (articulation criterion). Furthermore, the student's old situation and new situation were presented together in a contrastive table (contrastiveness criterion). By offering the user the possibility to learn more about the decision via hyperlinks to more elaborated information, the student's understanding of the situation might increase as well (interaction criterion).

4. Methodology

The case selected is a symbolic AI decision-making tool used for deciding on student loans provided by DUO. The original and conceptual versions of the presentation format were subjected to an A/B test. The A/B test was included in an online survey using Qualtrics. Half of the subjects received the survey that included version A, the other half version B. Besides questions about the explainability of the presented version, the survey included questions that were used to measure the students' attitudes towards the use of ADS in the Dutch government.

Chat service WhatsApp was used for contacting around 100 students, being the target audience for the application studied. Some of the students forwarded the questionnaire to other students, resulting in 133 students who completed the survey.

4.1. Hypotheses

The case study was used to test the following hypotheses:

- There is no relation between one's trust in government and trust in computer systems within the government.
- The citizen's support for the deployment of AI by the government does not vary by case.
- The presentation format of a governmental decision will have no influence on the citizen's perceived satisfaction about that decision.
- The presentation format of a governmental decision will have no influence on the chance a citizen will accept that decision.
- The presentation format of a governmental decision will have no influence on the citizen's urge to object to or appeal that decision.

4.2. Outline of the Survey

First, the subjects were shown an introductory text that explained the current situation of AI use by the Dutch government and the purpose of the research.

Thereafter, a five-point Likert scale, ranging from strongly disagree (1) to strongly agree (5), was used to determine the participants' attitudes. The participants were asked to rate how strongly they agreed with specific statements on the use of symbolic AI in government.

Subsequently, participants were asked to evaluate a disposal of an automated decision from DUO. One original disposal was obtained from the agency itself; the other one was a more interactive disposal that was created specifically for this study and included all factors that, according to theory, would enhance explainability. The participants were randomly assigned to one of the two versions and were then asked questions to survey their satisfaction with the disposal.

Before distribution, the survey was checked by three individuals to ensure understandability.

4.3. Participants

For finding subjects for the A/B test and the survey, a convenience sample was taken. The sample selection resulted in 133 subjects responding and completing the survey. The students recruited were enrolled in various universities and colleges in the Netherlands. From the total group, 60 students (45.1%) were female, and 73 students (54.9%) were male. All the participants were aged between 18 and 30, with an average age of 23.46 years (SD = 1.78). Most of the students were currently enrolled in an academic master's programme (49.6%), followed by academic bachelor students (28.6%), and 14 respondents were enrolled in a bachelor's programme at a university of applied sciences (10.5%). Additionally, there was one student enrolled in an applied sciences master's programme (0.8%) and one student from college (0.8%). Thirteen participants noted that they were currently not in school (9.8%). The next section discusses the data preparation and analysis, and the results are then discussed.

5. Analysis and Main Findings

The 133 students involved in the A/B test were split into one group of 68 persons who received the survey on the original letter and 65 who received the survey on the conceptual letter. A check for sampling independence between the two groups was then performed. No difference in gender ($\chi^2(1) = 0.013$, p = .910), age (t(131) = 0.662, p = .509) or education level between the groups ($\chi^2(5) = 5.161$, p = .397) was found.

Our first hypothesis was rejected as we found a correlation between the *trust in government* and the *trust in computer systems within government* (F(1,131) = 14.137, p < .0005, R^2 = .097, b = 0.333, t(131) = 3.760, p < .0005).

The respondents were asked for what tasks they support the deployment of computer systems for governmental use. Students stated that they support the use of computer systems for the optimisation of traffic flows (91.7%), the calculation of student finance (84.2%) and the calculation of tax assessment (80.5%). Only 34.6% of the students have the opinion that automated systems should be used for the rejection or grant of visas. Cochran's Q shows that agreement ratios for these four purposes are not identical (Cochran's Q(3) = 144.437, p < .0005). Post-hoc McNemar tests with Bonferroni correction showed that the students' support for automated systems for visa decisions is significantly lower than the three other variables. Therefore, we colclude that the students' support for the deployment of AI in government varies by use.

Since the dependent variables do not follow a normal distribution in either condition (*Original Disposal*: Shapiro-Wilk W(68) = .941, p = .003, *Conceptual Disposal*: Shapiro-Wilk W(65) = .936, p = .002), the t-test cannot be used. Therefore, a non-parametric Mann-Whitney U test is preferred to analyse the difference between the clarity of the two letters. One of the major findings of this study is that students are more satisfied with the conceptual disposal than the original disposal (U = 1082.5, z = 5.112, p < .0005). Furthermore, respondents also agreed with the statement *'I prefer an interactive (clickable) letter.'* With an average score of 3.80 on the five-point Likert scale, this was also significantly higher than the neutral value of 3.0 on the five-point Likert scale (t(132) = 10.805, p < .0005). Therefore, this study finds that students will be more satisfied with a more interactive letter than the original letter from DUO.

Furthermore, it is shown that the letter type (original or conceptual) has a significant influence on the acceptance of the decision. Respondents agree significantly more easily with the statement *'The content of the letter convinces me to agree with the decision.'* when receiving the conceptual letter (U = 1550, z = 3.331, p = .001). Therefore, the letter type, the presentation format of the governmental decision, has a significant influence on the acceptance of the decision by the student.

No significant difference between the two letter conditions was found in the urge to object to or appeal the decision (U = 1967, z = 1.186, p = .235). However, the explanation in the conceptual letter was found to be more beneficial for the support and argumentation of a potential objection or appeal (U = 1577, z = 2.979, p = .003). Also studied was the way in which the students agreed with the statement that a good explanation of the decision would help to reduce the chance of objection or appeal. With an average score of 4.02 on the five-point Likert scale, this is significantly higher than neutral, which has the value 3.0 (t(132) = 13.319, p < .0005). Therefore, it can only be stated that the citizen's willingness to object to or appeal the decision might only be reduced by offering a better explanation.

6. Conclusion

The adoption of new AI technologies by governments bring challenges such as the potential bias in the algorithms exploited and, certainly in case of data-driven, sub-symbolic AI approaches, the general lack of explainability of the decision-making processes supported by those algorithms. As a result, a renewed interest in XAI emerged. Equally important is the transformation in the way governments interact with their citizens thriving for higher effectivity and costs reduction leading to AI-usage in a wide variety of previously manually operated tasks. This study aims to contribute to this growing area of research by exploring the principles of explanations, and it offers a framework that strives to assess the quality of a given explanation. When analysing a Dutch disposal, it seems that the government is already doing a great job with a bright, interactive and straightforward letter. However, the way the government currently interacts with the citizens can be significantly improved.

In order to achieve a better understanding of the citizen, a digital letter should be compatible with existing knowledge of the citizen; the parts of the letter have to fit together and use as few causes possible; and the letter should be written clearly, provide contrastive information and offer the opportunity to interact.

Several conclusions can be drawn from the quantitative study. A significant relation between one's trust in the government and the trust in computer systems used by that government was found. The citizen's support for the deployment of AI by the government varies per use or case, and more research is necessary to better understand why. This research demonstrates that students will be more satisfied with a more interactive letter than the current original letter from DUO. Furthermore, it can be concluded that a clearer explanation of the decision will lead to a greater likelihood of accepting that decision, which also confirms the previous studies as discussed in section 2.1. Therefore, governments can increase the acceptance rate of citizens by improving the clarity of their explanations, and this can create a new field of interest in *explanation optimisation*. Lastly, the study found that letter type has no significant influence on the urge to object to or appeal the governmental decision. On the contrary, a good explanation of an automated governmental decision was found to help to reduce the citizen's willingness to object to or appeal that decision.

The study also reconfirms that while investments in AI supporting various tasks of public administrations are merely driven by the need for improving efficiency and effectiveness. It is important to keep in mind that explainability, transparency, accountability and auditability are essential to governmental processes.

7. Discussion

There are several limitations that need to be addressed for this study. First, this study mainly focuses on the adoption of rule-based AI-systems within the Dutch government. Data-driven, sub-symbolic AI technologies have become more popular but have even larger problems with explainability and fairness. At this moment very few governmental agencies within the Netherlands make use of data-driven sub-symbolic AI-technologies for their decision-making. Governmental agencies such as the Dutch Tax and Customs Administration (De Belastingdienst) stated that they were using sub-symbolic AI for var-

ious fields such as the prediction of fraud, and other agencies are either exploiting or considering the use of such technologies for similar purposes. This authority however did not want to provide materials on their reasoning mechanisms for this research because they were perceived to be confidential (intended lack of transparency). Therefore, a decision was made to collaborate with DUO, which provided materials on the reasoning mechanisms of their algorithms. The further adoption of data-driven AI-technologies would only raise the importance of XAI. Future studies in this field should also include such data-driven AI-technologies, as they are the most problematic in terms of explainability, fairness and transparency.

Acknowledgments. We would like to thank the Canadian Research Council sponsoring the ACT Project.

References

[1] W. Samek, T. Wiegand and K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, *arXiv preprint arXiv:1708.08296* (2017).
[2] D. Gunning, Explainable Artificial Intelligence, *Defense Advanced Research Projects Agency (DARPA)* (2017).
[3] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen and K.-R. Muller, How to explain individual classification decisions, *Journal of Machine Learning Research* **11**(Jun) (2010), 1803–1831.
[4] M. van Kempen, Motivering van automatisch genomen besluiten, *Knowbility* (2019).
[5] A. Dhasarathy, S. Jain and N. Khan, When governments turn to AI: Algorithms, trade-offs, and trust, *McKinsey&Company* (2019). https://www.mckinsey.com/industries/public-sector/our-insights/when-governments-turn-to-ai-algorithms-trade-offs-and-trust.
[6] M. Carrasco, S. Mills, A. Whybrew and A. Jura, The Citizens Perspective on the Use of AI in Government, *Boston Consulting Group* (2019).
[7] J.C. Giarratano and G. Riley, *Expert Systems*, PWS Publishing Co., 1998. ISBN ISBN 0534950531.
[8] AINED, *AI voor Nederland: vergroten, versnellen en verbinden*, 2018.
[9] V. Homburg, *Understanding e-government: Information systems in public administration*, Routledge, 2008. ISBN ISBN 1134085028.
[10] Y. Lv, Y. Duan, W. Kang, Z. Li and F.-Y. Wang, Traffic flow prediction with big data: a deep learning approach, *IEEE Transactions on Intelligent Transportation Systems* **16**(2) (2014), 865–873.
[11] B. Corydon, V. Ganesan, M. Lundqvist, E. Dudley, D.-Y. Lin, M. Mancini and J. Ng, Transforming Government Through Digitization, *McKinsey & Company* (2016). https://www.mckinsey.com/˜/media/McKinsey/Industries/Public Sector/Our Insights/Transforming government through digitization/Transforming-government-through-digitization.ashx.
[12] Dutch Digital Government, NL DIGIbeter, *Digital Government Agenda* (2018). https://www.nldigitalgovernment.nl/document/digital-government-agenda-2/.
[13] M. Reid, Rethinking the Fourth Amendment in the Age of Supercomputers, Artificial Intelligence, and Robots, *West Virginia Law Review* **119** (2017), 863–890.
[14] R.V. Schuwer, *Het nut van kennissystemen*, 1993. doi:10.6100/IR394707.
[15] J. Haugeland, *Artificial Intelligence: The Very Idea*, MIT Press, 1985. ISBN ISBN 0262580950.
[16] P. Smolensky, Connectionist AI, symbolic AI, and the brain, *Artificial Intelligence Review* **1**(2) (1987), 95–109. doi:10.1007/BF00130011.
[17] H. Lieberman, Symbolic vs. Subsymbolic AI, *MIT Media Lab* (2016). http://futureai.media.mit.edu/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx.pdf.
[18] H. Mehr, Artificial Intelligence for Citizen Services and Government, *Harvard Ash Center Technology & Democracy* (2017). https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.
[19] T.D. Kelley, Symbolic and Sub-Symbolic Representations in Computational Models of Human Cognition: What Can be Learned from Biology?, *Theory & Psychology* **13**(6) (2003), 847–860.
[20] J. Angwin, J. Larson, S. Mattu and L. Kirchner, Machine Bias, *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[21] G. Sileno, A. Boer and T. van Engers, The Role of Normware in Trustworthy and Explainable AI, 2018. http://arxiv.org/abs/1812.02471.
[22] Raad van State, Ongevraagd advies over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen, *Kamerstukken II 2017/18, 26643, nr. 557* (2018). https://www.raadvanstate.nl/adviezen/zoeken-in-adviezen/tekst-advies.html?id=13065.
[23] European Union, Regulation 2016/679: General Data Protection Regulation, *Official Journal of the European Communities* (2016), 1–88. ISBN 9251032718. doi:http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf.
[24] E.H. Shortliffe and B.G. Buchanan, *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, Addison-Wesley Publishing Company, 1985. ISBN ISBN 0201101726.
[25] E. Horvitz, D. Heckerman, B. Nathwani and L. Fagan, The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning, in: *Proceedings of Medinfo, Washington, DC*, 1986, pp. 27–31.
[26] J.E. Mercado, M.A. Rupp, J.Y.C. Chen, M.J. Barnes, D. Barber and K. Procci, Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management, *Human Factors* **58**(3) (2016), 401–415. doi:10.1177/0018720815621206.
[27] English Oxford Dictionaries, Definition of 'explanation' in English, 2019. https://en.oxforddictionaries.com/definition/explanation.
[28] R. Neches, W. Swartout and J. Moore, Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development, *IEEE Transactions on Software Engineering* **SE-11**(11) (1985), 1337–1351. doi:10.1109/TSE.1985.231882.
[29] T.R. Roth-Berghofer and J. Cassens, Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems, in: *International Conference on Case-Based Reasoning*, Springer, 2005, pp. 451–464.
[30] J.L. Herlocker, J.A. Konstan and J. Riedl, Explaining collaborative filtering recommendations, in: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, 2000, pp. 241–250. ISBN ISBN 1581132220.
[31] F. Sørmo, J. Cassens and A. Aamodt, Explanation in case-based reasoning-perspectives and goals, *Artificial Intelligence Review* **24**(2) (2005), 109–143.
[32] R. Ye and P. Johnson, The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice, *MIS Quarterly* **19**(2) (1995), 157–172. doi:10.2307/249686. http://www.jstor.org/stable/249686.
[33] D. Doran, S. Schulz and T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, *arXiv preprint arXiv:1710.00794* (2017).
[34] W. Pieters, Explanation and trust: what to tell the user in security and AI?, *Ethics and information technology* **13**(1) (2011), 53–64.
[35] J.Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia and M. Barnes, Situation awareness-based agent transparency, Technical Report, 2014.
[36] A. Falcon, Aristotle on Causality, in: *Stanford Encyclopedia of Philosophy*, 2008.
[37] S.E. Toulmin, The Uses of Argument, *Cambridge University Press* (1958). ISBN 0521827485.
[38] P. Thagard, Explanatory coherence, *Behavioral and brain sciences* **12**(3) (1989), 435–467.
[39] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* (2018).
[40] J.C. Zemla, S. Sloman, C. Bechlivanidis and D.A. Lagnado, Evaluating everyday explanations, *Psychonomic Bulletin and Review* **24**(5) (2017), 1488–1500. doi:10.3758/s13423-017-1258-z.
[41] N. Pennington and R. Hastie, *The story model for juror decision making*, Cambridge University Press Cambridge, 1993. ISBN ISBN 0521419883.
[42] S.J. Read and A. Marcus-Newhall, Explanatory coherence in social explanations: A parallel distributed processing account, *Journal of Personality and Social Psychology* **65**(3) (1993), 429.
[43] T. Lombrozo, Simplicity and probability in causal explanation, *Cognitive Psychology* **55**(3) (2007), 232–257.
[44] P. Lipton, Contrastive explanation, *Royal Institute of Philosophy Supplements* **27** (1990), 247–266.
[45] D.J. Hilton, Conversational processes and causal explanation, *Psychological Bulletin* **107**(1) (1990), 65.
[46] I. Nunes and D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, *User Modeling and User-Adapted Interaction* **27**(3–5) (2017), 393–444.
[47] R. Hoffman, S. Mueller, G. Klein and J. Litman, Metrics for Explainable AI: Challenges and Prospects, *XAI Metrics* (2018).

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190305

Deep Learning for Detecting and Explaining Unfairness in Consumer Contracts

Francesca LAGIOIA [a,b,1], Federico RUGGERI [c,1], Kasper DRAZEWSKI [a], Marco LIPPI [d], Hans-Wolfgang MICKLITZ [a], Paolo TORRONI [c] and Giovanni SARTOR [a,b]

[a] *European University Institute, Law Department, Italy*
[b] *CIRSFID, Faculty of Law, Alma Mater Studiorum, University of Bologna, Italy*
[c] *DISI, University of Bologna, Italy*
[d] *DISMI, University of Modena and Reggio Emilia, Italy*

Abstract. Consumer contracts often contain unfair clauses, in apparent violation of the relevant legislation. In this paper we present a new methodology for evaluating such clauses in online Terms of Services. We expand a set of tagged documents (terms of service), with a structured corpus where unfair clauses are liked to a knowledge base of rationales for unfairness, and experiment with machine learning methods on this expanded training set. Our experimental study is based on deep neural networks that aim to combine learning and reasoning tasks, one major example being Memory Networks. Preliminary results show that this approach may not only provide reasons and explanations to the user, but also enhance the automated detection of unfair clauses.

Keywords. Unfair clause detection, deep learning, memory networks

1. Introduction

As the pool of services existing solely in cyberspace rapidly grows, the number of online contracts concluded by clicking 'I agree' on a popup window or merely by using a given service also grows. Be it for shortage of time or information overload, most of these contracts are entered into without reading. Experiments on users reading Terms of Service (ToS) and privacy policies have indicated that users take under a minute to scroll through the contract before voicing their agreement, where it should have taken them at least 15 minutes to read [1]. Left to their own devices, consumers have neither the time nor the means to analyze every online contract they enter into, not to mention manually keeping track of any changes to the contract they are bound by. The weakness of their position becomes even more obvious when contrasted with the massive processing power and sophisticated machine learning algorithms used by businesses, tasked with collecting,

[1] The first two authors contributed equally to the work and are the corresponding authors. Francesca Lagioia: francesca.lagioia@eui.eu. Federico Ruggeri: federico.ruggeri6@unibo.it

processing, aggregating, and analyzing user data for the purposes of profiling, assessing risk, predicting group behavior and supporting other forms of analysis and intervention.

As the economist Ken Galbraith noted as early as in the 1950s, effective protection of consumers (and, more generally, of weaker parties) against the overbearing power of big business requires not only legal regulation and public supervision, but also the active countervailing power of consumers and their associations [2]. As the power of big business is today largely based on advanced technologies, and increasingly on AI, an effective social response also needs the support of AI [3]. In the consumer law and data protection law domains, some AI-powered applications have been developed, for instance to detect discrimination in commercial practices; extracting, categorizing and summarizing information from privacy documents; and assisting users in processing and understanding their contents [4]. A further contribution in this direction is provided by CLAUDETTE, a user-end tool and web service which uses machine learning to identify and grade potentially unfair clauses in ToS contracts. According to the Unfair Contract Terms Directive (UCTD), a "term" or "clause" is unfair if, "contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer".[2] This definition is further specified by an Annex containing an "indicative and non-exhaustive list of the terms which may be regarded as unfair" (art. 3.3) and by over 50 ECJ decisions [5].

CLAUDETTE was trained on a corpus of 50 terms of service contracts. These documents were annotated by lawyers who identified potentially unfair clauses and classified them depending on the category of unfairness. Even with this small dataset, the system was already able to achieve an average accuracy of around 80 percent in identifying potentially unfair clauses when tested on new documents [6]. The training set was later extended to 100 documents, which enabled an improvement in precision.

We are now working to enable CLAUDETTE to deal with rationales (reasons why a clause is considered unfair), for two parallel purposes: to improve its performance in detecting and classifying unfair clauses and to provide legal reasons why a clause is classified in a certain way.

The tasks of linking unfair clauses and rationales is a challenging one since the distinction between unfair and fair instances of behaviour or rules is not completely theorized. Human analysts and decision makers usually rely on their intuition, trained on their experience with relevant examples. However, humans are also able to provide explanations for their intuitions of unfairness, appealing to standards, rules and principles, possibly expressed by cases, and most significantly by judicial precedents. This capacity is usually lacking in most automated classifiers [7] available today, though a number of projects aim to improve the interpretability and the explainability of AI systems [8, 9]. Rationales are important in providing transparency and explainability, but may also play a role in learning. Contrary to the usual assumption of a conflict between performance and explanability, we will show that in some cases the acquaintance with explanations can improve the performance of a classifier. This paper follows and combines the results of our earlier work. In particular, we present a new small structured corpus, consisting of a knowledge base of rationales for the legal qualification of unfairness, used as a support for reasoning; some experimental results obtained by applying a new deep neural network model; and the extension of the classification task to a more informative classification of such clauses, now supported by forms of reasoning on context.

[2] See the Council Directive 93/13/EEC on Unfair Terms in Consumer Contracts, art. 3 (1).

2. Methodology

Legal experts can detect unfair clauses by relying on multiple sources of knowledge, such as the applicable legal regulations, the relevant judicial cases, their trained common sense. They also use these sources also for generating rationales (explanations). While a system aimed at recognizing unfair clauses cannot be expected to reason like a lawyer, it should however be expected to provide results that match the assessments that a trained lawyer would give after carefully reading the document containing such clauses.

In CLAUDETTE we have adopted a supervised machine learning approach, based on a training set of documents annotated by domain experts [6]. The system compares clauses classified as fair or unfair, and, based on such a comparison, it develops its implicit concept of unfairness. Such an approach has delivered very encouraging results, as noted above. However, the legal knowledge used by the system is restricted to the annotations (category and unfairness level) provided by the experts, which does not directly point to the rationales behind the annotations.

The aim of this work is to study whether the introduction of explicit domain knowledge, in particular a KB of rationales, can further improve the performance of our system, by enabling it to exploit rationales for unfairness in a forward-looking way. This corresponds to the idea that human lawyers use rationales not only to provide explanations for intuitions of unfairness, but also to guide such intuitions, pointing to general features relevant to unfairness, that may be shared by other similar clauses.

To provide a computable model for the forward-looking use of rationales we rely on a particular category of deep learning models, denoted as memory augmented neural networks (MANNs) [10, 11]. MANNs are a type of a recurrent neural network (RNN) that introduces an external memory block as a support for reasoning. Given an input, the model checks whether the memory contains some slot that is related to that input (e.g., through a similarity measure). Subsequently, the memory content is extracted and coupled with the given input to accomplish the classification task. Formally, we can distinguish between two phases: (i) the *memory addressing* sub-process, where the network computes some representation of the input to operate with the memory; (ii) the *reasoning* sub-process, where the new content is distilled for the resolution of the task. In our domain, the first step consists in comparing the clause to be evaluated with the relevant rationales in the memory, whereas the second step is to use the rationales to assess the category and the level of unfairness of the clause.

MANNs have been widely used for complex tasks where reasoning about the context of the given inputs plays a key role: question answering [12, 13, 14, 15], sentiment analysis [16], reading comprehension [17, 18, 19], graph analysis and navigation [20, 21]. In this paper we initially explore a simple variant [12] of the general concept of MANNs, named end-to-end memory network, since all the core operations (memory addressing and reasoning) are differentiable, and thus the model can be trained just like any other deep network. In the context of consumer contracts, we define the task of unfair clause detection as a binary classification task [6], where a given input clause can be labelled as either fair or (potentially) unfair.

The knowledge base stored in the memory consists of a fixed collection of possible rationales for unfairness. These rationales were provided by legal experts, based on their experience or on the case law (see Section 3), and linked by the same experts to the clauses they apply to. When analyzing an input clause, the system accesses the knowl-

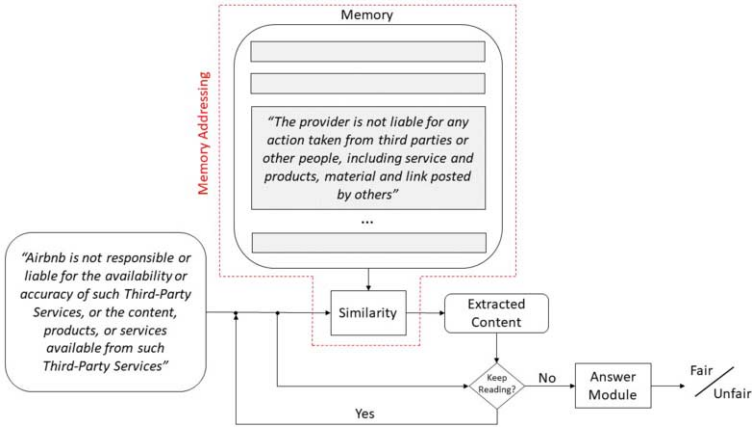


Figure 1. Architecture of the proposed model for unfairness detection. Each given input clause to be classified is firstly compared with the memory block via a similarity metric. By doing so, pertinent content is extracted from the memory and coupled with the input clause. Subsequently, based on the chosen fixed amount of read iterations, the model either repeats the procedure described so far with the newly modified input, or uses content gathered so far to produce a prediction via a dedicated answer module.

edge base to retrieve the rationales that best match such input. Note that the system can relate each potentially unfair statement to multiple rationales. The links between statements and rationales established by the system provide a simple criterion for qualitative analysis of the model. In particular, if clause-rationale links were given for some samples, it would be possible to compare such references made by the experts with those made by the MANN. In broad terms, the method employed by the MANN to classify a given input clause is as follows. The MANN iteratively performs reading operations based on a similarity metric between each memory slot and the input clause. Subsequently, the MANN extracts memory content by combining all memory slots, and attributing to each slot a weight proportional to the similarity score. Next, the extracted content is added to the current input to build a representation that can possibly be distilled as a new input for another iteration (reasoning phase). In this way, past iterations are always taken into account during memory reading. Eventually, after the last memory iteration, the network operates on the distilled input to predict a fair/unfair label. Note that the same memory slot may be read multiple times in order to properly exploit its content, since a distribution of weights is applied to the whole memory block. Figure 1 illustrates the architecture.

3. The Dataset

In our previous research[6], we produced a dataset consisting of 50 relevant online consumer contracts, i.e., Terms of Services (ToS) of online platforms. The dataset now consists of 100 ToS. Such contracts were selected among those offered by some of the major players in terms of global relevance, number of users, and time the service was established. To train the ML classifier, these ToS were analyzed and marked in XML. We focused on eight categories of clauses, which most often are unlawful or unfair, i.e., clauses establishing: (1) jurisdiction in a state different from the consumer's; (2) choice of a law

other than the consumer's; (3) limitation of provider's liability; (4) provider's right to unilaterally terminate the contract and/or access to the service; (5) provider's right to unilaterally modify the contract/service; (6) arbitration on disputes arising from the contact; (7) provider's right to unilaterally remove consumer content from the service, including in-app purchases; (8) acceptance of contract by the mere use of the service, even when the consumer has not read the contract or explicitly agreed to it [6]. As reported by [22] and as our research indicates [5], such categories are widely used in ToS for online platforms. For the purposes of this study we focused on *limitation of liability* (LTD). Clauses falling under this category stipulate that the duty to pay damages is limited or excluded for certain kinds of losses and under certain conditions. One reason for focusing on limitation of liability is that LTD is the category for which we have the largest number of problematic clauses in our dataset. More precisely, our corpus contains a total of 21,063 sentences, 674 of which contain a potentially or clearly unfair clause (note that the total number of sentences containing a potentially unfair clause, in any category, is 2,346). In particular, clauses excluding liability for broad categories of losses or causes of them were marked as potentially unfair, including those containing blanket phrases like "to the fullest extent permissible by law". Conversely, clauses meant to reduce, limit, or exclude the liability for physical injuries, intentional harm, or gross negligence were marked as clearly unfair [6, 5]. The second observation concerns the particular difficulty of detecting unfair LTD clauses. Our classifier has shown lower performance on such clauses in comparison to other categories [6]. Moreover, focusing on a single category of unfairness makes it easier to circumscribe a dedicated knowledge base for testing the MANNs. However, this does not affect the significance of our experiments, since unfair limitation of liability can be identified on the basis of several different rationales.

An initial analysis enabled us to identify 21 legal rationales for (potentially) unfair limitation of liability, which map different questionable circumstances under which the ToS reduce or exclude liability for losses or injuries. For each rationale we defined a corresponding identifier `[ID]`. The rationales have been formulated by two independent legal experts, each adopting different approaches. The first approach was more synthetic, and produced a smaller number of broad grounds of unfair exclusion of liability (Table 1). The second approach was more analytical, and produced many explanations, each describing multiple kinds of unfairly excluded losses or damages (Table 2). The two lists of rationales were then merged and used together for running the experiments.

Table 1.: Legal rationales for the legal qualification of unfairness (synthetic approach).

ID	Legal Rationale
blanket_phrase	The limitation of liability uses a blanket phrase to the fullest extent permissible by law, any indirect or incidental damages, liability arising out of or in connection with these Terms or similar.
srv_con_liab	Liability is excluded in cases related to availability, usability or legality of service, website and/or user's content.
vir_malware	Liability is excluded for data loss, corruption or damage whether caused by viruses, trojan horses, malware or other malicious activity.
physical_harm	Liability is excluded also in cases of physical or personal injuries.
third_party	Liability is excluded for the actions and/or services of third parties.

Table 2.: Legal rationales for the legal qualification of unfairness (analytical approach).

ID	Legal Rationale
extent	since the clause states that to the fullest extent permissible by law the provider is not liable.
discontinuance	since the clause states that the provider is not liable for any technical problems, suspension, disruption, modification, discontinuance, limitation of services and features.
compharm	since the clause states that the provider is not liable for harm or damage to hardware and software, including viruses, worms, trojan horses, or any similar contamination or destructive program.
anydamage	since the clause states that the provider is not liable for any special, direct and/or indirect, punitive, incidental or consequential damage, including negligence, harm or failure.
amount	since the clause states that the compensation for liability or aggregate liability is limited to, or should not exceed, a certain amount.
thirdparty	since the clause states that the provider is not liable for any action taken from third parties or other people, including service and products, material and link posted by others.
security	since the clause states that the provider is not liable for any damage deriving from a security breach, including any unauthorised access.
disclosure	since the clause states that the provider is not liable for damages resulting from disclosure of data and personal information.
reputation	since the clause states that the provider is not liable for reputational and goodwill damages or loss.
anyloss	since the clause states that the provider is not liable for any loss resulting from the use of the service and or of the website, including lost profits, data, opportunity.
awareness	since the clause states that the provider is not liable whether or not he was, or should have been, aware about the possibility of any damage or loss.
contractfailure	since the clause states that the provider is not liable for any failure in performing contract and terms obligations, breach of agreement.
unilateral	since the clause states that the provider is not liable for any unilateral change or unilateral termination.
dataloss	since the clause states that the provider is not liable for any loss of data.
grossnegligence	since the clause states that the provider is not liable for gross negligence.
injury	since the clause states that the provider is not liable for personal injury and death.

Each unfair limitation of liability clause in the training set has been indexed with on or more identifiers of rationales that apply to the specific clause. As an example consider the following clause taken from the terms of service of Badoo (last updated on 11 September 2018) and previously classified as potentially unfair:

> To the fullest extent permitted by law, Badoo expressly excludes:all conditions, representations, warranties and other terms which might otherwise be implied by statute, common law or the law of equity; and any liability

```
incurred by you arising from use of Badoo, its services or these Terms,
including without limitation for any claims, charges, demands, damages,
liabilities, losses or expenses of whatever nature and howsoever direct,
indirect, incidental, special, exemplary, punitive or consequential damages
(however arising including negligence), loss of use, loss of data, loss caused
by a computer or electronic virus, loss of income or profit, loss of or damage
to property, wasted management or office time, breach of contract or claims
of third parties or other losses of any kind or character, even if Badoo has
been advised of the possibility of such damages or losses, arising out of or in
connection with the use of Badoo.
```

The clause above has been linked to the following identifiers one the analytical approach: `ID: extent, anydamage, compharm, anyloss, awareness, contractfailure, dataloss`. Conversely, on the synthetic approach, the clause has been associated to the following `ID: blanket_phrase, vir_malware`. The link between rationales and clauses will be used in future experiments to instruct the system so that it can provide an explanation for the unfairness of particular clauses.

4. Experimental Results

Our experiments use the proposed dataset of consumer contracts, focusing solely on LTD clauses. We employ a 10-fold cross-validation as both an evaluation and a calibration method for our models of interest. In particular, the whole corpus is first split into 10 subsets, named folds. Each fold is then used, in turn, as the test set, whereas the union of the other folds is further split into a training set and a validation set, exploited for hyper-parameter tuning. Concerning calibration, we consider two simple baselines that were also tested in some of our previous work [6] about unfairness detection in consumer contracts: (i) a network comprised of stacked recurrent neural network layers, a variant of RNNs referred to as Long Short-Term Memory network (LSTM), used extensively in the deep learning community; and (ii) a network defined as a stack of convolutional neural networks (CNNs). Moreover, we consider the current state-of-the-art solution for this task [6], featuring at its core a Support Vector Machine (SVM). Among all these models, the MANN is the only one that leverages an external knowledge base. The only input of LSTM, CNN, and SVM is the clause to be classified.

The memory network we propose follows the architecture described in [12] with minimal differences. With respect to the system illustrated in Section 2, the model performs six iterations over the memory before producing an answer. Qualitative analysis of this behaviour is reported below. Just like in [12], the similarity operation between the input clause and each content stored in memory is implemented as a simple dot product between the two sentence-embedding vectors, numerical representations of the corresponding texts. Far more complex implementations have been adopted in the literature [13, 20, 23, 24, 25], and we will consider them in future extensions. For the present study we decided to adopt the simplest similarity operation, because of the exploratory nature of our investigation. Once extracted, the distilled memory vector, defined as the weighted sum of read contents, is summed with the current input and used as input for the next iteration, as shown in Figure 1. As a last stage, a stack of fully connected layers is used to predict the classification label, given the latest memory-enhanced input.

To account for performance variations due to different initial configurations, we address model stability by repeating the cross-validation routine a sufficient amount of

Table 3. Results on 10-fold cross-validation. Performance measures are macro-averaged.

Model	Precision	Recall	F_1
Baseline LSTM	37.55	**88.93**	51.51
Baseline CNN	43.43	87.90	56.27
Memory Network	**68.36**	84.31	**64.33**
State of the art SVM	52.52	81.57	63.7

times. In our experimental setting we fix the number of repetitions to 10. In this way, it is possible to gather insight about performance variance and select at the same time the best-performing results. Input sensitivity is a crucial factor for detection systems, especially when the task is centred on infrequent or hard-to-detect anomalies. Furthermore, models were early stopped based on validation loss scores. Hyper-parameters calibration was accomplished via the same evaluation method, but without repetitions.

Table 3 reports the results of the proposed experimental setting. In particular, for all the models we report precision, recall, and F_1 scores, macro-averaged over the ten folds.[3] From the collected results, it is evident that baseline models that exclusively leverage the input-clause content fail to correctly classify the majority of legal violations in consumer contracts. On the other hand, the MANN model shows a strong improvement in performance with respect to the baselines. This indeed corroborates the rationale that background knowledge is a crucial element for this task. The proposed model exploiting an external memory shows even slightly better results than the state-of-the-art SVM.

Differently from other described architectures, modelling explicit comparison in memory networks presents the advantage of directly visualizing the interaction level of the model with respect to its memory blocks. This opens up the possibility of understanding *what* the model believes to be useful for the detection task. As an example, Figure 2 shows the overall memory usage over all folds. It is clear that the model does not exploit all the memory slots equally. The most used memory is the one stating *"The limitation of liability uses a blanket phrase like to the fullest extent permissible by law, any indirect or incidental damages, liability arising out of or in connection with these terms, or similar"*, which is one of the most general explanations, also providing several examples. Overall, the eight most used memories account for over 45% of the cases labeled by our experts, which is very interesting, since we point out that no information about which explanations were linked to which clauses was given during training. The latter would be called a *strong* supervision for memory networks, and we plan to use it in future work.

5. Conclusions

This paper investigates the use of memory-augmented deep learning models, for the automated detection of potentially unfair clauses, with a focus on limitation of liability.

This study was motivated by two main goals. The short-term goal was aimed at verifying whether the use of a knowledge base of rationales can improve the system performance in unfairness detection, while improving the reliability of the classification task. Our results are very encouraging: using a relatively small set of rationales, and no information about the link between explanations and clauses, the proposed MANN markedly

[3]For all the neural models, we have exploited multi-start by training ten different networks for each fold, and selecting the best network for each fold according to the validation performance.

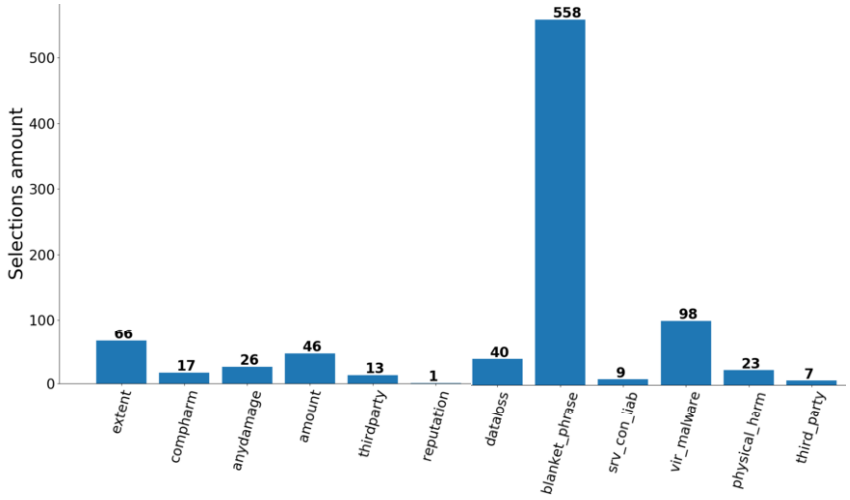


Figure 2. Cumulative memory distribution across all test folds. Slots never selected are not shown.

outperforms other neural models, while slightly improving over the state-of-the-art SVM. Moreover, memory-enhanced architectures inherently allow qualitative in-depth analyses of the model's behaviour, facilitating task-related investigations concerning relevant issues, such as trustworthiness and stability. Nonetheless, further steps are required, such as the annotation of memory targets for each input clause, to be exploited both to train the model towards task objectives, and to directly assess behaviour comparison. We are also working on the construction of a larger knowledge base of rationales relatively to other categories of unfair clauses, with the intention of improving training.

In the future, we plan to exploit the memory-enhanced architecture proposed in this study so as also to provide meaningful and trusted explanations for the users of CLAUDETTE, namely consumers, their organisations, and enforcement authorities.

References

[1] J.A. Obar and A. Oeldorf-Hirsch, The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services, *Information, Communication & Society* **0**(0) (2018), 1–20.

[2] K. Galbraith, American Capitalism The Concept of Countervailing Power, *American Journal of Agricultural Economics* **34**(4) (1952), 569–572.

[3] M. Lippi, G. Contissa, F. Lagioia, H.-W. Micklitz, P. Pałka, G. Sartor and P. Torroni, Consumer protection requires artificial intelligence, *Nature Machine Intelligence* **1**(4) (2019), 168.

[4] P. Pałka and M. Lippi, Big Data Analytics, Online Terms of Service and Privacy Policies, *Research Handbook on Big Data Law edited by Roland Vogl* (2019).

[5] H.-W. Micklitz, P. Pałka and Y. Panagis, The empire strikes back: digital control of unfair terms of online services, *Journal of consumer policy* **40**(3) (2017), 367–388.

[6] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor and P. Torroni, CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service, *Artificial Intelligence and Law* **27**(2) (2019), 117–139.

[7] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger and A. Wood, Accountability of AI under the law: The role of explanation, *arXiv preprint arXiv:1711.01134* (2017).
[8] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* (2018).
[9] S. Wachter, B. Mittelstadt and C. Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR, *Harv. JL & Tech.* **31** (2017), 841.
[10] J. Weston, S. Chopra and A. Bordes, Memory networks, *arXiv preprint arXiv:1410.3916* (2014).
[11] A. Graves, G. Wayne and I. Danihelka, Neural Turing machines, *arXiv preprint arXiv:1410.5401* (2014).
[12] S. Sukhbaatar, J. Weston, R. Fergus et al., End-to-end memory networks, in: *Advances in neural information processing systems*, 2015, pp. 2440–2448.
[13] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus and R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: *International conference on machine learning*, 2016, pp. 1378–1387.
[14] A. Bordes, Y.-L. Boureau and J. Weston, Learning end-to-end goal-oriented dialog, *arXiv preprint arXiv:1605.07683* (2016).
[15] A. Bordes, N. Usunier, S. Chopra and J. Weston, Large-scale simple question answering with memory networks, *arXiv preprint arXiv:1506.02075* (2015).
[16] D. Tang, B. Qin and T. Liu, Aspect level sentiment classification with deep memory network, *arXiv preprint arXiv:1605.08900* (2016).
[17] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes and J. Weston, Key-value memory networks for directly reading documents, *arXiv preprint arXiv:1606.03126* (2016).
[18] F. Hill, A. Bordes, S. Chopra and J. Weston, The Goldilocks principle: Reading children's books with explicit memory representations, *arXiv preprint arXiv:1511.02301* (2015).
[19] J. Cheng, L. Dong and M. Lapata, Long short-term memory-networks for machine reading, *arXiv preprint arXiv:1601.06733* (2016).
[20] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S.G. Colmenarejo, E. Grefenstette, T. Ramalho et al., Hybrid computing using a neural network with dynamic external memory, *Nature* **538**(7626) (2016), 471.
[21] W. Zaremba and I. Sutskever, Reinforcement learning neural turing machines-revised, *arXiv preprint arXiv:1505.00521* (2015).
[22] M. Loos and J. Luzak, Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers, *Journal of consumer policy* **39**(1) (2016), 63–90.
[23] C. Xiong, S. Merity and R. Socher, Dynamic memory networks for visual and textual question answering, in: *International conference on machine learning*, 2016, pp. 2397–2406.
[24] M. Henaff, J. Weston, A. Szlam, A. Bordes and Y. LeCun, Tracking the world state with recurrent entity networks, *arXiv preprint arXiv:1612.03969* (2016).
[25] J. Pavez, H. Allende and H. Allende-Cid, Working memory networks: Augmenting memory networks with a relational reasoning module, *arXiv preprint arXiv:1805.09354* (2018).

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190306

A Comparison of Two Hybrid Methods for Analyzing Evidential Reasoning

Ludi VAN LEEUWEN, Bart VERHEIJ
Department of Artificial Intelligence, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen

Abstract. Reasoning with evidence is error prone, especially when qualitative and quantitative evidence is combined, as shown by infamous miscarriages of justice, such as the Lucia de Berk case in the Netherlands. Methods for the rational analysis of evidential reasoning come in different kinds, often with arguments, scenarios and probabilities as primitives. Recently various combinations of argumentative, narrative and probabilistic methods have been investigated. By the complexity and subtlety of the subject matter, it has proven hard to assess the specific strengths and points of attention of different methods. Comparative case studies have only recently started, and never by one team. In this paper, we provide an analysis of a single case in order to compare the relative merits of two methods recently proposed in AI and Law: a method using Bayesian networks with embedded scenarios, and a method using case models that provide a formal analysis of argument validity. To optimise the transparency of the two analyses, we have selected a case about which the final decision is undisputed. The two analyses allow us to provide a comparative evaluation showing strengths and weaknesses of the two methods. We find a core of evidential reasoning that is shared between the methods.

Keywords. Bayesian networks, case models

1. Introduction

Reasoning with evidence is difficult. This is especially pertinent in court, where reasoning correctly about evidence can mean the difference between a rightful conviction, or a wrongful imprisonment. To safeguard against errors, three tools for the rational analysis have been investigated in the literature: argument-based, scenario-based, and probabilistic [1,2,3,4]. In argumentative analyses, the emphasis is on argument structure, defeat and evaluation [5,6,7,8,9]. In scenario methods, with roots in legal psychology, the emphasis is on the construction and comparison of coherent explanatory scenarios and their relation to the evidence [10,11,12,13,14]. Probabilistic tools analyze how hypothetical events are probabilistically related to the evidence and to evidential updating, in particular by using Bayesian networks [15,16]. Hybrid approaches investigate, for instance, combinations of scenarios and arguments [17], evidential Bayesian networks [15,18,16], scenarios and probabilities [19,20]. Comparative case studies for assessing the relative merits of approaches are as yet rare. A recent valuable effort to this effect is the study of the Simonshaven case using different methods (upcoming issue 'Models of Rational Proof in Criminal Law' in the journal *Topics in Cognitive Science*, editors Henry Prakken, Floris Bex and Anne Ruth Mackor).

In this paper, two methods recently proposed in AI and Law are compared and evaluated: Bayesian networks with embedded scenarios [21], and case models that provide a formal analysis of argument validity [22]. For this, we develop two analyses of a murder case, one for each method. Since we are developing both analyses ourselves, we can aim for optimal similarity, increasing comparability (in contrast with the analyses in *Topics in Cognitive Science*, each developed by a separate team). The case is based on a real case,[1] simplified for present purposes. To improve transparency, we have selected a case with undisputed conclusion:

> On October first, 2002, N, a 25-year-old student is found dead in her apartment. There are signs of violence: bullet casings and blood. Before she died, she had called a friend. The friend reported a normal conversation, then heard a 'good morning', followed by yells and loud sounds, before the call dropped. A suspect was soon identified: P, the son of the landlord, who also lived in the apartment. He fled to Poland before he could be apprehended, and was only arrested in 2003. The court found P guilty of the murder of N in 2004.

2. Methods compared

2.1. Bayesian networks with embedded scenarios

A Bayesian network is a directed acyclic graph with associated conditional probabilities, and represents a joint probability distribution [24]. Bayesian networks can be used to avoid common fallacies in probabilistic reasoning [25]. The probability distribution can be found by elicitation techniques [26], although the lack of data makes objective priors difficult to find [27].

Probabilistic tools and the scenario approach are combined in [21] to construct a Bayesian network via scenario idioms. A scenario idiom consists of a boolean scenario node, and child-nodes representing aspects of that scenario. When the scenario node is true then all child-nodes must also be true. This ensures coherence, and transfer of evidential support. In the method, mutually exclusive scenarios are modeled via a constraint node (see [28]). Child nodes can represent abstract aspects that a court needs to prove, like motive or opportunity. Aspect nodes can be connected to other aspect nodes, and must be supported by evidence nodes. Evidence nodes are conditional on the aspect nodes.

2.2. Case Models

Case models are a formal tool for the analysis of coherent, presumptive and conclusive arguments using a preference ordering of cases [22]. The formalism is inspired by the connections between the three approaches to evidence. A case model can be constructed by adding evidence piecewise (argumentative) to construct coherent hypotheses (scenarios) of varying credibility (probabilistic).

A case model consists of a set of cases C, and their preference ordering $\leq$. Cases combine hypothetical events and evidence. The preference ordering depends on the coherence, conclusiveness, and presumptive validity of the arguments [22] of the cases.

[1] Rechtbank Utrecht, see case ECLI:NL:RBUTR:2004:AO3150, also used in [23].

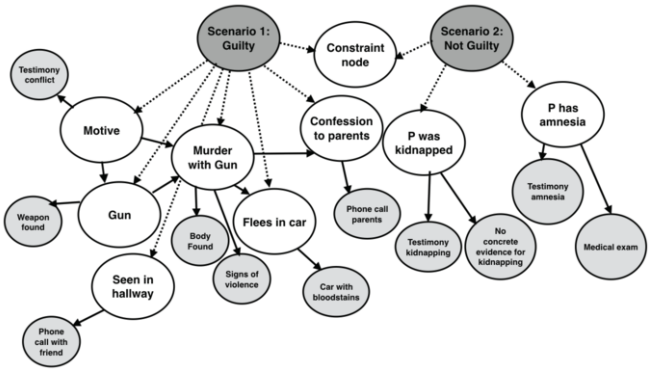


Figure 1. The Bayesian network of the case. The dark grey nodes are the scenario nodes: links between the scenario nodes and the aspect nodes are represented by dotted lines. The white nodes are the aspect nodes. The light grey, small nodes are the evidence nodes.

3. Models

In the following section, the methods for creating the Bayesian network and the case model are discussed.

3.1. Bayesian Networks

Following the method described by [21], two different scenarios of the case were constructed. The scenarios were modeled in a Bayesian network. The probability tables of each node were determined. The different nodes were turned off and on to see how each piece of evidence influences the probabilities in the scenario nodes.

3.1.1. Step 1: Create scenarios

Scenario 1 This scenario is based on the arguments of the prosecution. Suspect P murdered victim N with a gun. P had a motive, he was angry about an earlier conflict. He also had an illegal gun. N had been at home, on the phone with a friend. The friend testified that she heard N greet someone, followed by the sounds of gunfire and screaming. This greeting places P at the scene, as the other tenants had already left for work. After the murder, P flees in N's car, leaving blood traces behind. He flees to Poland. When he is in Poland, he makes several phone calls to his parents. In these phone calls, he confesses that he did something to N.

Scenario 2 This scenario is based on P's testimony. In this scenario P has been kidnapped, and he also has amnesia. P does not remember killing N, or where he was the morning of the crime.

3.1.2. Step 2: Creating the nodes and connections of the Bayesian Network

The complete network structure is shown in Figure 1. This is a diagram representation of the network that was created with GeNIe and AgenaRisk. The two scenario nodes were implemented first, connected by a constraint node.

Table 1. The probability table for gun. Numbers are based on the base rate of gun ownership, 6% in the Netherlands,[3] and the (debatable) assumption that people with a motive are more likely to own a firearm.

	Probability of having a gun
Scenario1 and Motive	0.2
Scenario1 and ¬Motive	0
¬Scenario1 and Motive	0.2
¬Scenario1 and ¬Motive	0.06

The aspect nodes for **scenario 1** are: `motive`, which represents the motive of P, supported by testimony of both his parents and N's friends about the conflict; gun, which represents the gun P had in his home, supported by the weapon being found, `seen in hallway`, which places P at the scene, and is supported by N's phone call with her friend. These three nodes are parent nodes of `murder with gun`, which represents P's murder of N, and is supported by N's body being found and signs of violence at the scene, like blood traces and bullet shells. The `murder with gun` node has two child-nodes: `flees in car`, representing how P flees, supported by N's car being found, which has been used after she died, and had her blood in it, and `confession to parents` node, representing P's confession to his parents over the phone, supported by his phone call. The node `motive` is also the parent of gun, as these are not two independent events: the probability of having a gun is not independent of the probability of having a motive.

The nodes for **scenario 2** are: the aspect node `kidnapped`, which represents P's kidnapping by unknown persons, which is supported by his testimony, evidence node `testimony P`, but detracted by evidence node: `no concrete evidence for kidnapping`, which represents that there is no concrete evidence, apart from P's testimony, that he was kidnapped. The aspect node `amnesia`, is supported by evidence node `P's testimony`, which represents P's testimony that he doesn't remember anything, and the node `medical investigation found no amnesia`, which represents the fact that there was no physical cause for amnesia as determined by a doctor.

3.1.3. Step 3: Creating the probability tables

Every node has an associated table, containing the probabilities of the node, conditioned on the values of the parents. Table 1 shows the probability table for the gun node, which depends on the value of the scenario node, and the motive node. The probabilities in the nodes are based on subjective choices. The constraint node has a value of NA when it was not the case that exactly one scenario was true.

3.1.4. Evidence flow through the network

By turning the evidence nodes off and on, the cumulative effect of different pieces of evidence on the probabilities of different scenarios is shown (Table 2). Presumption of innocence was modeled by having the prior probability of the guilty scenario node set to 50%, and the prior probability of the non-guilty scenario set to 50%, following [27].

3.2. Case Models

A case model (Figure 3) is created through a visual exploration of evidence (Figure 2). In this case study, evidence was collected from the court case. Then, the visual interpretation was created, where evidence was added step-by-step (in the same order as the nodes

Table 2. Rounded down cumulative evidence in nodes for scenario 1, guilty and scenario 2, not guilty, evidence is turned on in the same order as evidence is added to the case model. The probability of one scenario does not affect the probability of the other scenario if there are no nodes that belong to both scenarios.

Evidence	P(Scenario 1) in %	P(Scenario 2) in %
Start	50	50
Body found	43	56
Signs of violence	76	24
Weapon found	83	16
Phone call with friend	83	16
Testimony kidnapping	75	24
Testimony amnesia	64	35
Car with bloodstains	75	25
Testimony conflict	75	25
No concrete evidence of kidnapping	96	4
Medical investigation found no amnesia	99	1
Phone call parents	close to 100	close to 0

were turned on in (Table 2). From the visual interpretation, different hypotheses were collected. The hypotheses were then joined with maximally coherent evidence in order to create cases [22].

3.2.1. Step 1: Visual interpretation of the case model

A body is found. At this point, there is no evidence to assume a crime. However, there are (`signs_of_violence`), including bullet wounds and a found gun (`weapon_found`), so the victim was murdered with a gun.

Except for the victim, P was the only person in the house, and he was heard on the phone (`phone_call_with_friend`), so he is a suspect, and either guilty, or not guilty. P was then interviewed, and testified that he had been kidnapped, and that he had amnesia (`testimony_kidnapping`), (`testimony_amnesia`). The hypothesis of P not being guilty is further subdivided: either he is not guilty and he is telling the truth about the kidnapping and the amnesia, or he is not guilty and something else happened.

More evidence is added: N's (`car_with_bloodstains`) was found, moved after she was already dead, suggesting that P fled in her car. N's parents also testified about a conflict between P and N, which offers a motive (`testimony_conflict`). P's testimony conflicts with the results of a (`medical_examination`), which shows no physical cause for amnesia, as well as (`no_concrete_evidence`) of any kidnapping. The last piece of evidence is (`phone_call_parents`), he confesses that he did something to N in a phone call to his parents.

3.2.2. Step 2: Collect hypotheses

Every case in the case model has a hypothesis. This hypothesis can be found in the columns of the case model. This case model has the following three hypotheses:

1. `P_is_guilty`
2. ¬`P_is_guilty` ∧ `P_was_kidnapped` ($\neg P \wedge K$)
3. ¬`P_is_guilty` ∧ ¬`P_was_kidnapped` ($\neg P \wedge \neg K$)

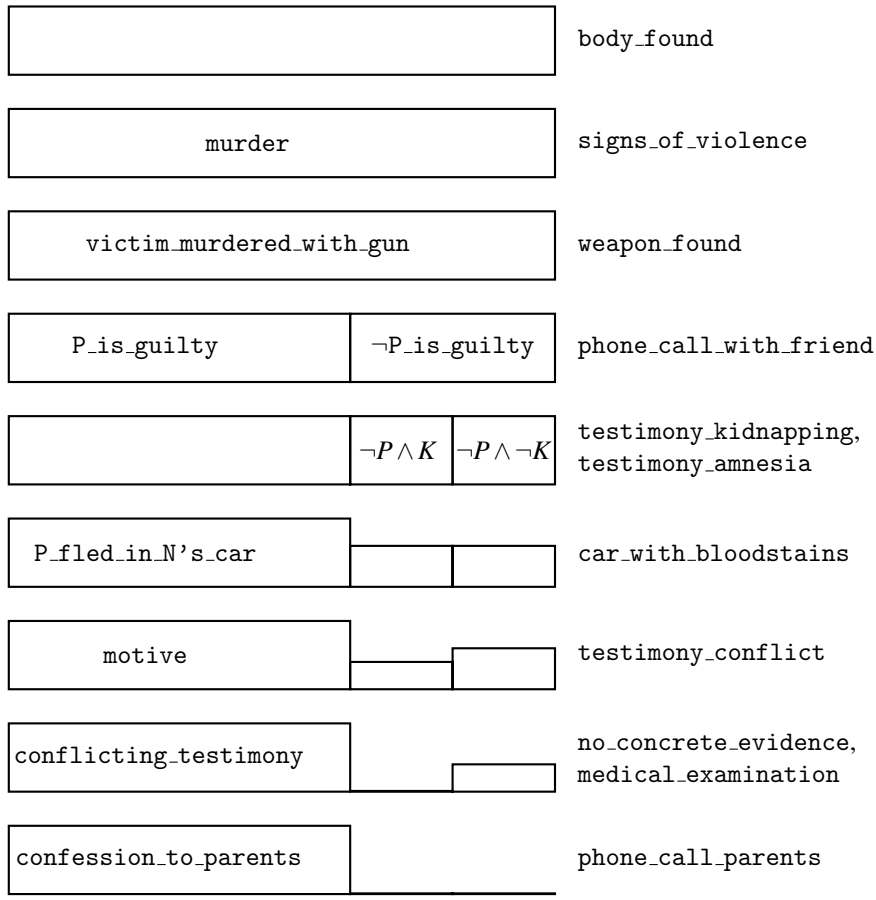


Figure 2. The case model creation, adding evidence chronologically

3.2.3. Step 3: Create cases by adding evidence to hypotheses

To create the cases, each hypothesis is extended with the subset of evidence that is coherent with the hypothesis. The evidence that is common to all three hypotheses: body_found ∧ signs_of_violence ∧ weapon_found ∧ phone_call_with_friend, is represented in these cases by E, for conciseness. There are 7 cases in total, shown in Figure 3.

1. P_is_guilty ∧ murder ∧ victim_murdered_with_gun ∧ P_fled_in_N's_car ∧ motive ∧ conflicting_testimony ∧ confession_to_parents ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ testimony_conflict ∧ no_concrete_evidence ∧ medical_examination ∧ phone_call_parents.
2. ¬P_is_guilty ∧ P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ ¬ car_with_bloodstains.

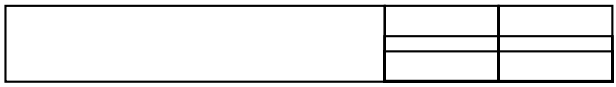


Figure 3. The final case model creation

3. ¬P_is_guilty ∧ P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ ¬ testimony_conflict .
4. ¬P_is_guilty ∧ P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ testimony_conflict.
5. ¬P_is_guilty ∧ ¬P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ ¬ car_with_bloodstains.
6. ¬P_is_guilty ∧ ¬P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ testimony_conflict ∧ ¬ no_concrete_evidence ∧ ¬ medical_examination.
7. ¬P_is_guilty ∧ ¬P_was_kidnapped ∧ murder ∧ victim_murdered_with_gun ∧ E ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ testimony_conflict ∧ no_concrete_evidence ∧ medical_examination.

The preference ordering is, as represented by the areas of the different boxes: $1 > 2 \sim 4 \sim 5 \sim 7 > 3 \sim 6$.

The arguments (T, body_found) and (T, body_found ∧ signs_of_violence) are coherent, conclusive and presumptively valid, as everyone agrees that a body was found, and violence was committed.

The argument (body_found ∧ signs_of_violence ∧ testimony_conflict, P_is_guilty) is presumptively valid and coherent, but not conclusive, as (body_found ∧ signs_of_violence ∧ testimony_conflict, ¬P_is_guilty ∧ P_was_kidnapped) is also coherent.

The argument (body_found ∧ signs_of_violence ∧ weapon_found ∧ phone_call_with_friend ∧ testimony_kidnapping ∧ testimony_amnesia ∧ car_with_bloodstains ∧ testimony_conflict ∧ no_concrete_evidence ∧ medical_examination ∧ phone_call_parents, P_is_guilty) is coherent, conclusive, and presumably valid.

4. Comparative evaluation

We have provided two analyses of one case using very different formal methods. Also of the Simonshaven case, both a Bayesian network [27] (but not with embedded scenarios as we did here, following [21]) and a case model [29] were made. However, as these were prepared by separate authors, these analyses are based on rather different selections of what is modeled about the case and how. Here we have aimed to optimise similarity between the two models in order to allow for a more specific comparative evaluation. Also the Simonshaven case can be considered as a 'hard case' with a disputable outcome, whereas we selected a case with an undisputed outcome.

The Bayesian network with embedded scenarios (the BNS model) consists of a directed acyclic graph (with associated conditional probability tables) modeling the evidence/events and their probabilistic dependencies. In contrast, the case model (CM) con-

sists of sentences and a preference ordering modeling coherent combinations of evidence and hypothetical events. The ordering models comparative credibility of the cases and can be given a probabilistic interpretation. The two models hence provide very different ways to connect qualitative and quantitative modeling styles.

The two analyses provide a perspective on the stepwise influence of the evidence that is well aligned, as can be seen by comparing Table 2 and Figure 2. The design of the CM model (representing the stepwise construction of a theory about the case) influenced the choice of numbers for the BNS model, thereby optimizing the alignment.

The BNS model allows fine-grained numeric estimates of relevance and strength, where the CM model uses a cruder ordering. The effects of numeric propagation in the BN model are not easy to predict and interpret (cf. the influence of the CM model on the BNS model). For both methods, it is not obvious how to choose differences (numbers and ordering, respectively).

Both the BNS and the CM model can model conflicts of the evidence. In the BNS model, adding evidence can have a positive or negative effect on the probability of a scenario, and in the CM model evidence can match and exclude hypothetical scenarios.

The coherent clustering of events in scenarios is modeled in the BNS model using scenario nodes and constraint nodes (cf. [28]). In the CM model, clusters of evidence with scenarios are modeled by mutually exclusive cases.

The scenario nodes add to the explanation of the BNS model. Also turning evidence on and off (as in a BN software tool) helps to uncover the influence of the evidence on hypotheses. However the meaning of results and how they come about is not always transparent (why is an outcome 10%? why 50%?). The CM model's construction (Figure 2) allows for an explanation that as said could support alignment with Table 2. The final decision in the model has a transparent explanation, but the choice of ordering remains an issue.

For justifying a decision, the BNS model allows for a choice after picking a threshold posterior probability (e.g., 95%), providing a clean and precise model of justification. However, there is no obvious choice of threshold. In the CM model, justification of a decision has the form of coherence after exclusion of all alternatives considered. For both, the question remains whether there are unconsidered, unmodeled alternatives that could change the decision.

On ease of modeling, choosing dependencies and numbers for the BNS model was not easy, but once built, the role of the evidence on the hypothetical outcomes can be directly tested. The CM model was easy to construct. It seems positive that indeed consistency with the probabilistic BNS model was possible (as suggested by the theoretical fact that CM models allow for a numeric, probabilistic interpretation). At the same time, it is not clear whether the focus on only an ordering in the CM model reduces the expressiveness allowed by a BNS model, where subtle interaction effects can be modeled.

5. Summary and conclusion

Both Bayesian networks with embedded scenarios and case models can be used as hybrid tools to combine probabilistic, scenario and argumentation approaches to investigating hypotheses and evidence. In this paper, we have analyzed one case (with an undisputed outcome) using these two methods aiming for optimal similarity by using comparable modeling elements.

Evidence in Bayesian networks is relevant across the whole network, have a precise interpretation of evidence strength, and are straightforward in their use, once they have been created. Relations between different pieces of evidence is done by making subjective probabilities explicit, and truth is decided based on threshold probabilities. Coherence is not an inherent feature to Bayesian networks. Modelling a Bayesian network and deciding on the probabilities is a challenge.

In case models, evidence can be more local, confined to relevant cases. How to include quantitative data using only the ordering of cases is not clear. Coherence is inherent, and justification and explanation seem more similar to human reasoning than the conditional probabilities of the Bayesian network. Modelling a case model visually is straightforward, although extracting the cases and the ordering is not.

Both methods have limitations: the subjective probability assessment, especially in combination with nodes with a large probability table (nodes with many parents), is a problem in Bayesian networks. Coherency is similarly subjective, as well as the lack of expression of evidence strength in case models. However, both methods might help resolve inconsistent or incorrect reasoning with evidence by guiding the reasoners to consider conflicting pieces of evidence and overall coherence of each hypothesis or scenario.

We saw that the evidential progression modeled in two different ways (Table 2, Figure 2) could be well aligned, suggesting that such reasoning provides a shared core of evidential reasoning in the different methods. That progression is also at the heart of the modeling style in [25].

Reliable probability elicitation methods [26] are needed for the proper use of Bayesian networks. However, as the probabilities needed are often unobservable, giving appropriate probabilistic assessments can be very difficult. One idea to improve the understanding of probabilistic assessments would be to create a multi-agent simulation, with certain known incidences of different crimes. Then, different ranges of probabilities, based on those in this simulation, can be used in Bayesian networks to test different scenarios inside these worlds.

The case model theory does not exclude the possibility of quantification. However, it is unclear how this would work in practise: would this mean that every part of the case model is quantified, or can there be a mixed approach? Modelling a case that is more reliant on statistical inferences (for example, with DNA evidence), would be useful in finding the limitations of quantification in case models.

A useful, stricter evaluation of the methods could come from a systematic comparative analysis of other cases, more complex than the present one, like in the Simonshaven case studies, but then with carefully guarded model similarity.

References

[1] T. Anderson, D. Schum and W. Twining, *Analysis of Evidence. 2nd Edition*, Cambridge University Press, Cambridge, 2005.

[2] H. Kaptein, H. Prakken and B. Verheij (eds), *Legal Evidence and Proof: Statistics, Stories, Logic (Applied Legal Philosophy Series)*, Ashgate, Farnham, 2009.

[3] A.P. Dawid, W. Twining and M. Vasiliki (eds), *Evidence, Inference and Enquiry*, Oxford University Press, Oxford, 2011.

[4] M. Di Bello and B. Verheij, Evidential Reasoning, in: *Handbook of Legal Reasoning and Argumentation*, G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini and D.N. Walton, eds, Springer, Dordrecht, 2018, pp. 447–493.

[5] D.N. Walton, *Legal Argumentation and Evidence*, The Pennsylvania State University Press, University Park (Pennsylvania), 2002.
[6] F.J. Bex, H. Prakken, C.A. Reed and D.N. Walton, Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations, *Artificial Intelligence and Law* **11**(2/3) (2003), 125–165.
[7] T.F. Gordon and D.N. Walton, Proof Burdens and Standards, in: *Argumentation in Artificial Intelligence*, I. Rahwan and G.R. Simari, eds, Springer, 2009, pp. 239–258.
[8] H. Prakken and G. Sartor, A Logical Analysis of Burdens of Proof, in: *Legal Evidence and Proof: Statistics, Stories, Logic*, H. Kaptein, H. Prakken and B. Verheij, eds, Ashgate, 2009, pp. 223–253, Chapter 9.
[9] F.H. van Eemeren, B. Garssen, E.C.W. Krabbe, A.F. Snoeck Henkemans, B. Verheij and J.H.M. Wagemans, Chapter 11: Argumentation in Artificial Intelligence, in: *Handbook of Argumentation Theory*, Springer, Berlin, 2014.
[10] W.L. Bennett and M.S. Feldman, *Reconstructing Reality in the Courtroom*, London: Tavistock Feldman, 1981.
[11] N. Pennington and R. Hastie, Reasoning in explanation-based decision making, *Cognition* **49**(1–2) (1993), 123–163.
[12] W.A. Wagenaar, P.J. van Koppen and H.F.M. Crombag, *Anchored Narratives. The Psychology of Criminal Evidence*, Harvester Wheatsheaf, London, 1993.
[13] J. Keppens and B. Schafer, Knowledge Based Crime Scenario Modelling, *Expert Systems with Applications* **30**(2) (2006), 203–222.
[14] R.J. Allen and M.S. Pardo, The Problematic Value of Mathematical Models of Evidence, *Journal of Legal Studies* **36**(1) (2007), 107–140.
[15] A.B. Hepler, A.P. Dawid and V. Leucari, Object-Oriented Graphical Representations of Complex Patterns of Evidence, *Law, Probability and Risk* **6**(1–4) (2007), 275–293.
[16] N.E. Fenton, M.D. Neil and D.A. Lagnado, A General Structure for Legal Arguments About Evidence Using Bayesian Networks, *Cognitive Science* **37** (2013), 61–102.
[17] F.J. Bex, P.J. van Koppen, H. Prakken and B. Verheij, A Hybrid Formal Theory of Arguments, Stories and Criminal Evidence, *Artificial Intelligence and Law* **18** (2010), 1–30.
[18] J. Keppens, Argument Diagram Extraction from Evidential Bayesian Networks, *Artificial Intelligence and Law* **20** (2012), 109–143.
[19] M. Di Bello, *Statistics and Probability in Criminal Trials: The Good, the Bad and the Ugly. Dissertation*, Stanford University, Stanford, 2013.
[20] R. Urbaniak, Narration in Judiciary Fact-Finding: a Probabilistic Explication, *Artificial Intelligence and Law* **26** (2018), 345–376.
[21] C.S. Vlek, H. Prakken, S. Renooij and B. Verheij, A Method for Explaining Bayesian Networks for Legal Evidence with Scenarios, *Artificial Intelligence and Law* **24**(3) (2016), 285–324.
[22] B. Verheij, Proof With and Without Probabilities. Correct Evidential Reasoning with Presumptive Arguments, Coherent Hypotheses and Degrees of Uncertainty, *Artificial Intelligence and Law* **25**(1) (2017), 127–154.
[23] F.J. Bex and B. Verheij, Solving a Murder Case by Asking Critical Questions: An Approach to Fact-Finding in Terms of Argumentation and Story Schemes, *Argumentation* **26** (2012), 325–353.
[24] F. Taroni, C. Aitken, P. Garbolino and A. Biedermann, *Bayesian Networks and Probabilistic Inference in Forensic Science*, Wiley, Chichester, 2006.
[25] C. Dahlman, De-Biasing Legal Fact-Finders With Bayesian Thinking, *Topics in Cognitive Science* (2019), 1–17. doi:10.1111/tops.12419.
[26] S. Renooij, Probability Elicitation for Belief Networks: Issues to Consider, *The Knowledge Engineering Review* **16**(3) (2001), 255–269.
[27] N. Fenton, M. Neil, B. Yet and D. Lagnado, Analyzing the Simonshaven Case Using Bayesian Networks, *Topics in Cognitive Science* (2019), 1–23. doi:10.1111/tops.12417.
[28] N.E. Fenton, M.D. Neil, D.A. Lagnado, W. Marsh, B. Yet and A. Constantinou, How to Model Mutually Exclusive Events Based on Independent Causal Pathways in Bayesian Network Models, *Knowledge-Based Systems* **113** (2016), 39–50.
[29] B. Verheij, Analyzing the Simonshaven Case With and Without Probabilities, *Topics in Cognitive Science* (2019), 1–25. https://doi.org/10.1111/tops.12436.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190307

Similarity and Relevance of Court Decisions: A Computational Study on CJEU Cases

Kody MOODLEY[a,b,1], Pedro V. HERNANDEZ SERRANO[a], Gijs VAN DIJCK[b] and Michel DUMONTIER[a]
[a] *Institute of Data Science, Maastricht University*
[b] *Faculty of Law, Maastricht University*

Abstract. Identification of relevant or similar court decisions is a core activity in legal decision making for case law researchers and practitioners. With an ever increasing body of case law, a manual analysis of court decisions can become practically impossible. As a result, some decisions are inevitably overlooked. Alternatively, network analysis may be applied to detect relevant precedents and landmark cases. Previous research suggests that citation networks of court decisions frequently provide relevant precedents and landmark cases. The advent of text similarity measures (both syntactic and semantic) has meant that potentially relevant cases can be identified without the need to manually read them. However, how close do these measures come to approximating the notion of relevance captured in the citation network? In this contribution, we explore this question by measuring the level of agreement of state-of-the-art text similarity algorithms with the citation behavior in the case citation network. For this paper, we focus on judgements by the Court of Justice of the European Union (CJEU) as published in the EUR-Lex database. Our results show that similarity of the full texts of CJEU court decisions does not closely mirror citation behaviour, there is a substantial overlap. In particular, we found syntactic measures surprisingly outperform semantic ones in approximating the citation network.

Keywords. Text Similarity, Word Embeddings, Network Analysis, CJEU

1. Introduction

Within the setting of case law, the identification and citation of relevant court decisions to support judicial decision making is a central activity. Network Analysis methodology [1,2,3,4] has proven to be useful for *a posteriori* analysis of court decision citation behavior, for example, in identifying legal precedents and measuring the influence of decisions. However, an *a priori* understanding of what constitutes a relevant case (w.r.t. to a given case) remains a complex and multifaceted question. In law generally, the concept of relevance has been previously studied and there have been attempts to define it for Legal Information Retrieval (LIR) tasks [5]. However, to date, there has been no measurable specialisation of this definition for case law.

The publishing of court decisions online as full texts in databases such as EUR-Lex (`https://eur-lex.europa.eu`) and HUDOC (`https://hudoc.echr.coe.int`), and the advancement of text similarity algorithms [6,7,8], has enabled the automatic search

[1] Corresponding Author: Institute of Data Science at Maastricht University, Universiteitssingel 60 (1st floor), 6229 ER, Maastricht, The Netherlands; E-mail: kody.moodley@maastrichtuniversity.nl

and retrieval of similar (and potentially relevant) cases. Many such measures are implemented in proprietary software such as ROSS (`https://rossintelligence.com`) and Lex Machina (`https://lexmachina.com`). The commercial success of these platforms suggest that the algorithms have promising accuracy and, therefore, text similarity may prove to be a useful tool for computationally characterising case relevance. However, there are caveats to these technologies. One is that many of these platforms do not explain *why* they found particular cases relevant, and therefore, it is difficult to measure and benchmark their legal merit. In particular, we are interested in measuring *recall* or *completeness* of these algorithms (and to a lesser extent, their *precision* or *accuracy*).

In order to establish a benchmark for completeness, we need to capture an understanding of relevance in a legal context, case law in particular. One possible strategy to achieve this is to solicit legal experts to annotate court decision texts with information (e.g. legal principles, topics and arguments) that they use to evaluate case relevance [9]. While we advocate such an approach for the longer term, there are alternatives to explore in the interim that would yield equally interesting insights with lower demand on time and resources. One of these, which we adopt in this work, is to select our base understanding for relevance to be equivalent to *citation* as captured in the *court decision citation network* (CDCN for short). Accepting this notion of relevance, we compare it to several state-of-the-art text similarity measures applied to cases from the Court of Justice of the European Union (CJEU). We use these algorithms to generate what we call a *court decision similarity network* (CDSN) - an analogue of the CDCN in which links between decisions imply high textual similarity. The graphical difference between a CDSN and a CDCN is that the edges of a CDSN are undirected, whereas those in a CDCN are directed. The goal of our study is to evaluate the size of overlap between the CDSNs generated by selected text similarity algorithms and the CDCN. Our results contribute towards an answer to the question: *to what extent can state-of-the-art text similarity measures capture the citations in the CJEU CDCN?*

The remainder of the paper is organised as follows: in Section 2, we provide an overview of related work in relevance and textual similarity of court decisions. In Section 3, we introduce the methodology of our study which includes descriptions of the selected dataset, sampling strategy and text similarity algorithms. Section 4 discusses our main findings, Section 5 outlines the caveats, limitations and challenges of the evaluation, and Section 6 summarises what we learned in the study, our plans for extending the work, and the licensing and availability of the data and software used.

2. Related Work[o]

In terms of efforts to define relevance for legal information retrieval, van Opijnen & Santos [5] provide a conceptual framework to categorise and define dimensions of relevance. There are six types listed: *algorithmic, topical, bibliographic, cognitive, situational* and *domain.* While this work provides a foundation for defining legal relevance, to date there has not been any mechanism proposed for *measuring* these relevance dimensions for specific legal topics.

In a separate endeavour, van Opijnen [10] has also established a model for ranking importance of case law. In this work, the author arrives at predictors for whether a case will play a marked role in future legal debate (based on its discussion in the legal community from the point of inception). Malmgren [11] also studies the notion of

relevance in LIR and also within the context of CJEU decisions. However, it appears that both these efforts presume a futility in developing reliable computational algorithms for finding relevant cases, that only take the decision texts or content into account. One major reason being intrinsic subjectivity in the notion of what legal experts might consider relevant. Therefore, there are many studies that try to measure the importance and relevance of case law by means of studying the CDCN through the use of Network Analysis metrics [12,13,14]. Network Analysis was also validated as a useful way to measure relevance and importance for Dutch cases [15]. There are also many efforts to apply text similarity measures to find relevant cases in the literature. Sugathadasa et al. [16] apply deep learning to train a similarity classifier for cases from FindLaw (https://www.findlaw.com). In order to measure performance, ground truth is based on validation by legal experts. Raghav, K. [17] also provide a method to augment similarity analyses of cases based on Network Analysis with text similarity on the paragraph level. The authors found a very high agreement between citation metrics and paragraph similarity on their dataset of Indian Supreme Court judgements. Panagis et al. [18] performed an interesting study on CJEU decisions to identify what they call "implicit" citations. These are references between cases that are not explicitly stated in the cited instruments of the decision but those identifiable from the text. They use the *Tversky index* measure [19] to compare similarity of paragraphs between cases. This approach proved that the CDCN does not provide the full picture of relevant cases and provides motivation for further research into increasing recall of case retrieval.

3. Methodology

In this section we detail our methodology for constructing the CDCN and CDSNs in the study and how we calculated the size of their overlap.

Corpus selection and extraction: we selected to first study decisions by the CJEU as published in the EUR-Lex database. While we would like to extend our investigation to other case law corpora in the future, we focus on EUR-Lex initially because: 1) EUR-Lex judgements are translated into English (unlike many national case law databases), which provided our analysis team with a *lingua franca* through which to interpret and communicate the results of the text similarity algorithms, 2) While databases such as HUDOC also provide English translations of cases, EUR-Lex cases can be downloaded directly from their webpage in both XML and HTML formats which are more readily processable with software tools (as opposed to HUDOC cases available in PDF and Microsoft Word format). We extracted the full texts of all judgements and *orders* (abridged judgements) from EUR-Lex / CELLAR (the central data store of the EU publications office). We did this for all decisions until December 2018 (according to their document dates). We excluded decisions from the General Court, Civil Service Tribunal and Court of First Instance. This gave us a corpus of 13,828 decision texts in total across various topics. In addition to the full texts, we also extracted the citations (exclusively to other CJEU judgements and orders) and *subject matters* for each case, as reported in the metadata published on the EUR-Lex webpage for the case. Subject matters are keywords denoting legal topics that a case deals with (the topics are part of a classification system for EUR-Lex documents aligned with the evolution of EU policies). Details about how the extracted information is stored, published and licensed (for further research) is found in Section 6.

Case sampling strategy: analysing all 13,828 CJEU cases would require in the region of 95 million similarity checks for each algorithm that we evaluate (*n choose k* where n=13,828 and k=2). We therefore elected to focus on a sample subset of the CJEU CDCN. To be representative of the CJEU cases, we chose to sample variance in the citation frequency of a case (to avoid bias). For the selection of topics, there is an option to perform a similar sampling across the case topic distribution in the CJEU corpus. However, while the advantage of this approach gives us a sample that contains a broad variety of topics, it also presents a challenge. This is because we would like to generate human interpretable visualisations of the CDSNs. If we have many topics within a particular visualisation, it is more challenging to represent all of them in the CDSN while still retaining a graphical representation of the CDCN in which patterns are self-evident. Therefore, we selected three topics of cases for our evaluation based on their currently heightened societal relevance: 1) Data protection, 2) Social policy and 3) Public health. Extracting all cases concerning these topics, we had 42, 707 and 181 for data protection, social policy and public health, respectively. We calculated sampling size based on population size and margin of error. Selecting a sampling error of 10% and confidence of 95%, resulted in a sample size of 63, 85 and 29 cases for each topic, respectively. To ensure that we sample cases uniformly across citation frequency, we sorted them by number of citations. We then partitioned them into N quantiles equidistant from each other, where N is the sample size for the case topic. The cases located at each quantile then serve as the sample cases for our analysis.

Selection of text similarity measures: Text similarity algorithms generally fall into two broad categories: *syntactic* and *semantic* [8]. Syntactic measures are generally based on calculating and comparing the frequency of characters or words between texts. Semantic measures provide mechanisms to take into account context of words within the text - i.e., their neighbouring words. For this initial study, we chose to evaluate three methods in each category. For syntactic measures, we elected to evaluate *Term Frequency - Inverse Document Frequency* (TF-IDF) [20], *Jaccard distance*, and *N-grams* (N=5). For the N-grams method, we found that the overlap of similarity links and citation links in the CDCN continues to increase until N=5. Thereafter, the overlap starts to drop (hence we choose N=5). TF-IDF and N-grams provide a method for vectorising the CJEU case texts into *document vectors*. In order to measure similarity of documents, we need a vector distance measure. We elected to use the popular *cosine similarity* distance measure for these two methods. The only preprocessing applied to the texts was removal of *stop words*. The stop words removed were a combination of: 1) the set of all English language stop words available in the Natural Language Toolkit Python library (https://www.nltk.org), and 2) the set of words that occur most frequently in the case texts (those appearing in at least 90% of the documents), and 3) a selection of words and phrases which were identified by legal researchers as particular to the corpus (e.g. “Court of Justice”). For the semantic measures, we chose to implement *word embeddings* [21] as the primary means to vectorise the texts. In order to gain insight into the question of whether general or domain-specific word embeddings are more successful, we used three types: 1) a general model pre-trained on news articles - the *GoogleNews* vectors (https://code.google.com/archive/p/word2vec), 2) a more specialised model pre-trained on legal documents from the EU (including EUR-Lex) and the US, called *Law2Vec* [22], and 3) a model trained by us on all EUR-Lex judgements and orders until December 2018. We shall refer to these models in the sequel as the *GoogleNews, Law2Vec* and *CJEU* embeddings, respectively. Our CJEU embeddings were trained using the following steps: Firstly, we removed stopwords from each case in the corpus of 13,828

cases. We then used the Word2Vec model implementation offered by the Gensim (`https://radimrehurek.com/gensim`) Python library in order to train the word embeddings. We varied the following parameters: 1) the vector dimension size (2^n where n=[5,9]), the number of training epochs (increments of 5 from 5-50), the *window size* (increments of 5, from 5-20). Window size refers to the number of words to the left and right of a word in the text that the embedding model should consider as its "context". For vector size, we tried dimensions that are powers of 2 to speed up training time by making efficient use of memory. We found a vector size of 256, number of training epochs of 30, and window size of 5 for the CJEU embeddings provided highest overlap size with the CDCN. Hence, this is the model reported in the sequel. In terms of document distance measures, we considered two measures: cosine similarity and *word mover's distance* (WMD) [23], the latter has given state-of-the-art performance for various applications. WMD can only be calculated with word vectors and therefore cannot be used for TF-IDF and N-gram, which use *document* vectors.

Evaluation setup: in summary, we selected three syntactic text similarity measures for the evaluation: Jaccard distance, TF-IDF and N-grams (N=5), the latter two methods are applied with cosine similarity to calculate document similarity. For semantic measures, we selected three word embedding models: GoogleNews, Law2Vec and CJEU embeddings. With each of these models, we applied cosine similarity and WMD to calculate document similarity. This gives us nine methods in total for the evaluation. For each of our sample cases in each topic, we calculate the top 20 similar cases to it (according to the given method). The motivation for choosing 20 as an upper bound for the size of the similarity list is that we found 99% of CJEU judgements and orders in our corpus of 13,828 to have fewer than 21 citations (with a mean of 4.2). Computing the top 20 similar cases thus gives the algorithms the theoretical possibility to capture all the citations for 99% of the cases. While there are cases in the other 1% which have up to 55 citations, it would be computationally infeasible for us to compute the top 55 similar cases for all the sample cases, using all the algorithms. For each similarity link computed by the algorithms, we check in the CDCN (for the sample cases) if there is a citation link between these same cases. If there is a citation link, we count it as an overlap. We record the overlap counts per case, per case topic and per algorithm. The CDCN for the sample cases is defined as the subset of the full CDCN that contains only the sample cases and their *direct* citations (one link). We do not include links with a length of more than one in this initial study.

4. Results

The results of the overlap, which contribute towards the main research question of the study, are depicted in Figure 1:

			Total Overlap in Percentage of Sampled Cases			
Similarity Type	Similarity Method	Vectorization Method	Data Protection	Public Health	Social Policy	3 Topics Together
Syntactic	Cosine Similarity	N-grams (N=5)	38,1%	40,4%	39,9%	39,6%
Syntactic	Jaccard Distance	N/A	38,1%	27,2%	37,4%	35,0%
Syntactic	Cosine Similarity	TF-IDF	37,2%	22,3%	38,5%	34,2%
Semantic	Word Mover's Distance	Law2Vec Embeddings	6,5%	9,8%	20,1%	14,6%
Semantic	Word Mover's Distance	GoogleNews Embeddings	7,8%	7,9%	20,1%	14,4%
Semantic	Word Mover's Distance	CJEU Embeddings	3,0%	7,5%	15,6%	10,9%
Semantic	Cosine Similarity	CJEU Embeddings	3,9%	3,8%	4,7%	4,3%
Semantic	Cosine Similarity	Law2Vec Embeddings	1,7%	3,8%	3,0%	2,9%
Semantic	Cosine Similarity	GoogleNews Embeddings	2,2%	2,3%	3,1%	2,7%

Figure 1. Percentage overlap of the similarity links in the CDSNs with the citation links in the CDCN.

Figure 1 demonstrates that the overlap remains fairly consistent across the three topics and that we reach a 40% overlap in the best case. N-grams with N=5 proved to be the method with the largest overlap. It is a surprise that syntactic measures far outperform semantic measures. We also observed that though the semantic measures have far lower overlap with the CDCN, they do find overlaps which the syntactic measures miss. To be precise, 12% and 21% of the WMD and cosine similarity overlaps, respectively, are missed by the syntactic measures. There is also, interestingly, only an overlap of 13% between cosine similarity and WMD (see Figure 2).

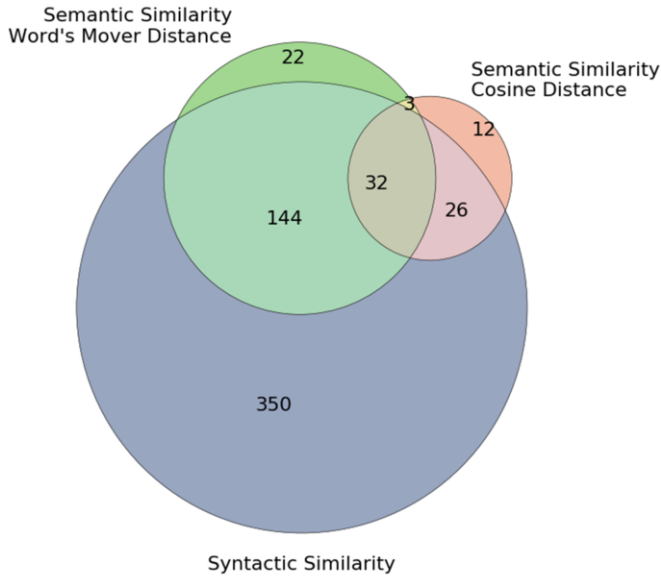


Figure 2. Venn diagram showing the degree of consensus among the algorithm categories concerning the CDSN and CDCN overlap.

It is a surprise that syntactic measures perform better because it was hypothesised that ambiguity in meaning would be an important factor in legal text. For example, the

words 'violation' and 'infringement', although semantically related, are syntactically distinct.

Semantic measures would still recognise this relationship, while syntactic measures do not. The CDSNs for the three methods having highest overlap with the CDCN are plotted in Figure 3 below:

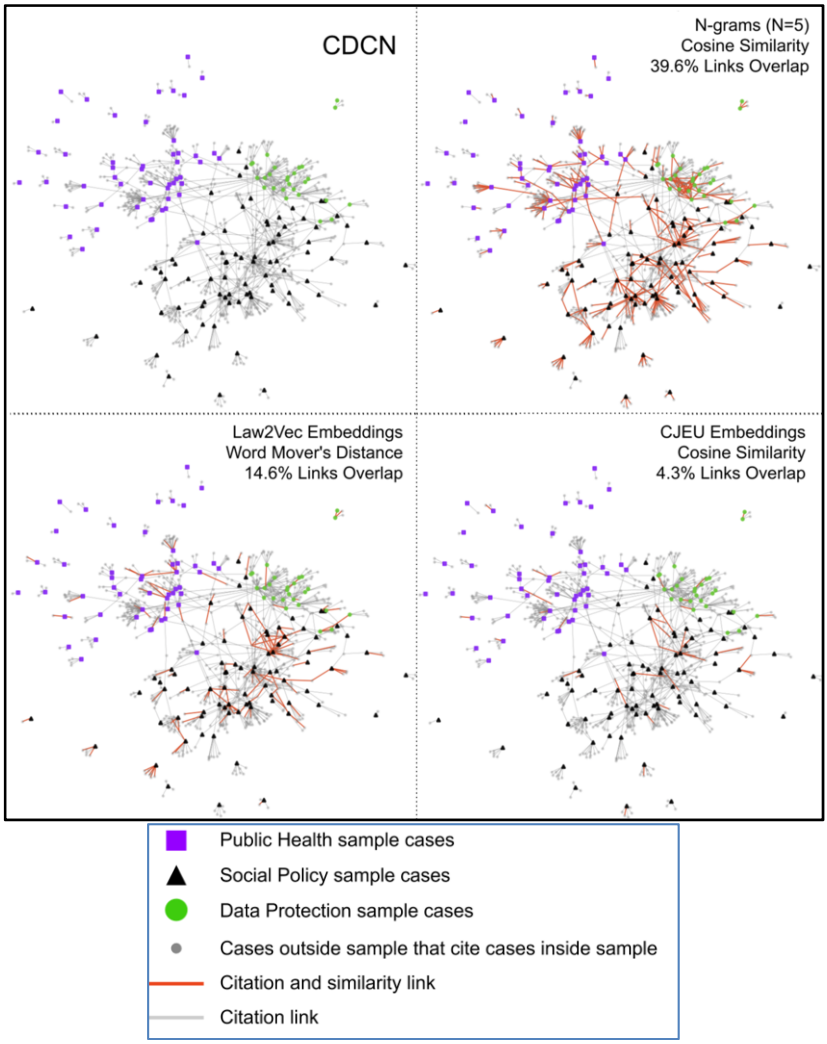


Figure 3. Visualisation of the CDSNs having the highest size of overlap with the sample cases CDCN (the best syntactic, cosine similarity and WMD methods are included).

Observing the difference between the networks for the CJEU and Law2Vec embeddings in Figure 3, we notice that there is very little improvement in overlap for the data protection cases. We also notice that the CDCN has a substantial number of cross-topic citations. It is confirmed for all methods that text similarity does not perform well at capturing these citations (most likely because the cases would be textually dissimilar,

reflecting their different legal topics). Another surprising finding is that there was no significant difference in performance between general word embeddings (trained on news articles) and those trained on legal text (Law2Vec). Cosine similarity was found to be a poor measure of case similarity (in the sense that agrees with the CDCN). WMD is a substantial improvement on cosine similarity but still far behind the performance of the syntactic measures. We also performed an analysis to verify the hypothesis that, given a case A, and two cases B and C which are similar to A with almost the same degree, the one which A will cite will generally be the more cited one. We found that similar cases that are also cited have on average 2 to 8 more citations than those that are not - regardless of the similarity score. We also found that, within the top 20 similarity list for each sample case, the probability of overlap with the CDCN is highest for the 8th similar case (on average for all algorithms). If we examine the individual methods, we find one outlier - Jaccard distance - which has the highest probability of overlap with the 13th similar case. Jaccard is also the outlier in terms of variance in where the overlap lies on the similarity list. 75% of the overlaps are found in the top 18 similar cases and 50% within the top 15. The results are slightly better for cosine similarity with top 13 and top 8 respectively. However, the most reliable method was WMD with 75% of overlaps coming in the top 10 and 50% within the top 5 respectively.

5. Challenges & Limitations

One of the main limitations of the study is that we only consider three legal topics. It remains an open question about whether these results would generalise to other topics. Another caveat is that we only compare similarity links with *direct* citations from the CDCN. In general, there may be multiple indirect paths between two nodes in the CDCN, and these paths could still capture relevance between cases. Because we don't capture these links, our calculated overlap sizes (Figure 1) represent a *conservative lower bound* on the actual number of overlaps. Nevertheless, it remains unclear what the maximum length of a path should be to still capture relevance between nodes.

It also remains an open question of how close we could ever get to reconstructing the citation network (purely from the content of court decisions). Some reasons include: not all court decisions are published online; not all relevant information about a case are published in the text; while the text does provide the legal arguments, topics and principles used in the case, it will often not depict *tacit* knowledge, information about the socio-economic and political climate in which the case was decided, nor the peripheral information about the parties involved; CJEU cases are substantively different from other court decisions in that they deal with fundamental EU law. E.g. two cases about free movement of goods can be textually quite different (one could be about wine and another about electrical appliances) but they might be similar in terms of related EU legislation concerning transportation of goods.

While we do not preprocess the texts (other than elimination of stopwords), this is more of a caveat than a limitation. The reason is that we plan to arrive at a computational signature for relevance that would be maximally explainable from an intuitive standpoint. We deliberately start with a naive implementation of algorithms so that they can be incrementally optimised systematically, thereby constructing a minimum viable algorithm. Finally, we adopted *citation* as the notion of relevance. However, this overlooks other notions of relevance (e.g. where cases are substantively related but the judge forgot to include a citation between them).

6. Conclusions & Future Work

We have presented an evaluation of selected state-of-the-art text similarity algorithms w.r.t. their ability to approximate relevance as captured by the CJEU citation network. We learned that we can approximate the CJEU citation network (at least for data protection, social policy and public health cases) using these algorithms with a completeness of up to 40%, with little to no preprocessing of the texts, and optimisation of the base algorithms. We also found that syntactic measures perform three times better than semantic measures overall for this task. Surprisingly, general word embeddings (GoogleNews) performed just as well as legal text word embeddings for the same task, while cosine similarity, as a document distance measure, performed poorly. We also observed that Word Mover's Distance was the most "consistent" document distance measure overall in that 75% of its overlapping cases came from the top 10-11 of its similarity list, and half of them came from the top 5. This is in contrast to all other methods tested, which had significantly more variance in the degree of textual similarity of the overlapping cases. Unsurprisingly, we also confirmed the generally acknowledged hypothesis that the higher the citation frequency of a case, the more likely it is to be cited. This was done by comparing the citation frequency of similar cases that are involved in a citation link vs. those that are not.

Our next steps will be to extend the study to understand if the findings we obtained generalise to: 1) other legal topics for cases in the CJEU network, and 2) other court decision corpora (e.g. ECHR decisions). We also plan to evaluate additional text similarity measures (both semantic and syntactic). From the semantic perspective, the recent *siamese networks* [25] appear to be promising, as well as the *Latent Dirichlet Allocation* (LDA) and *Latent Semantic Analysis* (LSA) methods for identifying abstract topics from text. Further syntactic approaches include *Dice's coefficient* and *Manhattan distance.* Finally, in this work, we adopted the notion of relevance captured by the CDCN. However, the presumption that citation frequency and centrality in the CDCN is a necessary condition for case relevance, is questionable if consistency of decision-making is the aim. Therefore, we would like to explore other definitions of relevance in future. One possible way to define relevance is to ask legal scholars which fragments of information in a case are most important to decide relevance. This information can be made machine processable through text annotation. We hope that these studies lead us closer to more reliable *computational signatures* of relevance for court decisions.

In the interests of promoting reproducibility, we have made all the data and software used to conduct our evaluation publicly available and accessible at the following digital object identifier (DOI) - (`http://doi.org/10.17605/OSF.IO/REBQX`). It is released under the GNU General Public License (GPL) v3.0 (`https://www.gnu.org/licenses/gpl-3.0.en.html`) which allows the distribution, modification and commercial use of the resources. However, it requires that all modifications made should be clearly stated, all source code for resulting works should be disclosed, and these works should also be released under the same license. The FAIR principles for data management [24] also advocate the interoperability and reusability of digital resources. Towards this, we have tried to document the resources we have produced in a manner that enables easier reproducibility of the study. We have used widely supported, platform-independent, data formats (CSV) and software standards (Python language with required libraries documented).

Jupyter Notebooks (`https://jupyter.org`) are also used to enable inline documentation, plots and segmented running of code.

Acknowledgements. The authors would like to thank Seun Adekunle and Andreea Grigoriu of Maastricht University for helpful discussions and inputs to the methods.

7. References

[1] J. Fowler et al. (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15(3), 324-346.
[2] Y. Lupu and E. Voeten (2012). Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights. *British Journal of Political Science*, 42(2), 413-439.
[3] M. Derlén et al. (2014). Goodbye Van Gend En Loos, Hello Bosman? Using Network Analysis to Measure the Importance of Individual CJEU Judgments. *European Law Journal*, 20(5), 667-687.
[4] D. van Kuppevelt and G. van Dijck (2017). Answering Legal Research Questions About Dutch Case Law with Network Analysis and Visualization. In *JURIX*, 302, 95-100, IOS Press.
[5] M. van Opijnen and C. Santos (2017). On the Concept of Relevance in Legal Information Retrieval. *Artificial Intelligence and Law*, 25(1), 65-87.
[6] R. Mihalcea, C. Corley and C. Strapparava (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *AAAI*, 775-780, AAAI Press.
[7] A. Islam and D. Inkpen (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery From Data*, 2(2), 10.
[8] W. H. Gomaa and A. F. Aly (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13), 13-18.
[9] O. Shulayeva, A. Siddharthan and A. Wyner (2017). Recognizing Cited Facts and Principles in Legal Judgements. *Artificial Intelligence and Law*, 25(1), 107-126.
[10] M. van Opijnen (2013). A Model for Automated Rating of Case Law. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, 140-149, ACM.
[11] S. Malmgren (2011). Towards a Theory of Jurisprudential Relevance Ranking. Using Link Analysis on EU Case Law, Graduate thesis, Stockholm University.
[12] A. Geist (2009). Using Citation Analysis Techniques For Computer-Assisted Legal Research In Continental Jurisdictions, Graduate thesis, University of Edinburgh.
[13] M. van Opijnen (2012). Citation Analysis and Beyond: in Search of Indicators Measuring Case Law Importance. In *JURIX*, 250, 95-104, IOS Press.
[14] A. Minocha, N. Singh and A. Srivastava (2015). Finding Relevant Indian Judgments Using Dispersion of Citation Network. In *WWW*, 1085-1088, ACM.
[15] R. Winkels et al. (2011). Determining Authority of Dutch Case Law. In *JURIX*, 235, 103-112, IOS Press.
[16] K. Sugathadasa et al. (2018). Legal Document Retrieval Using Document Vector Embeddings and Deep Learning. In *Science and Information Conference*, 160-175, Springer.
[17] K. Raghav, P. K. Reddy and V. B. Reddy (2016). Analyzing the Extraction of Relevant Legal Judgments Using Paragraph-level and Citation Information. *AI4JCArtificial Intelligence for Justice*, 30.
[18] Y. Panagis et al. (2017). Giving Every Case Its (Legal) Due. The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union Law. In *JURIX*, 302, 59-68, IOS Press.
[19] A. Tversky (1977). Features of Similarity. *Psychological Review*, 84(4), 327.
[20] J. Ramos (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 242, 133-142.
[21] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3, 1137–1155.
[22] I. Chalkidis and D. Kampas (2019). Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora. *Artificial Intelligence and Law*, 27(2), 171-198.
[23] M. Kusner, Y. Sun, N. Kolkin and K. Weinberger (2015). From Word Embeddings to Document Distances. In *International Conference on Machine Learning*, 957-966.
[24] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne and J. Bouwman (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3.
[25] J. Mueller and A. Thyagarajan (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In *AAAI*, 2786-2792, AAAI Press.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190308

Comparing Alternative Factor- and Precedent-Based Accounts of Precedential Constraint

Henry PRAKKEN
Department of Information and Computing Sciences, Utrecht University, and Faculty of Law, University of Groningen, The Netherlands

Abstract. In this paper several existing dimension-based models of precedential constraint are compared and an alternative is proposed, which unlike existing models does not require that for each value assignment to a dimension it is specified whether it is for or against the case's outcome. This arguably makes the model easier to apply in practice. In addition, it is shown how several factor- and dimension-based models of precedential constraint can be embedded in a Dung-style argumentation-based form, so that general tools from the formal study of argumentation become applicable.

Keywords. case-based reasoning, precedential constraint, factors, dimensions

1. Introduction

In the formal study of legal case-based reasoning dimensions (relevant aspects of a case that can have multiple values) have received increasing attention [4,10,7]. Much of this work concerns the idea of *precedential constraint*, that is, the question under which conditions a decision in a new case is determined by a set of precedents. One aim of this paper is to compare and assess existing dimension-based models of precedential constraint and to propose an alternative. The alternative is motivated by the observation that the requirement of existing models to specify for each value assignment to a dimension in a case whether it is for or against the case's outcome is often hard to apply in practice. Instead, I will propose a model in which all that needs to be specified is which change in value favours one outcome more and the other outcome less.

A second aim of this paper is to show how both factor- and dimension-based models of precedential constraint can be embedded in a Dung-style [5] argumentation-based form, so that general tools from the formal study of argumentation become applicable. Earlier similar attempts were [9,2], which formulated argument schemes for case-based reasoning with factors or dimensions in the context of the *ASPIC*$^+$ framework Unlike this work, I will model case-based reasoning 'stand-alone', without embedding in a more general theory of the structure of arguments and the nature of their relations. This will allow me to focus clearly on the essence and to remain close to relevant work of others.

Below I will, after presenting the formal preliminaries in Section 2, first reconstruct Horty's [6] factor-based result and reason models of precedential constraint as Dung-

style argumentation. The key idea is to define a similarity relation on precedents given a case to-be decided, and to use this relation to resolve attacks between arguments in an abstract argumentation framework. Then I will in Section 4 first adapt this embedding for Horty's [7] dimension-based result model, and then do the same for a dimension-based reason model inspired by Rigoni's [10] critique of Horty's model and for an alternative model addressing the pragmatic concerns with the Rigoni-style approach.

2. Formal Preliminaries

I first summarise the formal frameworks used in this paper. An *abstract argument framework* [5] is a pair $AF = \langle \mathcal{A}, attack \rangle$, where $\mathcal{A}$ is a set of arguments and *attack* a binary relation on $\mathcal{A}$. A subset $\mathcal{B}$ of $\mathcal{A}$ is *conflict-free* if no argument in $\mathcal{B}$ attacks an argument in $\mathcal{B}$ and it is *admissible* if it is both conflict-free and also defends itself against any attack, i.e., if an argument A_1 is in $\mathcal{B}$ and some argument A_2 not in $\mathcal{B}$ attacks A_1, then some argument in $\mathcal{B}$ attacks A_2. The theory of *AFs* identifies sets of arguments (called *extensions*) which are all admissible but may differ on other properties. For present purposes their differences do not matter much. What suffices is that the so-called *grounded extension* is always unique and thus captures a notion of 'justified arguments', i.e., those arguments that either directly or indirectly survive all attacks. Moreover, membership can be tested with an argument game between a proponent and an opponent of a given argument. The game starts with the proponent moving the argument to be tested and the players take turns after each argument: the opponent must attack the proponent's last argument while the proponent must one-way attack the opponent's last argument (in that the attacked argument does not in turn attack the attacker). A player *wins an argument game* iff the other player cannot move. An argument is *justified* (i.e., in the grounded extension) iff the proponent has a winning strategy in a game about the argument, i.e., if the proponent can make the opponent run out of moves in whatever way the opponent plays.

I next recall some notions concerning factors and cases often used in AI & law (e.g. in [6,10,7]), although with some differences in notation. Let o and o' be two outcomes and *Pro* and *Con* be two disjoint sets of atomic propositions called, respectively, the *pro*- and *con* factors, i.e., the factors favouring, respectively, outcome o and o'. The variable s (for 'side') ranges over $\{o, o'\}$ and $\overline{s}$ denotes o' if $s = o$ while it denotes o if $s = o'$. We say that a set $F \subseteq Pro \cup Con$ favours side s (or F is pro s) if $s = o$ and $F \subseteq Pro$ or $s = o'$ and $F \subseteq Con$. For any set F of factors the set $F^s \subseteq F$ consists of all factors in F that favour side s. A *fact situation* is any subset of $Pro \cup Con$.

The notion of a case can be defined in two ways. If all factors of a case c are supposed to be relevant to its outcome (as in Horty's [6] *result model* of precedential constraint), then it can be represented as a triple $(pro(c), con(c), outcome(c))$ where $outcome(c') \in \{o, o'\}$. Moreover, if $outcome(c) = o$ then $pro(c) \subseteq Pro$ and $con(c) \subseteq Con$ and if $outcome(c) = o'$ then $pro(c) \subseteq Con$ and $con(c) \subseteq Pro$. If, by contrast, a subset of the set of factors favouring a case's outcome can be sufficient for its outcome (as in Horty's [6] *reason model* of precedential constraint), then a case can be represented as a triple $(ppro(c) \cup con(c), pro(c), outcome(c))$, where $pro(c) \subseteq ppro(c)$ and where the above constraints on $pro(c)$ also hold for $ppro(c)$ (the factors 'potentially pro' c's outcome) and the other conventions and constraints are as above. Horty calls $pro(c)$ the 'rule' of the case. It consists of those pro-decision factors that according to the decision maker are jointly sufficient to outweigh all the con-decision factors in the case.

Given all this, a case base CB is a set of cases. Below I assume it clear from the context whether cases are represented for the result model or for the reason model.

I next summarise Horty's [6] result model of precedential constraint.

Definition 2.1 [Preference relation on fact situations.] Let X and Y be two fact situations. Then $X \leq_s Y$ iff $X^s \subseteq Y^s$ and $Y^{\overline{s}} \subseteq X^{\overline{s}}$.

$X <_s Y$ is defined as usual as $X \leq Y$ and $Y \not\leq X$. This definition says that Y is at least as good for s as X iff Y contains at least all pro-s factors that X contains and Y contains no more pro-$\overline{s}$ factors than X contains.

Definition 2.2 [Precedential constraint with factors: result model] Let CS be a case base and F a fact situation. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (X, Y, s)$ in CB such that $X \cup Y \leq_s F$.

I finally summarise Horty's [6] reason model of precedential constraint. The following definition says that a case decision expresses a preference for any pro-decision set containing at least the pro-decision factors of the case over any con-decision set containing at most the con-decision factors of the case. This allows *a fortiori* reasoning from a precedent adding pro-decision factors an/or deleting con-decision factors.

Definition 2.3 [Preferences from cases.] Let $(ppro(c) \cup con(c), pro(c), s)$ be a case, CB a case base and X and Y sets favouring $\overline{s}$ and s, respectively. Then

1. $Y <_c X$ iff $Y \subseteq con(c)$ and $X \supseteq pro(c)$;
2. $Y <_{CB} X$ iff $Y <_c X$ for some $c \in CB$.

Definition 2.4 [(In)consistent case bases.] Let C be a case base with $<_{CB}$ the derived preference relation. Then CB is *inconsistent* if and only if there are factor sets X and Y such that $X <_{CB} Y$ and $Y <_{CB} X$. And CB is *consistent* if and only if it is not inconsistent.

The final definition says that deciding a case for a particular outcome is forced if that is the only way to keep the updated case base consistent.

Definition 2.5 [Precedential constraint with factors: reason model.] Let CB be a consistent case base and (F, R, s) a case that is not in CB. Then, given CB, c is *allowed* iff $CB \cup \{c\}$ is consistent. Moreover, deciding F for s is *forced* iff for all cases $c = (F, R, outcome(c))$ it holds that $CB \cup \{c\}$ is consistent iff $outcome(c) = s$.

Horty [6] proves that his result and reason model are equivalent on the assumption that $pro(c) = ppro(c)$ for all cases c.

3. An Argumentation-Based Model of Precedential Constraint with Factors

In this section I define a similarity definition on the set of cases given a focus case (a case to be decided), to be used to resolve attacks between arguments in an argumentation framework. I then prove a relation between this similarity definition and Horty's factor-based models of precedential constraint. It suffices for this purpose to look at the relevant differences between a precedent and the focus case, which are those differences that are

a reason not to decide the focus case as the precedent. These are the situations in which a precedent can be distinguished in a HYPO/CATO-style approach [1], namely, when the new case lacks some features of the pro its outcome that are in the precedent or has new features con its outcome that are not in the precedent. Here it is relevant whether the two cases have the same outcome or different outcomes.

Definition 3.1 [Differences between cases with factors.] Let c and f be two cases. The set $D(c,f)$ of differences between c and f is defined as follows.

1. If $outcome(c) = outcome(f)$ then $D(c,f) = pro(c) \setminus pro(f) \cup con(f) \setminus con(c)$.
2. If $outcome(c) \neq outcome(f)$ then $D(c,f) = pro(f) \setminus con(c) \cup pro(c) \setminus con(f)$.

Intuitively, the fewer (with respect to set inclusion) the relevant differences between a case in the case base and the focus case are, the better it is. Below I formalise this by using the subset relation on sets of relevant differences with the focus case as a preference relation in an abstract argumentation framework in which arguments are cases.

Definition 3.2 [Case-based argumentation frameworks.] Given a case base CB and a focus case $f \notin CB$, an abstract argumentation framework $AF_{CB,f}$ is a pair $\langle \mathcal{A}, attack \rangle$ where:

- $\mathcal{A} = CB$;
- c attacks c' iff $outcome(c) \neq outcome(c')$ and $D(c',f) \not\subset D(c,f)$.

The idea is that a given fact situation F must be decided for s just in case there exists a justified argument for outcome S on the basis of the $AF_{CB,f}$ where $f = (F,s)$. So moving an argument in the grounded game is elliptic for 'the fact situation of the focus case must be decided as in this precedent since they are sufficiently similar'.

I next establish a formal relation between Horty's reason model of precedential constraint and the above argumentation-based reconstruction. Since Horty's result model is a special case of his reason model, this result also holds for the result model.

Proposition 3.3 Let $AF_{CB,f} = \langle \mathcal{A}, attack \rangle$ be an abstract argumentation framework defined by a consistent case base CB and a focus case f with fact situation F. Then deciding F for s is forced given CB iff there exists a case c with outcome s in CB such that $D(c,f) = \emptyset$.

Proof: Assume first that f is forced. Let $f = (F^s \cup F^{\bar{s}}, R, s)$. Then every case $f' = (F^{\bar{s}} \cup F^s, R', s)$ is inconsistent with the case base. Let $R' = F^{\bar{s}}$. Then since CB is consistent, by Observation 1 of [7] there exists a case $f'' = (X \cup Y, R'', s) \in CB$ such that $R'' <_{f'} F^{\bar{s}}$ and $F^{\bar{s}} <_{f''} R''$. The former priority entails that $R'' \subseteq Fs$. But then $pro(f'') \subseteq pro(f)$, so (1) $pro(f'') \setminus pro(f) = \emptyset$. The latter priority entails that $F^{\bar{s}} \subseteq Y$. But then (2) $con(f) \subseteq con(f'')$ so (2) $con(f) \setminus con(f'') = \emptyset$. Then observe that (1) and (2) together entail that $D(f'',f) = \emptyset$.

Assume next that there exists a $c \in CB$ with outcome s and such that $D(c,f) = \emptyset$. Then we have $con(c) <_c pro(c)$ and we have $pro(c) \subseteq pro(f)$ and $con(f) \subseteq con(c)$. But then we also have $con(f) <_c pro(f)$, so for every $R \subseteq con(f)$ we have $R <_c pro(f)$ and so $R <_c ppro(f)$. Any rule for deciding the facts of f for $\bar{s}$ requires adding a case $c' = (F,R,\bar{s})$ to CB but then $ppro(f) <_c R$ can be derived from CB, so CB is inconsistent. Moreover, this immediately implies that any case $f = (F,R,s)$ is consistent with CB. □

This proposition yields a simple syntactic criterion for determining whether a decision is forced. More generally it embeds Horty's models of precedential constraint in the formal theory of abstract argumentation. At present this embedding is still somewhat trivial, since an immediate consequence of Proposition 3.3 is that (assuming CB is consistent) deciding the fact situation of a focus case f for its outcome is forced iff there is a case in C for the same outcome that has no attackers in $AF_{CB,f}$. So dialogues for this case in the grounded game are trivial in that they stop after the proponent moves this case. However, there are ways to extend the present setup to yield more interesting dialogues, which can be explored in future research. One extension is with preferences between factors, so that cases with relevant differences could also be forced. Even more interesting is if these factor preferences can be argued for or if factors can be derived with further arguments.

4. Adapting the Approach to Dimensions

In this section I discuss various ways in which the above approach can be adapted to dimensions. I first show how Horty's [7] dimension-based result model can be embedded in an argumentation framework. I will not do the same for his dimension-based reason model, for two reasons. First, as Horty shows, his dimension-based reason model collapses into his result model, which arguably fails to capture the distinction between *ratio decidendi* and *obiter dicta* from common-law jurisdictions. Second, I agree with Rigoni [10] that Horty's model sometimes yields counterintuitive outcomes. For these reasons I will first formulate a reason model adapting ideas of Rigoni [10] and then present an alternative reason model motivated by some pragmatic concerns about Rigoni's approach.

4.1. Horty's Dimension-Based Result Model as Argumentation

I adopt from [7] the following technical ideas (again with some notational differences). A *dimension* is a tuple $d = (V, \leq_o, \leq_{o'})$ where V is a set (of values) and $\leq_o$ and $\leq_{o'}$ two partial orders on V such that $v \leq_o v'$ iff $v' \leq_{o'} v$. A *value assignment* is a pair (d, v). The functional notation $v(d)$ denotes the value of dimension d. Then a *case* is a pair $c = (F, outcome(c))$ such that D is a set of dimensions, F is a set of value assignments to all dimensions in D and $outcome(c) \in \{o, o'\}$. Then a case base is as before a set of cases, but now explicitly assumed to be relative to a set D of dimensions in that all cases assign values to a dimension d iff $d \in D$ (an assumption also made by Horty). Likewise, a fact situation is now an assignment of values to all dimensions in D. As for notation, $v(d, c)$ denotes the value of dimension d in case c. Finally, $v \geq_s v'$ is the same as $v' \leq_s v$.

From now on I will use as a running example the fiscal-domicile example introduced in [8] and also used by [4,10,7]. The issue is whether the fiscal domicile of a person who moved abroad for some time has changed. Let us consider two dimensions d_1, the duration of the stay abroad in months and d_2 the percentage of the tax-payer's income that was earned abroad during the stay. For both values, increasingly higher values increasingly favour the outcome *change* and decreasingly favour the outcome *no change*. So, for instance, $(d_1, 12m) <_{change} (d_1, 24m)$ and so $(d_1, 24m) <_{no\ change} (d_1, 12m)$. An example of a fact situation is $F = \{v(d_1) = 30m, v(d_2) = 60\%\}$ and an example of a case is $c = (F', change)$ where $F' = \{v(d_1) = 12m, v(d_2) = 60\%\}$.

In Horty's result model a decision in a fact situation is forced iff there exists a precedent c for that decision such that on each dimension the fact situation is at least as

favourable for that decision as the precedent. He formalises this idea with the help of the following preference relation between sets of value assignments.

Definition 4.1 [Preference relation on dimensional fact situations.] Let F and F' be two fact situations with the same set of dimensions. Then $F \leq_s F'$ iff for all $(d,v) \in F$ and all $(d,v') \in F'$ it holds that $v(d) \leq_s v'(d)$.

In our running example we have that $F' <_{change} F$ since F and F' are equal on d_2 while F is better for s on d_1.

Then adapting Horty's factor-based result model to dimensions is straightforward.

Definition 4.2 [Precedential constraint with dimensions: result model.] Let CS be a case base and and F a fact situation given a set D of dimensions. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (F', s)$ in CB such that $F' \leq_s F$.

So in our running example deciding F for *change* is forced.

I next embed Horty's dimension-based result model in an argumentation framework in a similar way as I did above for his factor-based result and reason model. First Definition 3.1 of differences between cases has to be adapted to dimensions. Note that unlike in the case of factors, there is no need to indicate whether a value assignment favours a particular side in the case since the $\leq_s$ ordering suffices for this purpose.

Definition 4.3 [Differences between cases with dimensions.] Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. The set $D(c,f)$ of differences between c and f is defined as follows.

1. If $outcome(c) = outcome(f) = s$ then $D(c,f) = \{(d,v) \in F(c) \mid v(d,c) \not\leq_s v(d,f)$.
2. If $outcome(c) \neq outcome(f)$ where $outcome(c) = s$ then $D(c,f) = \{(d,v) \in F(c) \mid v(d,c) \not\geq_{\bar{s}} v(d,f)$.

Let c be a precedent and f a focus case. Then clause (1) says that if the outcomes of the precedent and the focus case are the same, then any value assignment in the focus case that is not at least as favourable for the outcome as in the precedent is a relative difference. Clause (2) says that if the outcomes are different, then any value assignment in the focus case that is not at most as favourable for the outcome of the focus case as in the precedent is a relative difference.

In our running example, let $f = (F, change)$. Then $D(c,f) = \emptyset$. If $v(d_2, F)$ is changed from 60% to 50% then $D(c,f) = \{(d_2, 60\%)\}$ by clause 1. Next, let $g = (G, nochange)$ where $G = \{v(d_1) = 24m, v(d_2) = 60\%\}$. Then $D(c,g) = \{(d_1, 12)\}$ by clause 2.

With these definitions, Definition 3.2 of an abstract argumentation framework given a case base still applies to the setting with dimensions. This allows the following counterpart of Proposition 3.3.

Proposition 4.4 Let, given a set D of dimensions, $AF_{CB,f} = \langle \mathcal{A}, attack \rangle$ be an abstract argumentation framework defined by a case base CB and a focus case f with a fact situation F. Then deciding F for s is forced given CB according to Definition 4.2 iff there exists a case in CB with outcome s such that $D(c,f) = \emptyset$.

Proof: Consider first any $c = (F(c), s)$ in CB such that $D(c,f) = \emptyset$. Then for all $(d,v) \in F(c)$ and all $(d,v') \in F(f)$ it holds that $v(d) \leq_s v'(d)$, so $F(c) \leq_s F(f)$.

Suppose next f is forced. Then the proof is the same the other way around. □

4.2. A Dimension-Based Reason Model with Complete Rules

I next discuss how Horty's dimension-based result model can be turned into a dimension-based reason model. There are two features on which this can be done: by 'relaxing' an individual value assignment or by leaving some assignments out from a set of value assignments. In both ways a case is a triple $(c = (F(c), R(c), outcome(c))$, where $F(c)$ is as in the result model a value assignment to a given set D of dimensions and where $R(c)$, the rule of the case, is a set of value assignments that is in some way constrained by $F(c)$. In the first way, rule $R(c)$ consists of value assignments to each dimension in D such that for each element (d, v) in $R(c)$ and each element (d, v') in $F(c)$ it holds that $v(d) \leq_s v'(d)$. In other words, in this approach a rule of a case assigns to each of the case's dimensions a value that is at most as favourable for the case's outcome as its value in the case. Below I will call such a rule a *complete rule*. This idea is taken from Rigoni in [10], except that he also applies it to incomplete rules.

Definition 4.5 [Precedential constraint with dimensions: a reason model with complete rules.] Let, given a set D of dimensions, CS be a case base in which all cases have a complete rule and F a fact situation. Then deciding F for s is *forced* iff there exists a case $c = (F', R, s)$ in CB such that $R \leq_s F$.

This model does not collapse into the above result model. Suppose in the tax example that c has a fact situation $(\{v(d_1) = 30m, v(d_2) = 60\%\}$ and outcome *change* and consider again fact situation $F = \{v(d_1) = 24m, v(d_2) = 75\%\}$. Suppose the court in c ruled that with a percentage earned abroad of 60% a stay abroad of at least 12 months suffices for change of fiscal domicile. The rule of c then is $\{(d_1, 12m), (d_2, 60\%)\}$. Then in the reason model deciding F for *change* is forced, even though the stay abroad in F is shorter than in c, since it is still longer than its value in c's rule. By contrast, in the result model this difference suffices to make c distinguishable and deciding F for *no change* not forced.

The model also avoids an arguably counterintuitive feature of Horty's [7] model. In our example, if the rule of c is $\{(d_1, 12m)\}$ then in a new case in which the stay abroad is 24 months and the percentage of income earned abroad is 75% deciding for *change* is in Horty's model not forced by the precedent, since it is weaker for *change* than the precedent in that the stay abroad is not 30 but 24 months. However, as also argued by Rigoni in [10], this seems counterintuitive given that the court in the precedent ruled that 12 months abroad suffice for *Change* and given that the new case is stronger for this outcome in its only other dimension. With Definition 4.5 deciding for *change* is instead forced by c, since $\{(d_1, 12m), (d_2, 60\%)\} \leq_{change} \{(d_1, 24m), (d_2, 75\%)\}$.

One issue remains: Horty's factor-based reason model requires that courts select a rule in the new case that leaves the case base consistent when the case is added to it. In Horty's (and also Rigoni's [10]) model consistency is defined in terms of a preference relation between sets of reasons pro and con a decision (cf. Definition 2.3 above). However, the present model does not distinguish between pro and con value assignments, while still a notion of consistency is needed. Consider again the tax example with the two dimensions d_1 and d_2 and consider two precedents c_1 with rule $R_1 = \{(d_1, 12m), (d_2, 60\%)\}$ and outcome *change* and c_2 with rule $R_2 = \{(d_1, 8m), (d_2, 60\%)\}$ and outcome *no change*.

Consider next a fact situation F with $d_1 = 15$ and $d_2 = 60\%$. Then deciding F for *change* is forced. Suppose the court does so but formulates the rule $R_3 = \{(d_1, 10m), (d_2, 60\%)\}$. Then in a new fact situation equal to rule R_2 both deciding *change* and deciding *no change* would be forced, so adding $f = (F, R_3, change)$ would make it inconsistent in that for the same fact situation two opposite outcomes are forced. So a constraint on rule selection should be that it should leave a consistent case base consistent in this sense.

4.3. An Alternative Dimension-Based Reason Model

The second way in which the result model can be refined into a reason model is by allowing that the rule R of a case assigns a value to a subset of its fact situation, while still adhering to the constraint that the rule's values of dimensions are at most a favourable to the case's decision as their actual values in the case. Here I would like to follow a Rigoni-style approach, in order to avoid the counterintuitive consequences of Horty's approach. However, there is a pragmatic problem here, since Rigoni requires that for each value assignment it is indicated which side it favours. The problem is that, unlike with factors, this may be hard in practice, since often this will be context-dependent (likewise [3]). In our tax example, if a case with fact situation $(\{v(d_1) = 30m, v(d_2) = 60\%\}$ has outcome *change*, are both value assignments pro this outcome, or is one pro and the other con *change*? And if the latter, then which is pro and which is con? This is not easy to say in general. On the other hand, what is uncontroversial is that increasingly higher values for these dimensions increasingly support *change* and decreasingly support *no change*. For this reason I will instead explore an approach in which all that is needed is general knowledge about which side is favoured more and which side less if a value of a dimension changes, as captured by the two partial orders $\leq_s$ and $\leq'_s$ on a dimension's values.

Below for any two sets X and Y of value assignments, $Y^{|X}$ is the subset of Y that consists of value assignments to any dimension that is also assigned a value in X.

Definition 4.6 [Precedential constraint with dimensions: an alternative reason model with possibly incomplete rules.] Let, given a set D of dimensions, CS be a case base in which all cases have a possibly incomplete rule and F a fact situation. Then deciding F for s is *forced* iff there exists a case $c = (F', R, s)$ in CB such that $R \leq_s F^{|R}$.

So deciding F for s is forced iff there is a precedent for s such that F is at least of favourable for s on all dimensions in the precedent's rule.

Moreover, like with the reason model with complete rules, the constraint on rule selection is needed that adding a new case to a consistent case base should leave the case base consistent in that for no fact situation two opposite outcomes are forced.

To see how this definition works, consider again the tax example with dimensions d_1 and d_2 and consider precedent c with fact situation $v(d_1) = 30m, v(d_2) = 60\%$, with rule $R = \{(d_1, 12m)\}$ and with outcome *change*. Consider next a fact situation F with $v(d_1) = 24m, v(d_2) = 50\%$. Then deciding F for change is forced since $F^{|R} = \{(d_1, 24m)\}$ and we have that $R = \{(d_1, 12m)\} <_{change} \{(d_1, 24m)\}$. Note that deciding F for *change* is forced by c even though F is in one dimension weaker for *change* than c, namely in d_2. The point is that d_2 is not in c's rule.

Since a rule that assigns a value to all dimensions in D is a special case, the above example that shows that Definition 4.5 does not collapse into the dimension-based result

model also holds for this definition. Moreover, a counterpart of Proposition 4.4 can be obtained for this reason model by redefining the relevant differences between a precedent and a focus case as follows.

Definition 4.7 [Differences between cases with dimensions and possibly incomplete rules.] Let $c = (F(c), R(c), \mathit{outcome}(c))$ and $f = (F(f), R(f), \mathit{outcome}(f))$ be two cases. The set $D(c, f)$ of differences between c and f is defined as follows.

1. If $\mathit{outcome}(c) = \mathit{outcome}(f) = s$ then $D(c, f) = \{(d, v) \in F(f)^{|R(c)} \mid v(d, c) \not\leq_s v(d, f)$.
2. If $\mathit{outcome}(c) \neq \mathit{outcome}(f)$ where $\mathit{outcome}(c) = s$ then $D(c, f) = \{(d, v) \in F(f)^{|R(c)} \mid v(d, c) \not\geq_{\bar{s}} v(d, f)$.

Clause (1) says that if the outcomes of the precedent and the focus case are the same, then any value assignment in the focus case to a dimension in the precedent's rule that is not at least as favourable for the outcome as in the precedent is a relative difference. Clause (2) says that if the outcomes are different, then any value assignment in the focus case to a dimension in the precedent's rule that is not at most as favourable for the outcome of the focus case as in the precedent is a relative difference.

Proposition 4.8 given a set D of dimensions, $AF_{CB,f} = \langle \mathcal{A}, \mathit{attack} \rangle$ be an abstract argumentation framework defined by a case base CB in which all cases have a complete rule and let F be a fact situation. Then deciding F for s is forced given CB according to Definition 4.6 iff there exists a case $c = (F(c), R(c), \mathit{outcome}(c))$ in CB with the same outcome as f such that for any case $f = (F, R(f), s)$ it holds that $D(c, f) = \emptyset$.

Proof: As for Proposition 4.4 with $F(c)$ replaced by $R(c)$ and $F(f)$ replaced by $F(f)^{|R(c)}$.

On the other hand, this approach also has limitations. Consider again the last example. We saw that deciding fact situation F for *change* was forced by precedent c even though F is in one dimension weaker for *change* than c, since this is not in c's rule. This prevents that a decision maker can regard the fact that the percentage of income earned abroad was less in the new situation F than in the precedent an exception to the precedent's rule. In more general terms, in the factor-based reason model the idea of a rule has a clear intuition, namely, that the pro-decision factors in the rule are sufficient to outweigh the con-decision factors in the case. However, with dimensions this intuition does not apply, since the value assignments outside the rule do not necessarily favour the opposite outcome. All that can said is that by stating the rule the court has decided that, given the rule, the case's value assignments to the other dimensions are irrelevant. The question then is whether such a ruling is defeasible. If it is not, then every new case in which the dimensions in the precedent's rule have values that are at least as favourable to the decision as in the rule is constrained by the precedent regardless of possible differences on the other dimensions. If that is regarded as too rigid, then there are two options. The first is that value assignments to dimensions not in a precedent's rule can be a reason for distinguishing just in case in the new fact situation they are less favourable for the precedent's outcome than in the precedent. But then the model collapses into the reason model with complete rules. The second option is that *every* value assignment to a dimension that is not in the rule of the case can override the case's outcome. But then the problem with Horty's reason model reappears: in our last example any income percentage,

even a percentage higher than 60%, would suffice to distinguish c. It can be concluded that a Rigoni-style approach in which value assignments are always pro a particular outcome leads to finer-grained distinctions between forced and not-forced decisions than the present approach but is arguably harder to apply in practice.

5. Conclusion

In this paper I have shown how several factor-and dimension-based models of precedential constraint can be embedded in a Dung-style setting with abstract argumentation frameworks. Thus general tools from the formal study of argumentation become available for analysing and extending these models. In addition, I have critically analysed (variants of) some existing dimension-based models of precedential constraint. I argued that a pragmatic limitation of some of them is that they require the specification of information that may be hard to obtain in practical applications and I proposed an alternative without this limitation, although also with lesser ability to distinguish between situations in which a decision is or is not forced by a body of precedents..

In future research the dropping of some limited assumptions can be investigated, such as the assumption that every case assigns a value to every dimension of a given set of dimensions. Dropping this assumption allows the introduction of new dimensions in a case but may run into the same limitations as the above alternative reason-based model. Another issue for future research is the modelling of trade-offs between dimensions with preferences and/or values, as suggested by [4]. Arguably this paper's results on the embedding in a Dung-style setting are of value here.

References

[1] K.D. Ashley. Toward a computational theory of arguing with precedents: accomodating multiple interpretations of cases. In *Proceedings of the Second International Conference on Artificial Intelligence and Law*, pages 39–102, New York, 1989. ACM Press.

[2] K.D. Atkinson, T.J.M. Bench-Capon, H. Prakken, and A.Z. Wyner. Argumentation schemes for reasoning about factors with dimensions. In K.D. Ashley, editor, *Legal Knowledge and Information Systems. JURIX 2013: The Twenty-sixth Annual Conference*, pages 39–48. IOS Press, Amsterdam etc., 2013.

[3] T.J.M. Bench-Capon. Some observations on modelling case based reasoning with formal argument models. In *Proceedings of the Seventh International Conference on Artificial Intelligence and Law*, pages 36–42, New York, 1999. ACM Press.

[4] T.J.M. Bench-Capon and K.D. Atkinson. Dimensions and values for legal CBR. In A.Z. Wyner and G. Casini, editors, *Legal Knowledge and Information Systems. JURIX 2017: The Thirtieth Annual Conference*, pages 27–32. IOS Press, Amsterdam etc., 2017.

[5] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and *n*–person games. *Artificial Intelligence*, 77:321–357, 1995.

[6] J. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33, 2011.

[7] J. Horty. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27:309–345, 2019.

[8] H. Prakken and G. Sartor. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6:231–287, 1998.

[9] H. Prakken, A.Z. Wyner, T.J.M. Bench-Capon, and K. Atkinson. A formalisation of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25:1141–1166, 2015.

[10] A. Rigoni. Representing dimensions within the reason model of precedent. *Artificial Intelligence and Law*, 26:1–22, 2018.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190309

Legal Search in Case Law and Statute Law

Julien ROSSI [a,1], Evangelos KANOULAS [a,b]
[a] *Amsterdam Business School, University of Amsterdam*
[b] *Institute of Informatics, University of Amsterdam*

Abstract. In this work we describe a method to identify document pairwise relevance in the context of a typical legal document collection: limited resources, long queries and long documents. We review the usage of generalized language models, including supervised and unsupervised learning. We observe how our method, while using text summaries, overperforms existing baselines based on full text, and motivate potential improvement directions for future work.

Keywords. Legal Search, Case Law, Statute Law, Language Models, Transfer Learning

1. Introduction

The increasing amount of legal data available requires the ability to search and identify relevant information in this data, it calls for the ability to automate or assist in specialized retrieval tasks, as a service to either the public or practitioners. Legal Search is the specialized Information Retrieval (IR) task dealing with legal information relevant to a situation described factually and legally. The legal documents collections differ from generic IR collections by the level of professional knowledge needed to produce a labeled corpus [1]. This results in limited and specialized collections, each covering a narrow field of interest: lease contracts, financial products, court decisions, etc... Existing literature shows how Legal Search is an essential tool for practitioners and citizens in need of critical information [2,3,4], whereas it introduces difficult challenges that our work addresses.

In this paper, we focus on limitations of the applicability of generalized language models [5,6,7,8] in the context of the handling of legal information tasks. We observe that specific features of typical legal documents adversely affects those models, such as long texts, mixing of formal abstract concepts with casual verbatims in layman language, or dealing with an unclear formulation of relevance.

The development of Generalized Language Models such as BERT [7] has allowed the application of complex neural models to tasks even with limited data available, significantly improving over lexical methods, while past neural models were having difficulties to deal with smaller datasets, due to vanishing gradients, and were difficult to compute and implement on hardware [9].

[1] Corresponding Author: j.rossi@uva.nl

The core of these generalized language models is the separation between learning the language (pre-training) and adapting to a task (fine-tuning). The adaptation to legal tasks is complicated by the difference between the generic language and the legalese language, a difference in word usage and meaning, information structuring, or sentence complexity [10].

Our contribution is a method to improve Legal Information Retrieval, where the ranking problem is reformulated as a pairwise relevance score problem, modeled as a fine-tuning task of a generalized language model, and where we adapt long documents to limited input sequence length by summarizing. Furthermore we show how additional pre-training based on a narrow selection of texts can improve performance.

We will focus on the following questions:

- **RQ1** *Can we use summary encoding as a dense representation of long documents ?*
- **RQ2** *Can we use pre-training and fine-tuning of neural language model with limited data to learn a specific legal language ?*

In section 2, we will introduce our methods to answer the research questions, then in section 3 the tasks we will use for our work, then the research for each task will be described with results and analysis in sections 4 and 5. We will present our global analysis and conclusion in section 6, along with motivation for future research.

2. Methods

Our work reformulates the ranking problem into a pairwise relevance classification problem, where we train a classifier to separate relevant pairs (made of a query case and a noticed case) from irrelevant pairs of documents. For each query, we let the model infer the probability for the positive class for each pair made of the query and a candidate document, and then rank the candidate documents according to that score.

This method follows after the well-established probability ranking principle in IR [11]. For the tasks we consider, we argue that the relevance of a document with regards to a given query is independent from other documents of the collection, confirming the applicability of this framework to our tasks.

Within this framework, we introduce input pre-processing as well as model training steps that will help us address the research questions. We refer to [7] for details on the topology of BERT as well as for the setup of the pre-training task or the fine-tuning task.

2.1. Pairwise Embeddings

We make use of the capacity of BERT models to encode single dense embeddings for a pair of texts. These embeddings encode the interaction between the texts in the pair, so that it is a suitable input for a downstream pair classifier. In a BERT system, this corresponds to the embeddings of `[CLS]` token, inserted at the beginning of each input sequence.

2.2. Fine-Tuning of Pre-Trained Language Model

The fine-tuning is a supervised learning task for a complete model comprising a pre-trained BERT and a fully-connected layer used for classification. The learning is driven by the true classification labels of the examples, and all the weights of the combined model are updated during this training. The usual loss for classification is the Cross-Entropy. During this phase, the combined model learns how to perform the classification based on the pairwise embeddings.

2.3. In-Domain Additional Pre-Training of Language Model

This task is an unsupervised learning task, where an already pre-trained BERT model receives further pre-training with additional texts. We make the remark that BERT is a NLP system, with a high capacity of understanding at the language level. For example [12,13] demonstrate how the layers in BERT work as a hierarchical system with a capacity to discover syntactic features of the text, while the semantic features are assumed to emerge naturally from co-location and word contexts. In that regard, legal texts create a variety of domain specific languages, considering different domains such as contracts, court decisions, legislation, etc.

Legal languages differ from casual languages with unusual vocabulary (rare words, latin words, etc.), semantics (words where the legal meaning differs from casual usage), and syntactic features. We suggest that a pre-trained model would learn specific knowledge from a limited additional pre-training on legal domain-specific texts. We base our suggestion on the fact that within a specific legal domain, the text features will be fairly uniform. Following [14], we expect pre-training on the available material being more beneficial to the capacities of the system than only fine-tuning for the task. We expect to answer RQ2 with this method.

2.4. Summarization of Long Documents

This method will allow us to address our question RQ1, on long documents. As we base our models on BERT, we are constrained by a maximum input length of 512 WordPiece [15] tokens. WordPiece will subdivide each word into multiple tokens, we consider that this will nearly double the number of tokens observed under a standard Punkt [16] tokenizer as implemented in NLTK [17].

For this reason, we introduce an extractive summarization of the texts in the corpus using TextRank [18]. TextRank is an extractive summarization model, based on a graph ranking model operating at sentence level. As it extracts full sentences, it preserves the structure of a natural language and makes the summary fit for input to a Neural Language Model. We choose to limit the size of the summary to 180 words, so a concatenated pair of texts will not be longer than 512 WordPiece tokens. In the case this limit of 512 is not respected, then the sequence will drop the last tokens from the text of candidate case. We use the implementation from gensim [19].

3. Legal Retrieval Tasks

We will illustrate our work with 2 different tasks and their associated document collection, taken from COLIEE [20]. The 2 tasks illustrate different aspects of our research:

- Case Law Retrieval: Find past cases similar to a query case. In this task, the query and the documents are both long texts, and the amount of available data is limited. We will address RQ1 and RQ2 with this task.
- Statute Law Retrieval: Find law articles relevant to a situation. In this task, the query and documents are short snippets of text, and only a small amount of labeled data is available. We will address RQ2 with this task.

3.1. Case Law Retrieval

In this task, a given court case is considered as a query to retrieve supporting cases (also named 'noticed cases') for the query case. A noticed case supports the decision taken in the query case, although the final decision itself is irrelevant in our retrieval. What matters is the proximity of the themes that are tackled by the query case and the noticed cases, either legal themes or narrative themes.

The dataset is drawn from an existing collection of Federal Court of Canada case law, provided by vLex Canada[2]. Each query case is given a collection of 200 potential supporting cases (also named 'candidate cases'), these collections are provided labeled for the training dataset, and unlabelled for the unknown test dataset. The training dataset contains 285 query cases. All cases are relative to Citizenship and Immigration proceedings.

We consider the corpus as the complete collection of query cases, and candidate cases from the training dataset, Figures 1 and 2 illustrate the distribution of document length across the corpus, measured in number of words. With a median length of 2500 words, this dataset illustrates well the challenge of dealing with long documents and long queries, as formulated in our RQ1. The corpus is a limited resource with only 285 queries, which fits well with our RQ2.

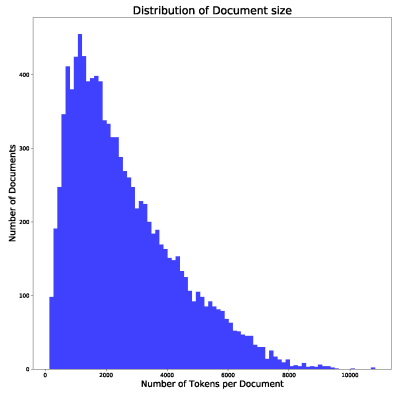


Figure 1. Distribution of the number of words per document

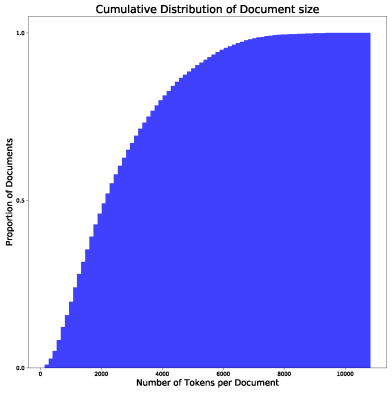


Figure 2. Cumulative distribution for the number of words per document

[2] http://ca.vlex.com

3.2. Statute Law Retrieval

In this task, a brief query has to be disputed with regards to the existing legislation, in this case limited to the Japanese Civil Code. The retrieval task identifies which articles of the Civil Code are relevant to the evaluation of the legal validity of the query.

The data for this task is compiled from Japanese Bar Exams, and provided both in Japanese and English. We focus on the English dataset from the 2018 edition. The training dataset contains 651 queries, while the test dataset is made of 69 queries. This amount of data will be a good fit for our RQ2. The queries are short snippets with an average of 40 words, with a maximum of under 120 words, while the average law article has 60 words, with 98% of articles having less than 200 words. We will not face the issue of long texts. We consider it a limited resource with few examples of relevant matches, as 94% of queries have 1 or 2 relevant articles, among around 1000 candidate articles.

3.3. Evaluation Metrics

In the COLIEE setting, the system decides for each query the number of candidates it returns, and it is evaluated based on the F1 or F2 score of that sub-list. In this paper, we will consider a more traditional approach with respect to Information Retrieval, grounded in a realistic use case for the Case Law Retrieval Task: the end user is a staff from a judge's office, and the queries are submitted to a computer software and returned as a ranked list. We use the same metrics for the Statute Law Retrieval Task.

We choose to report Precision at R (P@R), where R is the total number of relevant documents, Recall and Precision at k (R@k, P@k) and Mean Average Precision (MAP). For each metric, we will consider the macro average[3]. Metrics are computed with `trec_eval`[4], and Precision-Recall curve is plotted with `plot-trec_eval`[5]

4. Case Law Retrieval

4.1. Experimental Setup

Following the "Methods" section 2, we pre-process the collection by summarizing all texts. The final dataset is a collection of triplets made of query text, candidate text and relevance judgment. We will fine-tune a pre-trained BERT model for a binary classification task, using the relevance labels from the labeled training dataset. We will also perform additional in-domain pre-training, and then fine-tune this model for the same classification task. In the results section 4.3, the models appear under the names `FineTuned` and `PreTrained`.

For the in-domain pre-training, we use the entire corpus of court decisions as the pre-training corpus. This corpus has 18000 documents, and a total of 45 million tokens. The training will start with a pre-trained BERT model (bert-base-uncased), using the `Megatron-LM`[6] library. We run the pre-training on 1 GPU nVidia Tesla P40, with 24GB

[3] m_i is the value of the metric for sample $i \in [1,N]$, then $macro(m) = \frac{1}{N}\sum_{i=1}^{N} m_i$

[4] https://github.com/usnistgov/trec_eval

[5] https://github.com/hscells/plot-trec_eval

[6] https://github.com/cybertronai/Megatron-LM

onboard RAM during 24 hours, or 70000 iteration steps. We export this model for fine-tuning with `fast-bert`[7], during 10 hours for 4 epochs.

The labeled dataset is split 75%/25% between training data and evaluation data. The dataset is split according to the cases, so that the cases in the evaluation dataset are not in the training dataset. This split reflects properly the capacity of the system to generalize to unseen data. In this setting, there will also be candidate cases that are only in the evaluation dataset, and therefore unseen data.

For both models, after the fine-tuning step is finalized, the trained model computes the pairwise relevance score for each pair of query case and candidate case in the test dataset, the score for the positive class is used to rank the candidate cases of each query case.

Our implementation is based on libraries `pytorch-transformers`[8], using `PyTorch`[9] framework.

4.2. Baselines

We consider 2 lexical features baselines based on BM25. As we setup our BERT-based models to learn from the summaries of the full documents, we propose to have a baseline that represents as well this limitation in the availability of information:

- One baseline will use BM25 score as the pairwise relevance score, when the corpus contains all full texts documents
- Another baseline will use BM25 score as the pairwise relevance score, with a corpus made of the summaries of all documents

Further in the Results section, we will also introduce the Perfect Ranker, this ideal system ranks first all noticed cases, and then all unnoticed cases. It will provide us with an upper bound for the performances at each rank.

4.3. Results and Analysis

Table 1. Table of Results for All Systems

System	R@10	P@10	R@1	P@1	P@R	MAP
BM25 Summaries	0.14	0.07	0.02	0.11	0.11	0.14
BM25	0.76	0.36	0.32	0.70	0.68	0.73
PERFECT	*0.97*	*0.50*	*0.41*	*1.00*	*1.00*	*1.00*
Fine-Tuned	0.75	0.34	0.31	0.80	0.63	0.70
Pre-Trained	**0.81**	**0.39**	**0.34**	**0.90**	**0.73**	**0.79**

We first observe that all of our models outperform the baseline of BM25 with summarized texts (model named 'BM25Summ'), and perform at a level comparable to BM25 on full texts. We interpret this part as a validation of our approach, and of the underlying assumption that summarization was a way to capture the information necessary to establish relevance, therefore bringing a positive answer to our research question RQ1.

[7] https://github.com/kaushaltrivedi/fast-bert
[8] https://github.com/huggingface/pytorch-transformers
[9] https://pytorch.org/

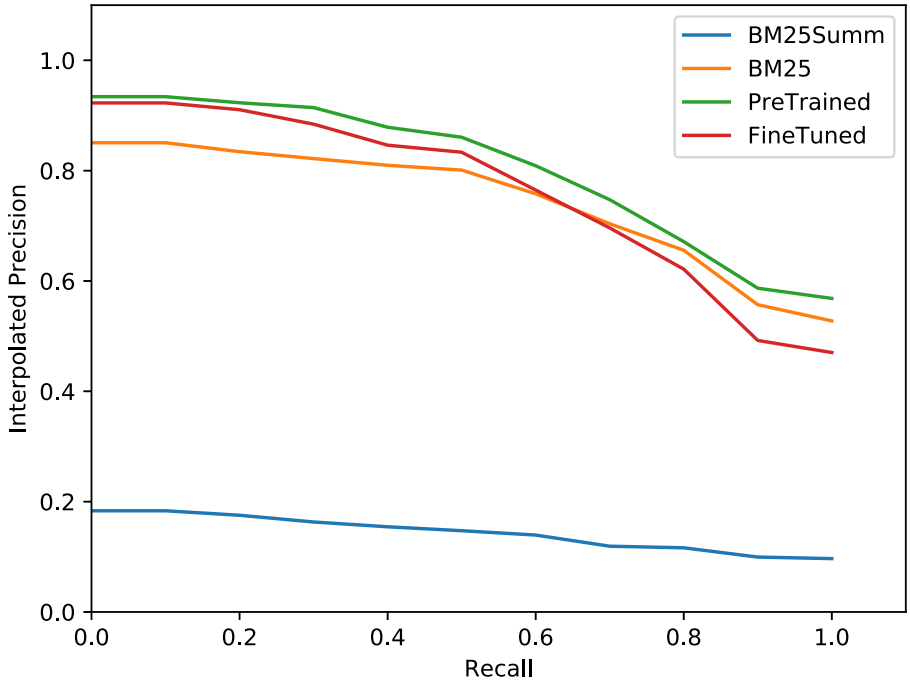


Figure 3. Precision-Recall curves

The Precision-Recall curves in Figure 3, show the skills of the introduced systems. It is noticeable that the `Fine-Tuned` system shows a consistent better performance at all levels of Recall.

With regards to statistical significance, we computed the p-values of an OLS regression model with the target metric as the response variable, and a binary dummy (False for the Baseline, True for the tested model) as the independent variable. We made use of the python library Statsmodels [21]. The p-value for the coefficient of the dummy indicates whether the observed difference in mean is significant or not. We use BM25 as the baseline, and observe a positive effect for both `PreTrained` and `FineTuned` models, while this effect is significant only for the `PreTrained` model.

We attribute this improvement to the knowledge gained during the additional pre-training. The amount of data available for our additional pre-training is orders of magnitude smaller that the material used for the initial pre-training, if we compare millions of words (our data) with tens of billions of words (Wikipedia, BookCorpus, etc.), so we consider this result as a positive answer to our research question RQ2.

5. Statute Law Retrieval

5.1. Experimental Setup

Following the ”Methods” section 2, we pre-process the collection by summarizing all texts. The final dataset is a collection of triplets made of query text, law article text and

relevance judgment. We will fine-tune a pre-trained BERT model for a binary classification task, using the relevance labels from the labeled training dataset. We will also perform additional in-domain pre-training, and then fine-tune this model for the same classification task. In the results section 5.3, the models appear under the names `FineTuned` and `PreTrained`.

For the in-domain pre-training, we used a collection of english translations of court decisions from the Japanese Supreme Court, scraped from their website[10]. This is a rather small dataset of 1500 texts, for a total of 230000 tokens. The training will start with a pre-trained BERT model (bert-base-uncased). We run the pre-training on 1 GPU nVidia Tesla P40, with 24GB onboard RAM during 2 hours, or 70000 iteration steps. We export this model for fine-tuning of 2 hours for 4 epochs.

For both models, after the fine-tuning step is finalized, the trained model computes the pairwise relevance score for each pair of query text and law article text in the test dataset, the score for the positive class is used to rank the law articles for each query.

5.2. Baselines

We use the SOTA from COLIEE 2018 as the baseline, namely `UB3` [22]. This system is based on Terrier[11] and uses TagCrowd[12] for query reduction by extracting keywords.

5.3. Results and Analysis

We report less metrics than for the previous task. As 98% of queries have 1 or 2 relevant articles, we do not report Precision for ranks higher than 1. Results from different systems are compiled in Table 2.

Table 2. Table of Results for All Systems

Name	R@5	R@10	R@30	MAP	P@1
UB3	0.7978	0.8539	0.9551	0.7988	*unknown*
FineTuned	**0.9010**	**0.9203**	**0.9686**	0.8246	0.7971
PreTrained	0.8913	0.9130	0.9444	**0.8321**	**0.8261**

Our work overperforms the baseline significantly. When focusing on only the models we introduce, we observe no significant differences in the evaluation metrics, with the exception of P@1 where the model with additional pre-training significantly improves over the "off the shelf" model.

In this context, the additional pre-training yield a better performance for some metrics, but not others. The observed significant improvement on the P@1 metric is of high interest in the setting of a Question Answering task. We will consider this as a conditional positive answer to our research question RQ2, as the improvement might realize only on some specific metrics.

[10] http://www.courts.go.jp/app/hanrei_en/search
[11] http://terrier.org/
[12] https://tagcrowd.com/

6. Conclusion

We have demonstrated how models based on Generalized Language Models could perform in the context of Legal Information Retrieval, in the presence of limited data. We introduce a document summarization step in order to accomodate the sequence length limitations of BERT. In this setting, our system significantly improves over a BM25 system operating on full text.

We introduced an additional step of pre-training for existing models, which provided significant improvement in the task with the largest amount of training material. We leave for future work the transfer of this method to other specific legal domains.

We consider for future work the possibilities of other new training tasks, considering Masked Language Modeling and Next Sentence Prediction as the way to establish a comprehension at semantic level, while other tasks would contribute to learn the deeper knowledge needed to achieve higher performance on the retrieval tasks.

Use of COLIEE Data

We present these research findings based on the COLIEE Dataset for the Legal Case Retrieval Task, in accordance with the "MEMORANDUM ON PERMISSION TO USE ICAIL 2019 PARTICIPANT DATA COLLECTION". While we use some competition models as baseline, we do not claim that this paper is an entry to the official COLIEE competition. In the "Evaluation Metrics" section 3.3, we elaborate on how the metrics used for competition ranking differ from the metrics we present. We refer to the proceedings of ICAIL 2019 [23] for further reading.

References

[1] P. Arora, M. Hossari, A. Maldonado, C. Conran, G.J. Jones, A. Paulus, J. Klostermann and C. Dirschl, Challenges in the development of effective systems for Professional Legal Search (2018).

[2] I. Nejadgholi, R. Bougueng and S. Witherspoon, A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases., in: *JURIX*, 2017, pp. 125–134.

[3] M. Maclean, J. Eekelaar and B. Bastard, *Delivering family justice in the 21st century*, Bloomsbury Publishing, 2015.

[4] J.-P. Boyd, *Alienated children in family law disputes in British Columbia*, Canadian Research Institute for Law and the Family, University of Calgary, 2015.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, Language Models are Unsupervised Multitask Learners (2019). https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[6] J. Howard and S. Ruder, Fine-tuned Language Models for Text Classification, *CoRR* **abs/1801.06146** (2018). http://arxiv.org/abs/1801.06146.

[7] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR* **abs/1810.04805** (2018). http://arxiv.org/abs/1810.04805.

[8] L. Weng, *Generalized Language Models*, 2019. https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html.

[9] E. Culurciello, *The fall of RNN / LSTM*, 2018. https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0.

[10] P. Tiersma, The nature of legal language (2008), 7–25.

[11] S.E. Robertson, The probability ranking principle in IR, *Journal of documentation* **33**(4) (1977), 294–304.

[12] Y. Lin, Y.C. Tan and R. Frank, Open Sesame: Getting Inside BERT's Linguistic Knowledge (2019). http://arxiv.org/abs/1906.01698.

[13] K. Clark, U. Khandelwal, O. Levy and C.D. Manning, What Does BERT Look At? An Analysis of BERT's Attention (2019). http://arxiv.org/abs/1906.04341.

[14] S. Ruder, M.E. Peters, S. Swayamdipta and T. Wolf, Transfer Learning in Natural Language Processing, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 2019, pp. 15–18.

[15] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes and J. Dean, Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, *CoRR* **abs/1609.08144** (2016). http://arxiv.org/abs/1609.08144.

[16] T. Kiss and J. Strunk, Unsupervised Multilingual Sentence Boundary Detection, *Comput. Linguist.* **32**(4) (2006), 485–525. doi:10.1162/coli.2006.32.4.485.

[17] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[18] F. Barrios, F. López, L. Argerich and R. Wachenchauzer, Variations of the Similarity Function of TextRank for Automated Summarization, *CoRR* **abs/1602.03606** (2016). http://arxiv.org/abs/1602.03606.

[19] R. Řehůřek and P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[20] C.O. Group, *COLIEE Website*, 2019. https://sites.ualberta.ca/~rabelo/COLIEE2019/.

[21] S. Seabold and J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: *9th Python in Science Conference*, 2010.

[22] M. Yoshioka, Y. Kano, N. Kiyota and K. Satoh, 'Overview of Japanese Statute Law Retrieval and Entailment Task at COLIEE-2018,', in: *Twelfth International Workshop on Juris-informatics (JURISIN 2018)*, 2018.

[23] Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019, ACM, 2019, ICAIL. ISBN 978-1-4503-6754-7. doi:10.1145/3322640.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190310

Legislative Dialogues with Incomplete Information

Guido GOVERNATORI [a] and Antonino ROTOLO [b,1]
[a] *Data61, CSIRO, Australia*
[b] *University of Bologna, Italy*

Abstract. This paper extends previous work by presenting a framework for modelling legislative deliberation in the form of dialogues with incomplete information. Roughly, in such legislative dialogues coalitions are initially equipped with different theories which constitute their private knowledge. Under this assumption they can dynamically change and propose new legislation associated with different utility functions.

Keywords. Legislation, Theory revision, Argumentation games, Dialogues

1. Introduction

This paper shows how to formally model legislative deliberation involving coalitions which express public interests. In this sense, it offers a conceptual and technical machinery suitable for designing new decision-support tools for e-Democracy. The contribution follows the general methodology of Governatori *et al.* [1] and extends [2]'s analysis to cover the case of deliberation with incomplete information.

As done with [2], we assume that the legislative procedure can be analysed into two different components: deliberation—the preparatory process of legislation, which runs in the form of a dialogue involving coalitions of agents—and voting (for a critique of this distinction, see [3]). Informally, the idea of legislative dialogue in [2] was the following:

- given an initial theory $\mathscr{T}_0$—intuitively corresponding to the current legislative corpus or a part of it—coalitions propose in a dialogue the legislative theory that amends $\mathscr{T}_0$ and that they would prefer;
- each theory is associated with an utility that measures the impact of the proposed changes given the utility of $\mathscr{T}_0$; the intended reading could be, for example, in terms of the consequence for the society if all agents would conform to such norms (as suggested by rule utilitarianism [4]);
- coalitions deliberate in a different way depending on which of the above theories are employed to compute the utility;
- we may have more rounds in which coalitions amend theories proposed earlier;
- the process does not require that coalitions are fixed during the debate.

[1]Corresponding Author: CIRSFID, University of Bologna, Italy; E-mail: antonino.rotolo@unibo.it.

Several rationality criteria can be introduced to guide the legislative dialogue and the amendments proposed by coalitions [1,2]. For the sake of simplicity, in this paper we only consider one type of utility maximisation among those proposed in [2]. The contribution of this paper goes far beyond [2]'s framework and shows how legislative dialogues work when we abandon the simplistic idea that coalitions in legislative dialogues have complete information, i.e., that the structure of the dialogue (typically, the set of all possible arguments) is common knowledge among the coalitions. This is clearly an oversimplification, as in many real-life contexts players in legislation do not know the entire structure of the argumentation game: in fact, each of them does not know what arguments its opponent will employ and thus takes part in the dialogue in a strategic way. While this thesis was previously defended for legal disputes (see [5]), the point has not yet been extensively analysed for modelling legislative dialogues.

Another contribution is an investigation on how the assumption of incomplete information interplays with the fact that coalitions search to express a majority within the set Ag of agents forming them. This is a very complex research issue, which we address here with some basic and preliminary remarks.

This paper is methodologically aligned with some general approaches developed in law and economics. In particular, we were inspired by the so-called Political Economy School [6], which is based on the following principles:

- private individuals respond to legal rules in an economic fashion;
- private individuals have predominantly self-interested preferences;
- the influence of legal rules is mediated the rational calculus of agents to maximise their preferences;
- public officials are also self-interested[2];
- legislation can be viewed as the product of interest group politics; the problem is then to form coalitions among interests.

While there is a large literature using argumentation for modelling joint deliberation among agents (see [7]), to the best of our knowledge no systematic investigation has been developed combining means-ends rationality principles, theory revision in the law and formal dialogues. The proposal of Shapiro and Talmon [3] is a recent exception, which shares with us the idea that the legislative process proceeds in rounds of deliberation focused on editing a legal text, but the authors do not consider utility criteria guiding the procedure; on the contrary, they analyse voting outcomes—which we do not discuss here—upon a range of conditions, including reaching consensus, a Condorcet-winner, a time limit, or a stalemate. More specifically, we are not aware of any work that combined approaches like the above with the assumption of dialogues with incomplete information.

The layout of the paper is the following. Section 2 recalls basic concepts introduced in [2]. Section 3 shows how legislative dialogues with incomplete information work. Section 4 offers some remarks on how majority dynamics of coalitions interplay with information asymmetries in legislation. Some conclusions end the paper.

[2]We should notice that this assumption does not necessarily mean that public officials work for their direct real benefits. Rather, we have to assume that they are faithful representatives of different social interests coming from groups of private individuals.

2. Background

In this section we recall basic concepts introduced in [2].

2.1. Building Blocks

Let us first give a basic language setting. A literal is a propositional atom or the negation of a propositional atom. Given a literal ϕ, its complementary literal is a literal, denoted as $\sim\phi$, such that if ϕ is an atom p then $\sim\phi$ is its negation $\neg p$, and if ϕ is $\neg q$ then $\sim\phi$ is q. If $Prop$ is a set of propositional atoms then $Lit = Prop \cup \{\neg p \mid p \in Prop\}$ is a set of literals. Rules have the form $\psi_1, \ldots, \psi_n \Rightarrow \phi$ $(0 \leq n)$, $\psi_1, \ldots, \psi_n, \phi \in Lit$. The set of all rules from this language is denoted by Rul.

A corpus of legislative provisions in a given legal system can be defined as a set of legislative rules equipped with priority criteria to rank such rules and solve possible conflicts between them:

Definition 1 (Legislative theory)**.** *A **legislative theory** is a tuple $\mathscr{T} = \langle \mathscr{R}, \succ \rangle$ where $\mathscr{R}$ is a set of rules, and $\succ \subseteq \mathscr{R} \times \mathscr{R}$ is a superiority relation over the rules.*

The legislative deliberation process involves a legislative body of lawmakers (such as the members of a parliament), which we generically call legislative agents, in short *agents*. During the deliberation process, agents can dynamically form coalitions. Typically, at the beginning of the deliberation, coalitions correspond to political-party groups in the legislative body.

Definition 2 (Legislative coalition)**.** *Let Ag be a finite set of agents. A **legislative coalition** in Ag is a subset of agents in Ag. The set 2^{Ag} of all coalitions is denoted by $\mathscr{C}$.*

For brevity we will often speak of coalitions instead of legislative coalitions.

When legislative agents, i.e., the members of the legislative body, argue about theories to govern their own society, they form coalitions proposing theories that represent social interests corresponding to the utility resulting from such theories.

Definition 3 (Coalition social theory utility distribution)**.** *Let $\mathfrak{T}$ be a set of theories, $\mathbb{V}$ an ordered set of values (on which the social utility functions are computed), and $\mathscr{C}$ the set of all legislative coalitions. A **coalition social theory utility distribution** is a function*

$$U : \mathfrak{T} \to \prod_{0}^{|\mathscr{C}|} \mathbb{V}.$$

Given a theory $\mathscr{T}$ and n agents, the function returns a vector of $2^n + 1$ values, which define the value of the theory for each possible coalition in Ag and where the first value, conventionally, indicates the aggregated welfare for all coalitions. Thus, the overall coalitions' utility corresponds in the vector to projection $\pi_0(U(\mathscr{T}))$, while the value of the theory for any specific coalition i corresponds to the projection on the i-th element of the vector, $U_i(\mathscr{T}) = \pi_i(U(\mathscr{T}))$.[3]

[3] In the remainder, $U_i(\mathscr{T})$ denotes the utility of any coalition $i \in \mathscr{C}$. Also, we abuse notation and write $U_{\mathscr{C}}(\mathscr{T})$ to denote the overall coalitions' utility, i.e., $U_j(\mathscr{T})$ where $j = \bigcup_{k \in \mathscr{C}} k$. Accordingly, the overall coalitions' utility corresponds in the vector to projection $\pi_0(U(\mathscr{T}))$.

In line with ideas developed, e.g., by rule utilitarianism, we can determine what is the value of a theory (for each coalition, in our case, and based on the context in which the theory is used) with respect to some inference mechanism [8]. In particular, an approach to articulate the way in which utility springs from any theory $\mathscr{T}$ can be based on the utility of conclusions that follow from arguing on $\mathscr{T}$.

For each literal l in a set *Lit* of literals and given a (possibly different) set of literals $\{l_1,\ldots,l_n\}$, we can define a function λ that assigns for each coalition i in $\mathscr{C}$ an utility value, i.e., the utility that the state of affairs denoted by l brings to i in a context described by $l_1,\ldots,l_n$.

Definition 4 (Coalition literal valuation)**.** *Let $\mathscr{C}$ and $\mathbb{V}$ be, respectively, a set of coalitions and an ordered set of values. A* ***coalition literal valuation*** *is a function*

$$\lambda : \mathscr{C} \times Lit \times \mathrm{pow}(Lit) \to \mathbb{V}.$$

If $E(\mathscr{T}) = \{c_1,\ldots,c_m\}$ is the set of conclusions of a theory $\mathscr{T}$, then a coalition utility can be given by agglomerating the values of all conclusions. Following an intuition from rule utilitarianism, the agglomeration can simply correspond to the sum of individual valuations with respect to any coalition i [4]:

$$U_i(\mathscr{T}) = \sum_{l \in E(\mathscr{T})} \lambda(i,l,E(\mathscr{T})). \tag{1}$$

2.2. Objectives of the Legislative Procedure

In [2] some objectives for the legislative procedure have been also proposed. Among them, one seems of paramount importance: *legislation must produce as an output an optimal theory from the utility point of view*. This requirement can amount to different rational criteria, which include in [2] those producing overall agents' utility optimal theories, i.e., theories maximising the coalitions' utility, or (strong) 'Pareto optimal theories', i.e., theories for which no coalition can be made better off by making some coalitions worse off, or 'maximin optimal theories', i.e., theories maximising the utility of the worst off coalitions, or, finally, theories satisfying Kaldor-Hicks efficiency, i.e., theories in which any coalitions which are made better off could in theory compensate those which are made worse off and so produce a Pareto efficient outcome.

Here, we just recall one of them, i.e., coalitions' utility optimality:

Definition 5 (Coalitions' utility optimal theory)**.** *Let $\mathscr{C}$ be a set of coalitions. A theory $\mathscr{T}^*$ is a* ***coalitions' utility optimal theory*** *amongst a set of theories $\mathfrak{T}$ iff there is no theory $\mathscr{T} \in \mathfrak{T}$ such that $U_{\mathscr{C}}(\mathscr{T}) > U_{\mathscr{C}}(\mathscr{T}^*)$.*

2.3. Legislative Amendments in Dialogues

As argued in [2], a legislative dialogue is the process through which coalitions propose their normative theories with the aim to improve on the current legislative corpus of provisions. The normative system resulting from the dialogue is taken to be justified and so it is suitable for the voting stage.

Several operations can be applied to the the current legislative corpus of provisions in order to revise it. Consider the following very basic operations [9]:

Definition 6 (Theory Revision [9]). *Let $\mathscr{T} = \langle \mathscr{R}, \succ \rangle$ be a legislative theory.*

*The **contraction** of $\mathscr{T}$ with respect to some set R of rules is defined as follows:*

$$(\mathscr{T})^{-R} = \langle \mathscr{R} - R, \succ' \rangle$$

where $R \subseteq \mathscr{R}$ and $\succ' = \succ - \{(r,s) \mid r \in R \text{ or } s \in R\}$.

*The **expansion** of $\mathscr{T}$ with respect to some set R of rules is defined as follows:*

$$(\mathscr{T})^{+R} = \langle \mathscr{R} \cup R, \succ' \rangle$$

where $\succ' = \succ \cup \{(r,s) \mid r \in R, s \in \mathscr{R} \text{ and } C(s) = \neg C(r)\}$.

Definition 6 identifies the legal ways through which legislative theories can be amended: coalitions propose possible amendments in dialogues.

3. Legislative Dialogues with Incomplete Information

Let us now define the notion of legislative dialogue with incomplete information. When legislative dialogues have incomplete information, this means that, while all coalitions share some common knowledge—in addition to the current legislative corpus of provisions (which is assumed to be known by all agents)—they know different ways in which such a corpus can be revised, i.e., they are initially equipped with different additional set of rules which constitute their private knowledge, being unknown by the other parties: each coalition does not know what rules are taken to be valid by the other parties in the game for revising the corpus. *Intuitively, that different coalitions have private knowledge means that they can operate strategically in the dialogue by having ways for promoting their view in the deliberation* and achieving the best results from their viewpoint. In the following definition we assume that all coalitions share knowledge of legislative rules in the legal corpus, plus, possibly, some more legal rules that can be used to revise such corpus. For clarity reasons, we will speak only of this additional set of rules as *the common knowledge of all coalitions* in the dialogue.

Definition 7 (Legislative Dialogue with Incomplete Information). *A **legislative dialogue with incomplete information** (henceforth, dialogue) d is a sequence of triples*

$$\langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle_{k=0,\ldots K}$$

where each

$$\mathscr{T}_k = \langle \mathscr{R}_k, \succ_k \rangle$$

is a legislative theory, and

$$\mathsf{Pr}_k = \{R_k^{i_j} \mid \forall i_j \in \mathfrak{C}, R_k^{i_j} \subseteq Rul, R_k^{i_j} \cap \mathscr{R}_k = \emptyset\}$$
$$\mathsf{Com}_k = \{r \mid r \in Rul, r \notin \mathscr{R}_k \cup (\bigcup_{\forall i_j \in \mathfrak{C}} R_k^{i_j})\}$$

are, respectively, the private knowledge of each coalition with respect to k, and the common knowledge of all coalitions with respect to k that in not contained in $\mathscr{T}_k$.

The dialogue d is such that

- *theory* $\mathscr{T}_0 = \langle \mathscr{R}_0, \succ_0 \rangle$ *is the initial theory;*
- *for each coalition* $i_j \in \mathscr{C}$, $R_0^{i_j}$ *is the initial private knowledge of* i_j*;*
- Com_0 *is the initial common knowledge of all coalitions which is not in* $\mathscr{T}_0$*;*
- *for every triple* $\langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle$, $k > 0$, *there is a set of theories* $\mathfrak{T}^k = \{\mathscr{T}_{i_1}^k, \ldots, \mathscr{T}_{i_n}^k\}$ *where* $\{i_1, \ldots, i_n\} \subseteq \mathscr{C}$ *(i.e., theories individually proposed by coalitions* $i_1, \ldots, i_n$*) such that each* $\mathscr{T}_{i_j}^k$ *is either*
  - $(\mathscr{T}_{k-1})^{-R}$ $(1 \leq j \leq n)$ *for some set* $R \subseteq \mathscr{R}_{k-1}$ *of rules, or*
  - $(\mathscr{T}_{k-1})^{+R}$ $(1 \leq j \leq n)$ *for some set* $R \subseteq R_{k-1}^{i_j} \cup \mathsf{Com}_{k-1}$ *of rules, so that the private knowledge of* i_j *with respect to* k *is* $R_{k-1}^{i_j} \setminus \mathscr{R}_k$*;*
- *triple* $\langle \mathscr{T}_{k+1}, \mathsf{Pr}_{k+1}, \mathsf{Com}_{k+1} \rangle$ *is such that the theory* $\mathscr{T}_{k+1} = Choice(\mathfrak{T}^k)$*, where*
  - *Choice is a function that selects theory* $\mathscr{T}_{k+1}$ *out of a non-empty set* $\mathfrak{T}^k$*;*
  - $\mathsf{Com}_{k+1} = \mathsf{Com}_k \cup (\bigcup_{i_j \in \{i_1, \ldots, i_n\}} \mathscr{R}_{i_j}^k \setminus \mathscr{R}_{k+1})$*;*
- *triple* $\langle \mathscr{T}_K, \mathsf{Pr}_K, \mathsf{Com}_K \rangle$ *is terminal iff* $\mathfrak{T}^K = \emptyset$*.*

Some (but not necessarily all) coalitions start the dialogue by proposing some revisions of the initial legislative theory. At each round of the dialogue the choice function obeys certain rational criteria (such as coalitions' utility maximisation) and aims at ensuring a utility improvement with respect to previous rounds. Legislative revisions proposed by coalitions, if they implement theory expansions, may resort to coalitions' private information, thus adding new rules to coalitions' common knowledge. Notice that all new rules that are proposed by coalitions but are not used for revising the current theory become anyway common knowledge.

Definition 8 (Theories proposed in a dialogue)**.** *The set of theories* $\mathfrak{T}^d$ *proposed in a dialogue* $d = \langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle_{k=0,\ldots K}$ *is* $\bigcup_{k \in \{0,\ldots K\}} \mathfrak{T}^k$*.*

We can note that theory $\mathscr{T}_k$ may be included in $\mathfrak{T}^k$, possibly leading to some sort of equilibrium. However, we are not interested in computing *equilibria* as we deal with principles and not with *moves* as in standard game theoretic approaches. For this reason, we rely on dialogues and not on games, though our dialogues may be seen as *mirroring* such games.

A dialogue is sound if, and only if, the choice function is sound. We concentrate on one sound *Choice* function:

Definition 9 (Coalitions' utility maximising choice)**.** *The choice function of a dialogue* $\langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle_{k=0,\ldots K}$ *is a* ***coalitions' utility maximising choice function*** *iff any theory* $\mathscr{T}_k$ $(2 \leq k)$ *is a coalitions' utility optimal theory amongst the set of theories* $\mathfrak{T}^{k-1}$*.*

Example 1 (Running example)**.** *Let us consider three fixed coalitions: coalition* i_1 *representing people with high incomes because of their high salary, coalition* i_2 *representing those with high incomes because of tax evasion, and coalition* i_3 *representing those with low incomes.*

Suppose the initial theory $\mathscr{T}_0$ *comprises the following:*

$$\begin{aligned}\mathscr{R} = \{&r_1 : \mathit{UpperClass} \Rightarrow \mathit{RaiseTax},\\ &r_2 : \mathit{TaxEvader} \Rightarrow \mathit{SeverePunishment},\\ &r_3 : \mathit{LowerClass} \Rightarrow \mathit{Subsidies},\\ &r_4 : \mathit{LowerClass}, \mathit{TaxEvader} \Rightarrow \neg \mathit{Subsidies},\\ &r_5 : \mathit{TaxEvader} \Rightarrow \mathit{PoorCountry},\\ &r_6 : \quad \Rightarrow \mathit{LowerClass},\\ &r_7 : \quad \Rightarrow \mathit{TaxEvader},\\ &r_8 : \quad \Rightarrow \mathit{InItaly}\}\\ \succ = \{&\langle r_4, r_3\rangle\}\end{aligned}$$

The conclusions of $\mathscr{T}_0$ *are the following:*

$$\begin{aligned}E(\mathscr{T}) = \{&\mathit{SeverePunishment}, \neg \mathit{Subsidies}, \mathit{PoorCountry},\\ &\mathit{LowerClass}, \mathit{TaxEvader}, \mathit{InItaly}\}.\end{aligned}$$

We also have the following: $\mathsf{Pr}_0 = \{R_0^{i_1}, R_0^{i_2}, R_0^{i_3}\}$ *where*

$$\begin{aligned}R_0^{i_1} &= \{r_9 : \mathit{UpperClass} \Rightarrow \neg \mathit{RaiseTax}\}\\ R_0^{i_2} &= \{r_{10} : \mathit{InItaly} \Rightarrow \mathit{Subsidies}\}\\ R_0^{i_3} &= \emptyset\end{aligned}$$

Finally, $\mathsf{Com}_0 = \emptyset$.

Consider, for example, coalition i_2 *and assume that the* λ *function is defined as follows (we omit the literals that are not logically derived):*

$$\begin{aligned}\lambda(i_2, \mathit{SeverePunishment}, E(\mathscr{T})) &= -10\\ \lambda(i_2, \neg \mathit{Subsidies}, E(\mathscr{T})) &= -5\\ \lambda(i_2, \mathit{PoorCountry}, E(\mathscr{T})) &= -2\\ \lambda(i_2, \mathit{LowerClass}, E(\mathscr{T})) &= 0\\ \lambda(i_2, \mathit{TaxEvader}, E(\mathscr{T})) &= 18\\ \lambda(i_2, \mathit{InItaly}, E(\mathscr{T})) &= 0.\end{aligned}$$

Hence, the overall utility of $\mathscr{T}_0$ *for* i_2 *is* 1. *Similarly, we could assume that* λ *works for coalitions* i_1 *and* i_3 *such that the overall utility for the former is* 3 *and* 1 *for the latter. If the global utility is the sum of individual coalitions utility, the utility distribution for* $\mathscr{T}_0$ *is* $[5,3,1,1]$.

What should coalition i_2 *do? Although it represents tax evaders (leading for them to a significant positive utility:* 15*) and their being free-riders, which makes poor the country, only slightly impacts on them personally (*-2*), the overall utility is positive but small. Hence, coalition* i_2 *knows that* $\mathscr{T}_0$ *can be improved. This can be done, for example, by directly working on rules leading to negative utilities, i.e., rules* r_2, r_4, r_5 *and* r_6. *For*

instance, i_2 could propose to amend theory $\mathscr{T}_0$ by expanding the theory and add the following rule from i_2's private knowledge:

$$r_{10} : \mathit{InItaly} \Rightarrow \mathit{Subsidies}$$

If the underlying semantics of reasoning is defeasible reasoning under grounded semantics [2], by expansion $(\mathscr{T}_0)^{+\{r_{10}\}}$ we would block $\neg$Subsidies and the overall utility of the new theory would be 6 for i_2.

Of course, this is i_2's view but the other coalitions play in the debate and work differently. Assume that the new theory $\mathscr{T}_1$ resulting from the debate involving all coalitions goes against the interests of coalition i_2, since the final utility distribution is $U(\mathscr{T}_1) = [8,2,0,6]$ (i.e., taxes are slightly raised for upper classes, tax evasion is more severely punished, and public subsidies are raised for lower classes). If the coalitions' utility maximising choice is adopted then $\mathscr{T}_1$ is elicited.

Assume that i_1 unsuccessfully proposed to add r_9, which was discarded through the deliberation choice. Clearly, the triple

$$\langle \mathscr{T}_1, \mathsf{Pr}_1, \mathsf{Com}_1 \rangle$$

is as follows:

- $\mathscr{R}_1 = \mathscr{R}_0 \cup \{r_{10}\}$*;*
- $R_1^{i_2}$ *in* Pr_1 *is now* $\emptyset$*;*
- $\mathsf{Com}_1 = \{r_9\}$.

Notice that all results proved in [1,2] hold, too, in the case of incomplete information.

Proposition 1. *The terminal theory of a dialogue d with a coalitions' utility maximising choice function is coalitions' utility optimal amongst the set of theories $\mathfrak{T}^d$ proposed in the dialogue if for any $\mathscr{T}_k$, it holds that $\mathscr{T}_k \in \mathfrak{T}^k$.*

Definition 10 (Coalitions' utility improving theory)**.** *Let $\mathscr{C}$ a set of coalitions. A theory $\mathscr{T}^*$ is a **coalitions' utility improvement** of a theory $\mathscr{T}$ iff $U_{\mathscr{C}}(\mathscr{T}^*) > U_{\mathscr{C}}(\mathscr{T})$.*

Proposition 2. *A theory is a coalitions' utility optimal theory amongst a set of theories $\mathfrak{T}$ iff there exist no coalitions' utility improvements in $\mathfrak{T}$ of the theory.*

Proposition 3. *The terminal theory of a dialogue d with a coalitions' utility maximising choice function is coalitions' utility optimal amongst the set of theories $\mathfrak{T}^d$ proposed in the dialogue and it is a coalitions' utility improvement of the initial theory, if for any $\mathscr{T}_k$, it holds that $\mathscr{T}_k \in \mathfrak{T}^k$, and there exists a theory $\mathscr{T}_k$ which is a coalitions' utility improvement of $\mathscr{T}_{k-1}$.*

4. Majority Dynamics and Incomplete Information

So far coalitions adopt only some type of means-ends rationality. However, deliberative procedures usually assume that some other basic constraints apply to them. In particular, *coalitions naturally search to express a majority* within the set Ag of agents.

As suggested in [2], we should notice that Definition 7 does not require that coalitions are fixed in the dialogue, but simply that at each turn in the dialogue some coalitions individually propose some revised theories. Hence, if the legislative body works on the basis of the *majority principle* as applied to the agents forming the coalitions, it is obvious that such coalitions could change during the dialogue.

This means that an additional criterion for dialogues can be added.

Definition 11 (Coalitions' majority optimal choice)**.** *The choice function of a dialogue* $\langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle_{k=0,\ldots K}$ *is a* ***coalitions' majority optimal choice function*** *iff any theory* $\mathscr{T}_k = \mathscr{T}_k^{i_j}$ $(2 \leq k)$ *amongst the set of theories* $\mathfrak{T}^{k-1}$ *is such that* $|i_j| > |Ag|/2$.

In other words, a coalitions' majority optimal choice ensures that each theory selected at each turn is proposed by a majoritarian coalition in Ag (since the size of the coalition i_j must exceed the half of the size of the set of agents). Definition 11 works with simple majority, but other requirements such as supermajority or unanimity can be easily implemented. It is easy to prove the following result also with incomplete information:

Proposition 4. *The terminal theory of a dialogue d with a coalitions' majority optimal choice is majority optimal amongst the set of theories* $\mathfrak{T}^d$ *proposed in the dialogue if for any* $\mathscr{T}_k$*, it holds that* $\mathscr{T}_k \in \mathfrak{T}^k$.

Of course, as done with utility maximisation (see Definition 10), we can imagine that dialogues aim at maximising majorities by reconfiguring coalitions during the debate.

Definition 12 (Majority improving theory)**.** *Let* $\mathscr{C}$ *a set of coalitions and* $i_j, i_k \in \mathscr{C}$*. A theory* $\mathscr{T}_*^{i_j}$ *is a* ***coalitions' majority improvement*** *of a theory* $\mathscr{T}^{i_k}$ *iff* $|i_j| > |i_k|$.

If Definition 5 applies to dialogues, coalitions improvement may impact of their private knowledge.

Let Ag be a set of agents:

$$Ag = \{ag_1, ag_2, ag_3\}.$$

All possible coalitions are trivially the following:

$$\begin{aligned} i_1 &= \{ag_1\} \\ i_2 &= \{ag_2\} \\ i_3 &= \{ag_3\} \\ i_4 &= \{ag_1, ag_2\} \\ i_5 &= \{ag_2, ag_3\} \\ i_6 &= \{ag_1, ag_3\} \\ i_7 &= \{ag_1, ag_2, ag_3\} \end{aligned}$$

Assume that in the dialogue only coalitions i_3 and i_4 propose at round 1 revisions of $\mathscr{T}_0$ and suppose that i_4 successfully revises theory $\mathscr{T}_0$ by adding rules from its private knowledge $R_0^{i_4}$, thus resulting into $\mathscr{T}_1$. Nothing prevents that some private information from $R_1^{i_3} \cup \mathsf{Com}_1$, when combined with $R_1^{i_4}$, may support a utility improvement, this time being based on the largest coalition i_7. If this happens, private information no longer exists, as all rules are common knowledge among agents and possible coalitions.

Proposition 5. *Let $\mathscr{C}$ a set of coalitions, $i_j, i_k \in \mathscr{C}$ and R any set of rules. If theory $\mathscr{T}_*^{i_j}$ is a coalitions' majority improvement of a theory $\mathscr{T}_\sharp^{i_k}$ in dialogue $\langle \mathscr{T}_k, \mathsf{Pr}_k, \mathsf{Com}_k \rangle_{k=0,\ldots K}$, where $\mathscr{T}_*^{i_j} = (\mathscr{T}_\sharp^{i_k})^{+R}$, then $\mathsf{Com}_*^{i_j} \subseteq \mathsf{Com}_\sharp^{i_k}$.*

5. Summary

In this paper we extended Governatori *et al.* [1,2]'s framework to the legal domain for modelling legislative deliberation with incomplete information. As done in [2], we assumed that the legislative procedure can be analysed into two different components: deliberation—the preparatory process of legislation, which runs in the form of a dialogue involving coalitions of agents—and voting—which was not discussed here.

The idea of legislative deliberation consists in revising the current legislative corpus or a part of it, where agents's coalitions propose in a dialogue legislative theories that amends such corpus. Each revision is associated with an utility that measures the impact of the proposed changes. Several rationality criteria can be described according to which coalitions deliberate.

In this sense, we argued that this work is methodologically aligned with some general approaches developed in law and economics. In particular, we were inspired by the so-called Political Economy School, where legislation can be viewed as the product of interest group politics and the problem is then to form coalitions among interests.

The current contribution extended this analysis by making [2]'s original framework more realistic: indeed, coalitions in the dialogue can be strategic and exploit in a convenient ways their private knowledge. Once the search of majorities is added to the framework, this integration exhibits some expected but interesting interactions with the dynamics of information asymmetries.

References

[1] G. Governatori, F. Olivieri, R. Riveret, A. Rotolo and S. Villata, Dialogues on Moral Theories, in: *Deontic Logic and Normative Systems - 14th International Conference, DEON 2018, Utrecht, The Netherlands, July 3-6, 2018.*, 2018, pp. 139–155.

[2] G. Governatori, A. Rotolo, R. Riveret and S. Villata, Modelling Dialogues for Optimal Legislation, in: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019.*, 2019, pp. 229–233. doi:10.1145/3322640.3326731.

[3] E. Shapiro and N. Talmon, Integrating Deliberation and Voting in Participatory Drafting of Legislation, *CoRR* **abs/1806.06277** (2018). `http://arxiv.org/abs/1806.06277`.

[4] J.C. Harsanyi, Rule utilitarianism and decision theory, *Erkenntnis* **11**(1) (1977), 25–53.

[5] G. Governatori, F. Olivieri, A. Rotolo, S. Scannapieco and G. Sartor, Two Faces of Strategic Argumentation in the Law, in: *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, 2014, pp. 81–90. doi:10.3233/978-1-61499-468-8-81.

[6] L.A. Kornhauser, Economic Rationality in the Analysis of Legal Rules and Institutions, in: *The Blackwell Guide to the Philosophy of Law and Legal Theory*, Wiley, 2008, pp. 67–79.

[7] K.A. et al. (eds.), Special issue on argumentation in multi-agent systems, *Argument & Computation* **7** (2016).

[8] J.C. Harsanyi, Morality and the Theory of Rational Behavior, *Social Research* **44** (1977), 623–656.

[9] G. Governatori and A. Rotolo, Changing legal systems: legal abrogations and annulments in Defeasible Logic, *Logic Journal of the IGPL* **18**(1) (2010), 157–194. doi:10.1093/jigpal/jzp075.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190311

Verifying Meaning Equivalence in Bilingual International Treaties

Linyuan TANG [a] andKyo KAGEURA [b]
[a] *Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Tokyo, Japan*
[b] *Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo, Japan*

Abstract. This paper examines to what extent distributional approaches to induce bilingual lexica can capture correspondences between bilingual terms in international treaties. Recent developments in bilingual distributional representation learning methods have improved bilingual textual processing performances, and the application of these methods to processing specialised texts and technical terms has increased, including in the legal domain. Here we face at least two issues. Firstly, whether technical terms follow the distributional hypothesis or not is both theoretically and practically a critical concern. Theoretically, corresponding technical terms in different languages are the labels of the same concept and thus their equivalence is independent of the textual context. From this point of view, the distributional hypothesis holds only when the terms totally bind the context. This leads to the second issue, i.e. to verify the extent to which word embedding models trained on texts with different levels of specialisation are useful in capturing cross-lingual equivalences of terms. This paper examines these issues by conducting experiments in which different models trained on the texts with different degree of specialisations are evaluated against three different sets of equivalent bilingual pairs in the legal domain, i.e. of legal terms, of sub-technical terms and of general words. The results show that models learned on large-scale general texts fall far behind models learned on specialised texts in representing equivalent bilingual terms, while the former models have better performances for sub-technical terms and general words than the latter.

Keywords. legal terminology, international treaty and convention, cross-lingual word embedding

1. Introduction

This paper explores to what extent distributional approaches to induce bilingual lexica can capture correspondences between bilingual legal terms in international treaties. We focus on English and Japanese pairs.

Multilingual communication in law is based on the presumption that all the authentic texts of a legal instrument are equal in meaning, effect, and intent [1]. The equal meaning presumption is secured from three aspects. First, the presumption is codified in Article 33(3) of the Vienna Convention on the Law of Treaties. Second, there is always an 'authentic texts' article of international treaties and conventions to prescribe that the original

texts, which are often written in Arabic, Chinese, English, French, Russian and Spanish, are equally authentic. Third, for some essential terms in the certain contexts, there is always a 'use of terms' article to define the meaning of terms for the purposes of the treaty. Technical terms play an essential role in achieving the equality in meaning; this is so because they represent concepts independent of the textual context. The equality in meaning basically holds for official translations of the treaties in different languages. For instance, Japanese official translations of treaties Japan ratified are supposed to portray exact meanings in the original authentic texts.

In natural language processing, measuring the extent to which two words are semantically similar is one of the popular research topics with a wide range of successful applications [2]. This is promoted by the emergence and rapid development of corpus-oriented distributional semantic models [3,4,5,6,7]. The application of these methods to processing specialised texts and technical terms has also increased, including in the legal domain [8,9]. However, whether or not distributional models can sufficiently capture the equivalence of meaning is yet to be fully examined.

As corresponding technical terms in different languages represent the same concept *independent of* the textual context [10,11], we have two issues of appropriateness in applying distributional semantic models to the legal domain:

1. to examine whether legal technical terms follow the distributional hypothesis or not in the first place;
2. to verify the extent to which word embedding models trained on texts with different levels of specialisations are useful in capturing cross-lingual equivalences of legal technical terms.

We defined three different evaluation sets of parallel bilingual lexical pairs occurring in international treaties in English and Japanese, i.e. of technical legal terms, of sub-technical terms, and of general words [12], and carried out experiments in which different bilingual word embedding models trained on the texts with different degree of specialisations, i.e. general corpus (Wikipedia here) and legal corpus, are evaluated against each of the three sets.

The results of our experiments also give insights into more socially oriented questions postulated from a different point of view. According to [1], the equal meaning presumption of treaties in different languages rarely holds. This may be partly due to the fact that interpretations of technical terms, even though theoretically they are supposed to be independent of the contexts, still are affected by their informal non-technical usages in wider discourse. Assuming that distributional representations of terms or words learned on texts reasonably reflect their prevalent meaning embodied in the discourse represented by the texts, the gap between distributional models and the evaluation sets can be interpreted as reflecting this factor.

2. Related work

2.1. Characteristics of bilingual words/terms in the legal domain

Lexical elements in specialised domains can be very different from those in general langauge. In terminology literature, a technical term is defined as a lexical item that repre-

sents a concept inside a specialised domain [1]. A word, in contrast, is essentially a syntagmatic unit that is located in the hierarchy of linguistic units. Thus handling terms requires onomasiological approach in which concepts are identified first and terms are regarded as their labels [11]. Translators need to keep this characteristic of terms [10,11].

Besides technical terms, there are also lexical items called 'sub-technical terms' that are "frequently shared by the general and specialised fields which either retain a legal meaning in general English or acquire a specialised one in the legal context," and English legal lexicon is characterised by the high ratio of these elements [12].

A theoretical overview of legal translation is given in [1]. The priority in translations of international treaties is "to achieve the greatest possible interlingual concordance so as to prevent any ambiguity that could result in international disputes, unnecessary litigation or legal uncertainty." Still,"an apparently harmless linguistic diversity" can later lead to major differences in interpretation.

2.2. Cross-lingual word embedding methods

The pioneering work that explored cross-lingual word similarity relation using word embedding models revealed that parallel corpora may have similar distributed structure and thus parallel word pairs may closely located in a common semantic space after a linear transformation for combining two monolingual word embeddings [13].

Models for learning cross-lingual representations developed so far can be broadly classified into joint methods and mapping methods [14]. Joint methods [15,16,17] simultaneously learn word representations for multiple language on parallel corpora, while mapping methods [18,13] independently train word embeddings in different languages and map them to a common space, supervised or unsupervised.

Recent unsupervised approaches show competitive results with their supervised counterparts [19,20]. Comparative analyses of mapping methods are conducted in [21]. They demonstrated that the performance of cross-lingual embedding models largely depends on the task at hand.

The limitations of mapping methods are investigated in [14]. They experimented with parallel corpora to compare offline mapping to joint learning. and observed that joint learning yields to more isomorphic embeddings and obtains better results in bilingual lexicon induction under ideal conditions. However, these embedding-based joint algorithms are also claimed to be "unable to outperform the traditional methods by a significant margin" in [22].

3. Experimental Design

The experiment in this paper can be viewed as a bilingual lexicon induction (BLI) task, a common practice of bilingual word embedding approaches, in which a bilingual dictionary is induced by linking each source word in the source language with its nearest neighbor(s) in the target language. The basic framework of the BLI task is that multiple models are evaluated against a given gold standard which is a list of pre-defined bilingual word pairs. Three essential components are involved: the corpus on which the distribu-

[1] We nevertheless use the word 'meaning' as an umbrella word to cover both 'concept' and 'meaning' for succinctness.

tion of words are based, the method with which the bilingual distributional models are trained using monolingual models, and the gold set of bilingual word pairs against which the models are evaluated.

In contrast to those current approaches, the primary goal of our experiment is to examine whether the distributional hypothesis holds in the first place because to apply those semasiological approaches to technical legal terms can be theoretically inappropriate. In spite of this theoretical inconsistency, the corpus-oriented distributional semantic models can still be useful in practice because legal texts contain words and terms with different degree of specialisations.

3.1. Gold sets of bilingual pairs

In order to carry out our investigation, we first constructed valid gold sets of equivalent bilingual pairs with different degree of specialisations. The procedure of constructing the gold sets is as follows:

1. We first collected 45 international treaties that Japan ratified, of which authentic Japanese versions exist.
2. We introduced two pre-defined bilingual dictionaries as base dictionaries, i.e. *Standard Legal Term Dictionary* [2] obtained from *Japanese Law Translation Database* [3] as the technical legal term dictionary, and *ground-truth bilingual dictionary* (Japanese-English, English-Japanese) provided in the Python library `MUSE` [4] as the general word dictionary, to identify bilingual elements in the treaties. We converted all the multi-to-multi pairs in both dictionaries into one-to-one pairs. Phrase pairs are excluded in order to fit the word embedding methods. Using the bilingual pairs extracted from the dictionaries, we constructed three bilingual word/term pairs from the treaties:
   - *General word* pairs which consist of pairs listed in the general word dictionary and not appeared in the technical legal term dictionary;
   - *Sub-technical term* pairs which consist of pairs listed in both technical legal term and general word dictionaries;
   - *Legal term* pairs which consists of pairs listed in the technical legal term dictionary but not in the general word dictionary.

Table 1. Basic statistics of two base dictionaries and three gold sets of bilingual pairs. The numbers of unique Japanese/English words are shown in 'Types (JA)/(EN)'. The diffrence in counts is due to polysemes.

	base dictionary		gold set		
	general word	technical term	general word	legal term	sub-technical term
#Pairs	25,918	1,861	1,642	386	306
#Types (JA)	20,968	1,287	1,257	306	275
#Types (EN)	22,484	1,272	1,536	320	282

Table 1 shows the basic quantities of gold sets and base dictionaries.

[2] March 2013 Edition. `http://www.phontron.com/jaen-law/index-ja.html`.
[3] `http://www.japaneselawtranslation.go.jp/?re=02`.
[4] `https://github.com/facebookresearch/MUSE`.

3.2. Corpora and training methods

Word distributions in specialised domains vary from domains and are different from distribution based on general corpus including Wikipedia. We thus used monolingual embedding models trained on the texts with different degree of specialisations. For general word embedding models, we used monolingual embedding models pre-trained on English and Japanese Wikipedia. For domain-specific models, we used the corpus consisting of bilingual legal texts.

As introduced in section 2.2, existing approaches to obtain bilingual embedding models can be roughly distinguished into joint methods, supervised mapping methods, and unsupervised mapping methods. According to [22], joint embedding models are unable to outperform the traditional non-embedding models by a significant margin on the BLI task. We thus limit the number of approaches to be compared from four (non-embedding, joint embedding, supervised mapping embedding, and unsupervised mapping embedding approaches) to three (non-embedding, supervised mapping embedding, and unsupervised mapping embedding approaches) for succinctness.

Monolingual embeddings We applied the pre-trained English word embedding model provided in the `MUSE` library as English monolingual general word embedding model. The model was trained on English Wikipedia with the fastText [3] method. We did not apply the same pre-trained Japanese model because the provided model can hardly detect parallel word pairs after bilingual word embedding training. Instead, we applied the pre-trained *Wikipedia Entity Vectors* [5] as our Japanese monolingual general word embedding model.

To train domain-specific models, we used two legal text corpora. One consists of 45 English-Japanese parallel international treaties and conventions (1945–2003) obtained from *The World and Japan Database* [6]. The other additionally contains Japanese laws and their English translations obtained from *Japanese Law Translation Database* besides the international treaties and conventions. Details of those two corpora are shown in table 2. The trained domain-specific embeddings are skip-gram models using Word2Vec (implemented in `gensim` [23]) with $dimension = 300$, $window.size = 10$, $negative.sample = 10$, $min.count = 3$ and other default settings [7].

Table 2. Basic statistics of the two domain-specific corpora.

Dataset	#Sentence pairs	#Tokens (EN)	#Tokens (JA)	#Types (EN)	#Types (JA)
Treaty only	15,186	277,321	329,533	7,131	5,987
Treaty and Law	277,635	9,783,174	10,330,284	26,570	23,911

[5] A 300-dimension Skip-Gram Negative Sample (SGNS) Word2Vec model trained on Japanese Wikipedia. `https://github.com/singletongue/WikiEntVec` (latest version).

[6] `http://worldjpn.grips.ac.jp/`.

[7] We examined different settings (5, 10, 15) of *window.size* and *negative.sample*. Due to the relatively tiny data size, those parameters lower than 10 led to worse performances of Japanese models.

Non-embedding bilingual approach The non-embedding method applied in this paper is IBM Model-1 (implemented in `NLTK` [24]; for more details, see [25]), a statistic alignment model developed for lexical translation, which is one of the traditional models that were used in [22], which confirmed its high performance. The probability that a source word is translated into a target word is based on the co-occurrence of word pairs in parallel sentence pairs. Due to the lack of parallel sentences of Wikipedia, we only use sentence pairs in the two domain-specific copora as input data.

Bilingual word embedding approaches We utilised the Procrustes approach (for the "orthogonal Procrustes problem", see [26]) as the supervised mapping method, and the adversarial Procrustes approach as the unsupervised mapping method [8] (implemented in the `MUSE` library).

The Procrustes approach (for more details, see [19,27]) transforms a given matrix (in this case, the matrix containing the embeddings of the source words in the parallel pairs) into another matrix (the matrix containing the embeddings of the target words in the pairs) by an orthogonal transformation matrix, so that the sum of squares of the residual matrix is a minimum. A subset of the ground-truth dictionary is used as anchor points for Procrustes (i.e. to gain a transformation matrix for mapping the source monolingual embedding to the target embedding) [9].

The unsupervised method uses an adversarial criterion [28] to learn the mapping without cross-lingual supervision including word-level parallel data (for more details, see [18,19,29]). After generating an initial proxy of the transformation matrix with adversarial training, a synthetic parallel vocabulary is built using this matrix to refine the mapping. The subsequent procedure is the same as the supervised method.

3.3. Evaluation methods

We evaluate the performance of each trained model against each gold set on BLI task. The measurement is the precision at k (@k, k is set to $1, 5, 10$ in this paper), that counts how many pairs of which the target word are ranked in the top-k nearest neighbors of its counterpart are extracted. The probability that a source word is translated into a target word is calculated by EM algorithm in IBM Model-1 [25]. While the retrieval metric used in the embedding approaches is cross-domain similarity local scaling (CSLS. For more details, see [19]). The CSLS increases the accuracy for word translation retrieval and consistently outperform cosine similarity on nearest neighbor retrieval, while not requiring any parameter tuning.

4. Results

Bilingual word embedding models trained on pre-trained monolingual general models are referred to as *Wiki* models. *JaL* models refer to the models that are trained on treaty

[8] We also tried to find cross-lingual nearest neighbors using unmapped pre-trained monolingual embeddings. The results prove that joint or mapping processes are necessary for the cross-lingual similarity task.

[9] Identical characters in both monolingual embeddings can be used as the training dictionary as well. Additionally, we modified the training dictionary by combining legal term dictionary with the default one. However, these two modifications achieve no notable performance improvement in all circumstances.

and law corpus in both non-embedding and embedding approaches, while *InT* models refer to the models that are trained solely on the treaty corpus in both approaches.

The adversarial Procrustes method (unsupervised embedding approach) failed to outperform the Procrustes method (supervised embedding approach). The EN→JA models (Japanese monolingual data as source data, while English data as target data) failed to outperform the JA→EN models in all circumstances except in the case of InT embedding models. Polysemes in the pairs may affect the results.

Due to the above reasons, we only report the BLI results (in table 3) of the EN→JA models in non-embedding approach (IBM Model-1) and the supervised mapping embedding approach (Procrustes) against three gold sets. We also derived the BLI scores against the union set of all the word/term pairs in the three gold sets (the gold set of 'all' pairs) from individual results of each case.

In all circumstances, while the precision increases largely from @1 to @5, the precision improvement from @5 to @10 is less remarkable. Therefore, we only refer to @1 and @10 in the following discussion.

Table 3. BLI results of both non-embedding (IBM Model-1) and embedding approach (Procrustes). Models are trained from Japanese to English. The gold set of 'all' is a union set of three gold pair set of general words, of sub-technical terms, and of legal terms. Due to the lack of sentence-level parallel Wikipedia data, we were unable to conduct the experiment in which Japanese/English Wikipedia corpus are trained in the non-embedding approach. The best scores against each gold set are in bold. The best scores of embedding models against each gold set are underlined.

Gold set	Model	Non-embedding				Embedding			
		@1	**@5**	**@10**	**#Pairs**	**@1**	**@5**	**@10**	**#Pairs**
All	Wiki					**42.8**	**63.4**	**68.1**	2,331
	JaL	28.3	51.0	57.7	1,926	22.3	30.8	34.1	2,175
	InT	27.9	48.3	54.3	1,599	5.9	8.1	8.7	1,701
(General dict.	Wiki					25.2	40.7	46.2	25,918)
General word	Wiki					**51.6**	**73.4**	**78.2**	1,642
	JaL	28.3	51.6	58.1	1,336	18.8	26.4	29.4	1,486
	InT	29.5	50.0	56.6	1,147	7.1	9.7	10.3	1,108
Sub-technical term	Wiki					42.5	64.7	70.6	306
	JaL	40.5	**69.6**	**78.8**	299	**47.7**	60.1	64.7	306
	InT	33.7	61.1	66.0	250	6.5	9.5	10.1	263
Legal term	Wiki					5.6	19.7	23.1	383
	JaL	**18.9**	**33.9**	**39.1**	291	16.8	26.4	29.8	383
	InT	16.3	31.1	35.0	202	0.3	0.3	0.5	330

4.1. With regard to the distributional hypothesis

We present the BLI result of the supervised Wiki embedding EN→JA model against the *general word dictionary* here as a baseline for evaluation. In the case of words in everyday usage, the supervised large-scale bilingual embedding model is able to identify parallel pairs with 25.2%@1 and 46.2%@10 precision.

The conclusion at the first glance could be that the language usage of international treaties distinctly follow the distributional hypothesis, since the Wiki embedding model has an almost 20% improvement on @k precision against the gold set of *all* pairs comparing with the baseline. However, the good performance is mainly contributed by high precision against the *general word* set. The performance against the *sub-technical term* set is similar to overall performance, and drops drastically to 5.6%@1 and 23.1%@10 precision against the *legal term* set.

Domain-specific models in the non-embedding approach outperform those in the embedding approaches with roughly 5∼20% improvements in almost all circumstances. The poor performances of the InT embedding models can be due to the tiny corpus size. In the case of JaL embedding models, despite the major precision deterioration against the *general word* set, @1 precision against the *sub-technical term* set increases from 40.5% to 47.7%, and is competitive with that in the non-embedding approach against the *legal term* set (18.9%/16.8%).

These observations demonstrate that embedding approaches can achieve overall high performance on the BLI task in the legal domain only because there exists a large amount of general words which follow the distributional hypothesis. The low performance of embedding models against *technical legal term* set will likely be obscured without the separation.

4.2. The capability of embedding models to capture meaning equivalence

Focusing on the scores achieved by embedding models (the right part of table 3), we have mixed results about the capability of those models to capture meaning equivalence of parallel pairs.

Against the *general word* set, the Wiki model achieves 51.6%@1 and 78.2%@10 precision which are 20% higher than those of the non-embedding domain-specific models. Against the *sub-technical term* set, the performance of the Wiki model is similar to the JaL model. Both models achieve 40%@1 and over 60%@10 precision, while the JaL model detect 5% more top-1 most similar pairs and 5% less top-10 most similar pairs than the Wiki model. Against the *technical legal term* set, the highest performance is achieved by the JaL model with 16.8%@1 and 29.8%@10 precision.

The general word embedding model can perform well on detecting the correct counterpart if the level of specialisation of the evaluated pairs is comparably low, while its performance fall significantly against the *technical legal term* set. Conversely, the domain-specific embedding model fails to capture similarity relation of *general word* pairs, while greatly outperforming the model learned on large-scale general texts.

5. Conclusions

We conducted experiments in which different models trained on the texts with different degree of specialisations are evaluated against three different gold sets of equivalent bilingual word/term pairs in the international treaties in order to verify the meaning equivalence of technical terms in the legal domain with the concern that those terms do not follow the distributional hypothesis.

The answer to our first research question that whether legal technical terms follow the distributional hypothesis is highly probably negative. The overall good performance

of the embedding models learned on large-scale general texts is mostly contributed by the detection of general word and sub-technical term pairs in the bilingual treaties. Even the embedding models learned on legal texts can hardly identify parallel technical legal terms.

In terms of the second issue that to which extent word embedding models are useful in capturing cross-lingual equivalences of legal technical terms, embedding models can achieve competitive performance comparing with non-embedding models in detecting the most similar bilingual counterparts. General word embedding models fall far behind domain-specific models in representing equivalent bilingual technical legal terms, while those have better performances on capturing the similarity relation of general words and sub-technical terms. Despite the theoretical inconsistency, domain-specific embedding models have potential to outperform general embedding models on this task as the degree of specialisations increases.

In conclusion, the gap between distributional models and the evaluation sets can be interpreted as reflecting the gap between semasiological assumption in distributional models and onomasiological characteristics of technical legal terms. Thus to distinguish technical terms from general words in advance should be a prerequisite for processing legal texts in distributional approaches.

Acknowledgments. We thank three anonymous reviewers for helpful comments. This work was supported by Grant-in-Aid for Scientific Research(S) Grant Number 19H05660.

References

[1] S. Šarcevic, Legal translation and translation theory: A receiver-oriented approach, in: *International Colloquium,'Legal translation, theory/ies, and practice', University of Geneva*, 2000, pp. 17–19. `https://www.tradulex.com/Actes2000/sarcevic.pdf`.

[2] J. Camacho-Collados, M.T. Pilehvar, N. Collier and R. Navigli, SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity, in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, 2017. doi:10.18653/v1/s17-2002.

[3] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching Word Vectors with Subword Information, *arXiv preprint arXiv:1607.04606* (2016).

[4] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018).

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient Estimation of Word Representations in Vector Space, *arXiv preprint arXiv:1301.3781* (2013).

[6] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[7] J. Pennington, R. Socher and C.D. Manning, GloVe: Global Vectors for Word Representation, in: *Proc. of EMNLP 2014*, 2014, pp. 1532–1543. `http://www.aclweb.org/anthology/D14-1162`.

[8] D.S. Carvalho, V. Tran, K.V. Tran and N.L. Minh, Improving Legal Information Retrieval by Distributional Composition with Term Order Probabilities, in: *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment*, K. Satoh, M.-Y. Kim, Y. Kano, R. Goebel and T. Oliveira, eds, EPiC Series in Computing, Vol. 47, EasyChair, 2017, pp. 43–56. ISSN 2398-7340. doi:10.29007/2xzw. `https://easychair.org/publications/paper/7Pv`.

[9] R. Nanda, A.K. John, L.D. Caro, G. Boella and L. Robaldo, Legal Information Retrieval Using Topic Clustering and Neural Networks, in: *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment*, K. Satoh, M.-Y. Kim, Y. Kano, R. Goebel and T. Oliveira, eds, EPiC Series

in Computing, Vol. 47, EasyChair, 2017, pp. 68–78. ISSN 2398-7340. doi:10.29007/psgx. https://easychair.org/publications/paper/RC.

[10] L. Bowker, Terminology and translation, in: *Handbook of Terminology*, Vol. 1, H.J. Kockaert and F. Steurs, eds, John Benjamins, 2015, pp. 304–323.

[11] K. Kageura, Terminology and lexicography, in: *Handbook of Terminology*, Vol. 1, H.J. Kockaert and F. Steurs, eds, John Benjamins, 2015, pp. 45–59. doi:10.1075/hot.1.04ter2.

[12] M.J. Marín Pérez, Measuring the Degree of Specialisation of Sub-technical Legal Terms through Corpus Comparison: A Domain-independent Method, *Terminology* **22**(1) (2016), 80–102. doi:10.1075/term.22.1.04mar.

[13] T. Mikolov, Q.V. Le and I. Sutskever, Exploiting Similarities among Languages for Machine Translation, *arXiv preprint arXiv:1309.4168* (2013).

[14] A. Ormazabal, M. Artetxe, G. Labaka, A. Soroa and E. Agirre, Analyzing the Limitations of Cross-lingual Word Embedding Mappings, *arXiv preprint arXiv:1906.05407* (2019).

[15] K.M. Hermann and P. Blunsom, Multilingual Models for Compositional Distributed Semantics, *arXiv preprint arXiv:1404.4641* (2014).

[16] T. Luong, H. Pham and C.D. Manning, Bilingual Word Representations with Monolingual Quality in Mind, in: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Association for Computational Linguistics, 2015. doi:10.3115/v1/w15-1521.

[17] I. Vulic and M.-F. Moens, Bilingual Distributed Word Representations from Document-aligned Comparable Data, *J. Artif. Int. Res.* **55**(1) (2016), 953–994. https://arxiv.org/pdf/1509.07308.pdf.

[18] M. Artetxe, G. Labaka and E. Agirre, Learning bilingual word embeddings with (almost) no bilingual data, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2017. doi:10.18653/v1/p17-1042.

[19] A. Conneau, G. Lample, M. Ranzato, L. Denoyer and H. Jégou, Word Translation Without Parallel Data, *arXiv preprint arXiv:1710.04087* (2017).

[20] C. Zhou, X. Ma, D. Wang and G. Neubig, Density Matching for Bilingual Word Embedding, *arXiv preprint arXiv:1904.02343* (2019).

[21] G. Glavas, R. Litschko, S. Ruder and I. Vulic, How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions, *arXiv preprint arXiv:1902.00508* (2019).

[22] O. Levy, A. Søgaard and Y. Goldberg, A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments, in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, 2017. doi:10.18653/v1/e17-1072.

[23] R. Řehůřek and P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, 2010, pp. 45–50.

[24] S. Bird, E. Loper and E. Klein, *Natural Language Processing with Python*, O'Reilly Media Inc., 2009.

[25] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra and R.L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, *Comput. Linguist.* **19**(2) (1993), 263–311. http://dl.acm.org/citation.cfm?id=972470.972474.

[26] P.H. Schönemann, A generalized solution of the orthogonal procrustes problem, *Psychometrika* **31**(1) (1966), 1–10. doi:10.1007/bf02289451.

[27] S.L. Smith, D.H.P. Turban, S. Hamblin and N.Y. Hammerla, Offline bilingual word vectors, orthogonal transformations and the inverted softmax, *arXiv preprint arXiv:1702.03859* (2017).

[28] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative Adversarial Networks, *arXiv preprint arXiv:1406.2661* (2014).

[29] M. Zhang, Y. Liu, H. Luan and M. Sun, Adversarial Training for Unsupervised Bilingual Lexicon Induction, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2017. doi:10.18653/v1/p17-1179.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190312

ERST: Leveraging Topic Features for Context-Aware Legal Reference Linking

Sabine WEHNERT [1], Gabriel CAMPERO DURAND and Gunter SAAKE
University of Magdeburg, Germany

Abstract. As legal regulations evolve, companies and organizations are tasked with quickly understanding and adapting to regulation changes. Tools like legal knowledge bases can facilitate this process, by either helping users navigate legal information or become aware of potentially relevant updates. At their core, these tools require legal references from many sources to be unified, e.g., by legal entity linking. This is challenging since legal references are often implicitly expressed, or combined via a context. In this paper, we prototype a machine learning approach to link legal references and retrieve combinations for a given context, based on standard features and classifiers, as used in entity resolution. As an extension, we evaluate an enhancement of those features with topic vectors, aiming to capture the relevant context of the passage containing a reference. We experiment with a repository of authoritative sources on German law for building topic models and extracting legal references and report that topic models do indeed contribute in improving supervised entity linking and reference retrieval.

Keywords. reference linking, entity resolution, topic models, information retrieval

1. Introduction

Nowadays, institutions and businesses face the challenge of understanding the implications of legal changes, as they occur. Often multiple experts for each jurisdiction monitor a broad spectrum of legal texts, which is a challenging task. In such work, the context of a legal entity and the current situation are determining the applicability of laws. Legal knowledge bases support users in understanding such contexts, drawing out their implications. However, the development of such systems is complex, since they often rely on hand-crafted domain knowledge, thus do not scale well and are difficult to maintain. Explainable machine learning methods are a promising alternative, as they can be efficient in large data analysis. In previous work [1], we introduced a method of extracting bottom-up domain knowledge from legal literature. This approach allowed us to leverage a diverse array of authoritative resources in the field, supporting our main goal of capturing context-dependent application of laws, by using keywords, chapter and section titles in the proximity of a cited law. In our work the extracted knowledge is represented by several concept hierarchies (one per book). Hierarchies need to be aligned, allowing the complete information about entities, as spread across the diverse information sources, to

[1]Corresponding Author: Sabine Wehnert, University of Magdeburg; E-mail: sabine.wehnert@ovgu.de. The work is supported by Legal Horizon AG, Grant No.:1704/00082 and by the DFG (Grant No.: SA 465/50-1).

be connected. For this linking task, we focus in this work on legal citations and refer to these entities as *references*. Fig. 1 depicts differently complex ways in which legal refer-

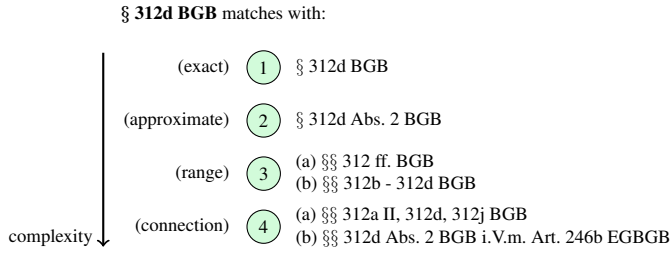


Figure 1. Complexity levels of legal reference patterns, illustrated by a reference to section 312d in the German civil code (BGB). Our linking task consists of detecting other references pointing to the same section.

ences can be found. Detecting these references, and linking them across sources is one of the core challenges in developing a bottom-up legal information system. The first kind of references are *exact* matches without the need for complex identification procedures. The second level represents *approximate* matches between references to the same section, where one reference can be more specific. Given the variability of references, applying approximate string matching could be cumbersome. The third level refers to references with a specified *range*, so that all elements within that range also need to be identified, as mentioned. The most complicated level are references comprised of multiple laws forming statute chains, with their sections indicated by one of the previous three levels. These references are only relevant in certain contexts. In addition, there might be references expressed in informal language, which refer implicitly to certain laws. These references too are highly complex and need to be identified to use the information about them. In sum, the different levels of complexity in expressing legal references pose a challenge for linking references and building legal information systems. In this paper we evaluate a machine learning solution to handle references, considering the complexity levels. We focus on two tasks: First, identifying references pointing to the same legal text (e.g., a norm) and second, retrieving valid references for a given context. In supporting these tasks, we study the applicability of topic models. Retrieving characteristic keywords for a document within a corpus is often solved by topic modeling. After grouping the documents into a given number of topics, the elements within each topic share common characteristics represented by their likelihood of containing certain keywords. In this paper we evaluate whether there are benefits to the two aforementioned tasks by extending each identified reference with a topic model vector corresponding to the text window in which the reference occurred. Our contributions are methodological and summarized by:

- First, we identify requirements for context-dependent legal reference linking: **E**xplainality, **R**eliability, **S**tability and **T**opical Relevance **(ERST)**.
- Second, we extract and resolve legal references found in German legal literature, in a supervised setting, showing the usefulness of adding topic modelling features. We report, across several types of classifiers, that topic features assist in legal entity resolution, when combined with standard features.
- Third, we combine traditional retrieval methods with topic features for legal reference retrieval, in an unsupervised retrieval setting. We report that topic features can indeed improve the relevance of returned laws with respect to the context.

The remainder of this paper is structured as follows: In Sec. 2, we introduce our requirements for bottom-up knowledge base alignment. We then describe our method of using topic modeling features to improve entity resolution and retrieval. Sec. 3 contains results of two experiments, and their respective evaluation. In Sec. 4, we build connections to other research covering the role of rule-based legal document annotation, similarity functions for legal entity resolution, probabilistic topic modeling and legal information retrieval. In Sec. 5 we conclude our work, motivating future research directions.

2. ERST Requirements for Legal Reference Management

In our compliance checking use case, we have a high recall requirement because missing even a single regulation can lead to high costs. In fact, building on this overriding requirement, we can identify a set of requirements for bottom-up knowledge base alignment: explainability, reliability, stability and topical relevance.

Explainability: We require *explainability* (i.e., the outcome of an application can be reasonably interpreted) in two regards: ground-truth generation and in the actual application. While a common demand is the explainability of applications, the ground truth which is used to train the algorithms should also contain an explanation (e.g., for the target label). The intuition behind this requirement is to provide enough resources to understand the original thought process leading to the label. This assists in feature engineering and designing applications that can offer the same level of explanation as the ground truth. If the ground truth is generated with rules, explainability can be easily achieved by indicating the rule which generated the instance. Another aspect of explainability are the features used by the application. While feature importance is easily determined in trained models, the choice of features can also be based on explainability.

Reliability: The purpose of legal entity resolution is the matching of legal named entities, such as person, organization, location and reference to their mentions in natural language text. More precisely, we frame this as a linking task of recognizing whether two mentions refer to the same entity. We distinguish legal reference entity types from other entities because the amount of variation in the citation pattern is not only restricted to common resolution cases, such as the use of abbreviations compared to the whole word. Legal references can be very specific, occasionally pointing to a part of a sentence in an article's paragraph. Our goal is to resolve references on an article basis, despite differences in citation granularity, see Figure 1. We name the requirement from our similarity function to properly convey matches, giving high recall, as *reliability*.

Stability: Given a collection of real-world documents, it is natural to assume that they could by grouped by underlying semantic themes. Topic modeling is a broad term that covers a series of statistical methods to describe documents according to such latent semantic groups. Through such methods each document in a collection can be described as a multinomial distribution over a number of discrete topics, while topics themselves are represented as multinomial distributions over a series of keywords. As a consequence, modeled topics can be compared by their probability of including given keywords, and documents can be compared and grouped by their probability of including a given topic. Some popular methods for building topic models are Latent Semantic Analysis, Latent Dirichlet Allocation (LDA), Correlated Topic Models and Non-negative Matrix Factorization. Building a topic model multiple times on the same corpus can lead to very dif-

ferent results: deviations in the top keywords per topic and their rank. For example, the use of Gibbs Sampling, Expectation Maximization or Variational Bayesian Inference for approximately inferring the distribution parameters that characterize an LDA model, are expected to converge to stable & reproducible results, however this might fail to occur based on the training (e.g., its duration) and model configuration (e.g. the number of topics chosen). Such variation is not desired for legal entity linking because a variation in topic quality can affect the overall use and interpretability of topic features. We therefore require measures to ensure *stability* when topic features are used. Some starting points to assist in studying the stability of topic modeling are hyperparameter optimization strategies [2] and evaluations over repeated runs (measuring coherence, perplexity).

Topical Relevance: Statutes are written in abstract, legal jargon to be applicable to many situations. The previously described references containing statute chains are only relevant in few situations. Having those references in a knowledge base, our goal is to align them only to those references that are sharing highly related concepts, thus satisfying a requirement we call *topical relevance*. An example of topical relevance is the reference "§ 286 BGB i.V.m. § 280 BGB", which specifies the breach of a duty combined with a default of the obligation. There are many contexts, in which those regulations can apply, such as the non-issuance of a job reference or the failed transfer of an asset against the negotiated terms. Those two situations occur in different settings, so that the topical connection to other references concerning labour law is only given in the former. For our use case of context-aware legal reference linking, the *topical commonalities* between the surrounding contexts of two references determine the likelihood of a connection, regardless of reference type. Since we consider a bottom-up knowledge acquisition process of concepts related to legal references, the contexts are available in natural language.

Legal Reference Management under ERST: In the following, we explain our methodology for legal reference management enhanced with topic models. We elaborate upon the legal reference resolution task, and then show how we enable context-aware legal reference retrieval. Figure 2 illustrates the workflow. First, we preprocess the legal literature corpus *(1)* and a document, which is compared to the remaining corpus to detect matching reference pairs. This document contains besides the legal references also the context in which they are considered. In the second step, we apply topic modeling on the literature corpus *(2a)* and annotate laws *(2b)* in the query document. After the annotation, context windows *(3)* around all legal references are extracted. Then, we use those context windows to infer a feature vector *(4a)* with the topic model. The references themselves are also featurized *(4b)* regarding the capitalization of the first token *(CAP)*, the length of the whole reference *(LEN)*, the type of reference *(TYPE)* and the token set similarity *(TSS)*. Finally, we train a model to link legal references *(5a)* using the features and can retrieve *(5b)* contextually relevant laws. We satisfy the requirement for *explainability* by first, using a rule-based approach for document annotation and second, identifying matching entities with rules for generating the ground truth in legal entity resolution. The purpose of this experiment is not to replicate the ground truth which is limited to the patterns indicated by the rules, but to analyse how well the models perform for differently complex types. Considering *reliability*, the similarity between two strings shall be detected regardless of length (due to differences in granularity) and order (due to different citation styles). For this, we apply token set similarity, as done by Cohen [3], for reference string comparison. This method is comparing the intersection (t_0) and the remainders of two sorted sets of strings (t_1, t_2) concatenated with t_0 against each other.

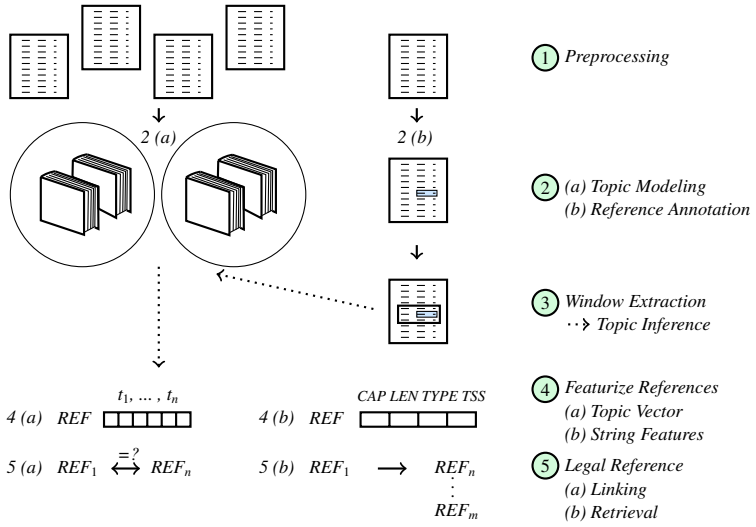


Figure 2. Overview of the featurization workflow for reference linking and retrieval.

The strings can have a different length because the comparison is allowed to end at the length of the shorter sequence. The *token set similarity (TSS)* is computed as follows:

$$\mathrm{TSS} = \max\left(\frac{|t_i| + |t_j| - \mathscr{L}}{|t_i| + |t_j|}\right), \quad i,\ j \in \{0,1,2\}, \tag{1}$$

where $|t_i|$ and $|t_j|$ are placeholders for the strings to be compared and $\mathscr{L}$ is the Levenshtein distance [4] between the strings. We consider the length of the intersection between both strings $|t_0|$, and the lengths of the two strings, $|t_1|$ and $|t_2|$, respectively. Three combinations are compared, $(t_0,\ t_1)$. $(t_0,\ t_2)$ and $(t_1,\ t_2)$ and the maximum score is the *TSS*. When we compute the token set similarity for the example strings from Figure 1, a score of 100 is returned for all strings except for 3 (a), where a score of 78 is obtained. In this case, the character "d" is missing and thus a match between §§ *312 ff. BGB* and § *312d BGB* is implied by the abbreviation *ff.*, referring to the following articles until the end of the section. That shows that the token set similarity is well suited for partial string matching irrespective of string length. For harder cases, such as 3 (a) (i.e., with the use of *ff.*), background knowledge is needed to resolve the correct number of regulations following. It is worth noting that we do not consider token order. That assumption may not hold for references of type 4 (b) (i.e., combinations), where the connections between the law books (e.g., BGB and EGBGB) and the respective section numbers (e.g., 312d and 246b) should not be lost. For those cases, the substrings of each reference can be matched separately with the token set similarity. Aside from the token set similarity, topic features are used for entity resolution because we assume semantically overlapping content across books. Having topic models in a productive setting, they shall be optimized regarding *stability*, in order to be interpretable and maintainable. For this, we refer to the different techniques summarized in Section 4. Given those preconditions, *topical relevance* can be a helpful indicator for identifying references pointing to the same entity in similar contexts. For the specific task of pair-wise classifier-based entity resolution, where classifiers are responsible for predicting if a pair is a match or

Table 1. Distribution of reference types based on extraction rules.

Norms	Court Decisions	EU-Directives	EU-Regulations	Combinations
83,661	4,277	3,869	730	122

non-match (i.e., same entity or not), traditionally similarity/distance-based features are used. In our case, we employ as features the absolute difference between topic vectors of the paired instances. Together with features capturing the capitalization of the first token *(CAP)*, the length of the string *(LEN)* and the type of reference *(TYPE)* - as shown in Table 1, the feature vector for entity resolution is formed. We train common classifiers on the binary classification problem. Another perspective on knowledge base alignment is the retrieval task. Here, we detect references with the same surface features using *TSS* and rerank the instances based on a reference context obtained from a query. The size of the context depends on the density of entities found in the corpus. We lemmatize the tokens and infer the topic vectors using the LDA model for each reference context. We reorder all retrieved references with the topic vector distance to the querying reference, thus increasing *topical relevance*.

3. Results and Evaluation

Evaluation Setup: For our experiments, we use a corpus of 193 German books which we manually grouped into 30 categories by their title, such as IT Security Law, Labour Law and Commercial Law. To obtain a similar granularity from the topic model, we run LDA for 200 iterations with a standard parameter configuration, setting the number of topics to 30. We specifically select an approach to LDA supported by Variational Bayes optimization as proposed by Hoffman et al [5], offering a reasonable runtime to facilitate repeated studies[2]. Table 2 provides representative words for the obtained topics from LDA. Notable outlier topics are criminal activities (28), chemicals (29) and consumers (30). There are significant overlaps between many topics, such as Credits (6) and Patent law (7). Since we could give all the topics an unambiguous label, we refrain from further optimization in this study. For optimal results and in productive settings, we nevertheless recommend to optimize LDA regarding topic stability (see Section 4). We adapted the reference extraction rules from previous work [1] to the Apache UIMA Ruta annotation tool[3] and extended them to other reference types [4]. Empirical checks resulted on average at roughly 90% reference coverage. Our reference annotation patterns are based on regular expressions and constrained by part-of-speech tags (POS). Hence, we obtain a distribution of references, as shown in Table 1. The 83,661 found norms cover patterns similar to the examples 1 - 4 (a) in Figure 1. We extracted 4,277 court decisions, such as *"EuGH NJW 2006, 2465"*. Most of the 3,869 entities of type EU-Directive occur in the following shapes: *"RL 29/2005/EG"* or *"Richtlinie über den elektronischen Geschäftsverkehr"*. Among the 730 EU-Regulations, common forms are *"VO 267/2010/EU"* and *"Verordnung über die Freizügigkeit der Arbeitnehmer"*. Combinations can contain all reference types, separated by an *"i.V.m."* (meaning: "in connection to"), see type 4 (b) in Figure 1.

[2]Gensim multi-core LDA: https://github.com/RaRe-Technologies/gensim
[3]https://uima.apache.org/ruta.html
[4]Implementation: https://github.com/anybass/HONto/tree/master/reference_linking

Table 2. Topics, given names and representative words in our corpus

Topic Nr.	Given name	Representative words
1	*International business*	abs, bgb, hgb, europäisch, bag, unternehmen, arbeitnehmer, betrvg, mitgliedstaat, corporate, kosten, international,...
2	*Compliance*	unternehmen, dabei, sowei, management, neu, daten, compliance, hoch, weit, beispiel, stellen, informationen,...
3	*Employment law*	bag, arbeitnehmer, arbeitgeber, betrvg, nza, bgb, kündigung, betriebsrat, arbeitverhältnis, gelten, tarifvertrag, besehen,...
Remaining topics	*Stock Enterprises (4), Commerce (5), Credits (6), Patent law (7), European law (8), Data privacy (9), Energy (10), Trade taxes (11), Income taxes (12), Traffic/Infrastructure (13), Business ethics (14), Commercial code (15), Insurance (16), Environment (17), Vacations/Working hours (18), Cyber-security (19), Control mechanisms (20), Stock market (21), Business taxation (22), Health (23), E-mobility (24), Audits (25), Online communication (26), Corporate governance (27), Criminal activities (28), Chemicals (29), Consumers (30)*	

Experiment 1 (Legal Reference Resolution): Following the steps in Fig. 2, we identify 92,659 references in our aforementioned legal literature corpus, corresponding to a natural occurrence of references to different types of legal entities (as shown in Table 1), and of the different complexity levels described in Fig. 1. Based on domain rules and an extent of manual verification, we identify 7,459,674 pairwise matches (i.e., only 0.173% of all possible matches). We split these matched pairs into training and test data (66%, 33%), randomly sampling from the non-match classes until the same number of items as the matched class is reached per split (i.e., for having balanced examples), and checking that non-matched pairs are not repeated. This leads to a test-train split of 5,307,699 / 9,576,279 labeled items. In terms of features, we enhance each reference with a topic vector that captures the probability of topic assignations using the window of 200 characters surrounding a reference (rounded up to complete words). Next, we use string features (i.e., CAP, LEN, TYPE and TSS, as mentioned in Section 2)), topic features (the absolute difference on each dimension of the topic vectors of the paired references) and a combination of topics and standard features. Table 3, shows the features as standard, topic model and combined, respectively. We evaluated the contribution of each feature for the supervised entity linking task; so we selected 4 different classifiers. As a baseline we use a Gaussian Naive Bayes (GNB) classifier (with no priors on the class distributions), due to its simplicity and few requirements on hyperparameters. We select random forest-based methods: XGBoost (XGB, eta: 0.3, max depth: 6, alpha: 0, lambda: 1), AdaBoost (Ada, decision-tree-based, max depth: 1, 50 estimators, lr: 1) and RandomForest itself (RF, with bootstrapping, using GINI criteria, min samples for split: 2, no depth limitations), due to their computational efficiency and potential for explainability. The overall F1 score shows a consistent trend of improving with topic model features and the combination with the standards. GNB performs the worst. RF performs the best, followed by XGB. When considering the scores of the entity types, it is shown that topic features alone cannot bring improvements in several of the classifiers evaluated. The only cases where the combination of feature types brings disadvantages are for our weakest classifer (GNB), or for the combination reference types, which constitute a little-represented class. Though the RF combined model is overall the best, with a consistent performance

		F1-Score	Norms	Court Dec.	EU-Dir.	EU-Reg.	Comb.
Standard	XGB	0.81	0.84	0.79	0.85	0.91	0.91
	Ada	0.81	0.83	0.77	0.83	0.91	0.91
	RF	**0.86**	**0.87**	**0.86**	**0.90**	**0.98**	**0.98**
	GNB	0.69	0.57	0.78	0.72	0.77	0.75
Topic model	XGB	0.83	0.83	0.76	0.86	0.84	0.79
	Ada	0.81	0.82	0.76	0.87	0.84	0.77
	RF	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**
	GNB	0.77	0.75	0.54	0.75	0.69	0.68
Combined	XGB	0.89	0.90	0.93	0.91	0.94	0.89
	Ada	0.89	0.89	0.79	0.93	0.91	0.88
	RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00!**	**1.00!**
	GNB	0.78	0.76	0.78	0.74	0.76	0.78

Table 3. F1-score and entity type-based accuracy for supervised legal entity resolution on our dataset, considering different types of features and classifiers. The exclamation mark indicates zero mislabeled entities.

across all classes, we note that in spite of having a grouped F1 score of 1.00, 6,581 norm instances, 3 court decisions and 16 EU-Directives were part of mislabeled pairs (out of the 5M tested pairs). Results suggest that there is room for improvement and serving better the less represented reference type is important for our approach to contribute to the overall performance of reference linking. Common error causes are ranges, missing whitespaces, errors from extraction rules and different citation granularities.

Experiment 2 (Retrieval of Context-Dependent Reference Connections): In this experiment, we test whether topic features can help to increase the relevance of retrieved references. We frame the task with a reranking objective and compute the distance between the topic vectors via *Jensen-Shannon Divergence (JSD)* [6] and *Maxium Absolute Difference (MAD)*. The Mean Absolute Difference behaved similar to JSD in our earlier experiments, so that we employ MAD instead. We randomly draw 14 queries from 122 references of the narrow context reference group 4 (b) (see Figure 1) consisting of the topic features and the reference. For these queries, we use TSS to generate candidates from all references and compute the topic-based distances. The ground truth is created by manually assigning a binary relevance label to all references returned by TSS (ranging from 8 to 246 hits), given their natural language contexts. We use r-Precision for evaluation, which returns the precision at position r where all relevant documents have been retrieved. Results indicate that it is worthwhile to rerank the data with topic features to obtain more relevant output. The best individual score was MAD with 62.3%, followed by JSD with 60.4%. The *Term Frequency - Inverse Document Frequency (TFIDF)* baseline yielded a score of 51.6%. A combination of MAD with TFIDF achieves the best r-Precision of 63%, whereas all metrics combined achieve a lower score of 61.1%. We observe a variance of the r-Precision regardless of the amount of candidates, that we attribute to a different granularity of the relevance label that the topic model does not serve.

Table 4. r-Precision based on JSD, MAD and TFIDF and combinations on 14 queries over our dataset.

rP(TFIDF)	**rP**(JSD)	**rP**(MAD)	**rP**(TFIDF, JSD)	**rP**(TFIDF, MAD)	**rP**(TFIDF, JSD, MAD)
0.516	0.604	0.623	0.596	0.630	0.611

4. Related Work

With regards to the *ERST* requirements, we explore related research, covering topic models for entity resolution and legal information retrieval. Topic models are used for entity resolution for more than a decade, see similar work on Wikipedia by Pilz et al. [7], as well as a latent dirichlet model by Bhattacharya et al. [8]. We find that topic features are a suitable technique for context-aware legal reference linking. Considering *explainability*, Glaser et al. [9] develop a system for German legal texts which disambiguates named entities to semantic roles using templates. Similar to their work, we extracted legal reference entities by using rule-based methods in Apache UIMA Ruta. Legal named entity recognition and resolution has been studied by Dozier et al. [10] for entities of judges, attorneys, companies, jurisdictions and courts. They apply well-founded techniques for resolution, such as blocking and *reliable* string similarity metrics for each entity type and train a Support Vector Machine (SVM) classifier. Van Opijnen et al. perform legal entity linking by using national and European Law Identifiers (ELI), which we consider for follow-up work [11]. Computing entity context similarity based on word embeddings is a state-of-the-art approach, but it can hardly be interpreted. Traditional bag-of-words representations often oversimplify sensitive natural language tasks. We consider features from topic models to be a viable trade-off between both worlds. Topics capture the contextual use of words and distances at this level of abstraction are well interpretable, as shown by Yurochkin et al. [12]. They define the *Hierarchical Optimal Topic Transport (HOTT)* measure, based on the Word Mover's Distance [13] between the word distribution per topic and the optimal transport between documents as distributions of topics. The topic model LDA uses a random inference process and thus suffers from instability. Many authors have addressed *stability*, e.g., by proposing a combination with non-negative matrix factorization [2,14] or a search-based parameter optimization using differential evolution [15]. LDA performance is strongly affected by hyperparameter tuning, therefore for each corpus a different setting is recommended [15]. When the corpus is extended in the future, the topics of new documents are inferred from the existing model, or a new topic model can be computed, optionally with must-link and cannot-link constraints to preserve the original structure [16]. Considering *topical relevance*, there are similar challenges in legal information retrieval in identifying the same application context of legal references. The system by Kim et al. [17] is based on well-known retrieval methods: stopword removal, lemmatization and TFIDF. A common problem occurs when there is no lexical overlap between the query and the statute. Although word embeddings and their newer contextual variants (e.g., XLNet [18]) may be a solution to this problem, they need to be adapted to the legal terminology and trained on a sufficiently large corpus.

5. Conclusion and Future Work

In this work, we pose four requirements for bottom-up knowledge base alignment: explainability, reliability, stability and topical relevance. We describe how those requirements can be fulfilled and perform experiments on legal reference linking and contextual retrieval. We find a benefit of using topic feature vectors with standard similarity metrics for legal entity linking, which can generate further viable candidates for contextual retrieval. Hence we validate the methodological choice of leveraging topic models trained

on legal literature, for creating contextual features for reference linking and retrieval. A common challenge for feature creation of domain-specific text data is the absence of word embeddings trained on a representative corpus; our topic feature vectors are a viable choice for smaller corpora. Combining topic models with word embeddings, e.g., using the *HOTT* method by Yurokchin et al. [12] can be worthwhile to investigate. Regarding supervised reference linking: Blocking and understanding better the behavior for less represented types are good avenues for continuing this research. Current approaches for knowledge base alignment use graph and word embeddings, which we want to test in follow-up work [19].

References

[1] S. Wehnert et al., Concept Hierarchy Extraction from Legal Literature, in: *Proceedings of the ACM CIKM 2018 Workshops*, CEUR-WS.org, 2018, to appear.

[2] D. Greene et al., How many topics? Stability analysis for topic models, in: *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part I*, Springer-Verlag, 2014, pp. 498–513.

[3] A. Cohen, FuzzyWuzzy: Fuzzy string matching in python, 2011. Retrievedfromhttps://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/.

[4] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: *Soviet physics doklady*, Vol. 10, 1966, pp. 707–710.

[5] M. Hoffman et al., Online learning for latent dirichlet allocation, in: *advances in neural information processing systems*, 2010, pp. 856–864.

[6] J. Lin, Divergence measures based on the Shannon entropy, *IEEE Transactions on Information theory* **37**(1) (1991), 145–151.

[7] A. Pilz et al., Named Entity Resolution Using Automatically Extracted Semantic Information., in: *LWA*, 2009, p. KDML–84.

[8] I. Bhattacharya et al., A latent dirichlet model for unsupervised entity resolution, in: *Proceedings of the 2006 SIAM International Conference on Data Mining*, SIAM, 2006, pp. 47–58.

[9] I. Glaser et al., Named entity recognition, extraction, and linking in german legal contracts, in: *Internationales Rechtsinformatik Symposium*, 2018.

[10] C. Dozier et al., Named entity recognition and resolution in legal text, in: *Semantic Processing of Legal Texts*, Springer, 2010, pp. 27–43.

[11] M. Opijnen et al., Beyond the experiment: the eXtendable legal link eXtractor, in: *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL)*, 2015.

[12] M. Yurochkin et al., Hierarchical Optimal Transport for Document Representation, *arXiv preprint arXiv:1906.10827* (2019).

[13] M. Kusner et al., From word embeddings to document distances, in: *International conference on machine learning*, 2015, pp. 957–966.

[14] M. Belford et al., Stability of topic modeling via matrix factorization, *Expert Systems with Applications* **91** (2018), 159–169.

[15] A. Agrawal et al., What is wrong with topic modeling? And how to fix it using search-based software engineering, *Information and Software Technology* **98** (2018), 74–88.

[16] Z. Zhai et al., Constrained LDA for grouping product features in opinion mining, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2011, pp. 448–459.

[17] M.-Y. Kim et al., Statute Law Information Retrieval and Entailment, in: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ACM, 2019, pp. 283–289.

[18] Z. Yang et al., XLNet: Generalized Autoregressive Pretraining for Language Understanding, *arXiv preprint arXiv:1906.08237* (2019).

[19] B.D. Trisedya et al., Entity alignment between knowledge graphs using attribute embeddings, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 297–304.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190313

Computer-Assisted Creation of Boolean Search Rules for Text Classification in the Legal Domain

Hannes WESTERMANN [a,1], Jaromír ŠAVELKA [b], Vern R. WALKER [c], Kevin D. ASHLEY [b] and Karim BENYEKHLEF [a]

[a] *Cyberjustice Laboratory, Faculté de droit, Université de Montréal*
[b] *ISP, School of Computing and Information, University of Pittsburgh*
[c] *LLT Lab, Maurice A. Deane School of Law, Hofstra University*

Abstract. In this paper, we present a method of building strong, explainable classifiers in the form of Boolean search rules. We developed an interactive environment called CASE (Computer Assisted Semantic Exploration) which exploits word co-occurrence to guide human annotators in selection of relevant search terms. The system seamlessly facilitates iterative evaluation and improvement of the classification rules. The process enables the human annotators to leverage the benefits of statistical information while incorporating their expert intuition into the creation of such rules. We evaluate classifiers created with our CASE system on 4 datasets, and compare the results to machine learning methods, including SKOPE rules, Random forest, Support Vector Machine, and fastText classifiers. The results drive the discussion on trade-offs between superior compactness, simplicity, and intuitiveness of the Boolean search rules versus the better performance of state-of-the-art machine learning models for text classification.

Keywords. Artificial Intelligence & Law, Text Classification, Semantic exploration, Boolean search, Natural language processing, Explainable artificial intelligence

1. Introduction

Reading, interpreting, and understanding legal texts is one of the most important skills of legal professionals. Lawyers, judges, students, and researchers alike spend a lot of time and effort learning, honing, and using the skill in reading statutory law, a legal case, a contract, or a journal article, while interpreting the document and applying the knowledge to solve a new problem. Such analysis is done on several levels—sometimes, an individual sentence carries the much needed information, while other times the reader has to study whole sections or the entire document to understand the important point. Further, the reader may have to understand different features, such as the facts of a legal case or the relevance of a sentence.

The ability to categorize the texts or their pieces into certain types (e.g., court reasoning, legal rule, facts) is an integral part of the analysis. It is therefore no coincidence

[1]Corresponding Author: Hannes Westermann, E-mail: hannes.westermann@umontreal.ca

that text classification is one of the big focus areas in the field of artificial intelligence and law (AI & Law). Automating the classification tasks has often become feasible due to advances in machine learning (ML) and natural language processing (NLP) methods. Typically, the research is conducted by manually labelling hundreds or thousands of documents. Once the annotation is completed, the researchers use the data to build ML models that are able to learn patterns in the annotated data and apply these to classify new unseen texts.

An automated approach has a host of advantages when compared to the tedious work of classifying the whole corpus without the help of a computer. However, there are some drawbacks as well. Firstly, it still takes a lot of effort to manually label the subset of documents required for training an ML classifier. Secondly, it can be difficult to explain the decisions of sophisticated models. This may sometimes lead to skepticism as to the suitability of such models to be used in practice. Possible over-fitting is yet another risk which needs to be taken into account. Sometimes the models work well on the annotated data, but fail to generalize to unseen documents.

In this paper, we present a method addressing these issues. We built a tool that allows annotators to create Boolean rules in a computer-assisted fashion. These rules could potentially be used for classification in domains with little available data, by incorporating human intuition into the process. Further, the rules created are more explainable than most machine learning models, while still performing reasonably well.

2. Prior work

According to Antonie and Zaiane [1] "a good text classifier [...] efficiently categorizes large sets of text documents in a reasonable time frame and with an acceptable accuracy, and [...] provides classification rules that are human readable for possible fine-tuning." One approach to text classification is to let a human expert define a set of logical rules based on his domain-specific knowledge of how to classify documents under a given set of categories [2]. Generating rules based on human expertise is time-consuming, expensive, and sometimes not feasible. However, the great advantage of such rules is that they often provide intuitive and meaningful explanation (justification) of the resulting classification.

Alternatively, one can apply various methods for inducting text classification rules automatically including such methods as decision trees or associative rule mining [2]. The latter employs an iterative search of a database to discover the most frequent sets of k items (k-itemsets) that are associated with the documents sharing a particular classification; a logical rule based on a k-itemset should support the classification with a confidence above a certain threshold. The potentially very large number of rules are then pruned using various techniques [1]. A disadvantage of such automatically learned rules is that they may not correspond to expert intuitions about texts in the domain.

Various hybrids of manual and automated methods are possible. For example, Yao, et al. [3] evaluated a medical clinical text classification method that employed rules to identify trigger phrases such as disease names and alternatives. They used the trigger phrases to predict classes that had very few examples. For the remaining classes they trained a knowledge-guided convolutional neural network (CNN) with word embeddings and medical feature embeddings.

In Walker et al. [4], the authors investigated the task of automatically classifying, within adjudicatory decisions in the United States, those sentences that state whether the conditions of applicable legal rules have been satisfied or not ("Finding Sentences"), by analyzing a small sample of classified sentences (N = 530) to manually develop rule-based scripts, using semantic attribution theory. The methodology and results suggested that some access-to-justice use cases can be adequately addressed at much lower cost than previously believed. Our work extends that effort by developing a platform for efficiently improving the classification rules in the iterative fashion.

3. Boolean Search Rules

We propose and evaluate a novel hybrid combination of manual and automated construction of text classification rules. Our CASE system helps annotators select relevant terms, create Boolean text classification rules, and evaluate and improve them in an iterative manner. Depending on the use case, the resulting rules may prove very useful—especially where explanatory power and compactness are important.

"Boolean search rules" are an appealing method for classifying documents because such rules are familiar to anyone who works with legal information retrieval systems. They make it possible to search for single words (such as "veteran"), which would return all cases containing the word. Further, it is possible to logically combine several rules, using the OR, AND, and NOT operators. OR returns texts with either of the two words while AND requires both of them to be present. NOT excludes texts containing a particular word. In our case, we are using the FTS5 search engine integrated into the SQLite Database [5] to process our queries. This allows us to build complex queries, combining different logical operators, that are executed very rapidly.

3.1. Existing methodologies to create Boolean rules

There have been previous attempts of using Boolean search rules in AI & Law. However, without the methods presented in this paper the process can be long and laborious. A recent attempt at creating such search rules was made by Walker et al. [4]. The researchers tested whether distinctive phrasing in legal decisions enables the development of automatic classifiers on the basis of a small sample of labeled decisions, with adequate results for some important use cases. Certain words, such as "finds", were found to closely correspond to a sentence having the rhetorical role of a finding of fact. Two such rules were tested, leading to an F1 score of 0.512 in identifying such sentences. Testing new hypotheses, observing the results, and comparing the results of new classification rules against the old ones, was a time-consuming and laborious process. In this paper, we introduce a tool that makes such a process more efficient.

4. Methodology

In this paper, we test the hypothesis that Boolean search rules created by humans with the assistance of a computerized tool can prove useful in building text classifiers in the legal domain. To test the hypothesis we created such rules on four datasets of case texts, and compared the results to those obtained by using ML methods. The process is described in this section.

4.1. Datasets

We selected four existing datasets created within the AI & Law community to evaluate our methodology. These are presented below.

4.1.1. Veterans Claims Dataset (sentence roles)

Walker et al. [4] analyzed 50 fact-finding decisions issued by the U.S. Board of Veterans' Appeals ("BVA") from 2013 through 2017, all arbitrarily selected cases dealing with claims by veterans for service-related post-traumatic stress disorder (PTSD). For each of the 50 BVA decisions in the PTSD dataset, the researchers extracted all sentences addressing the factual issues related to the claim for PTSD, or for a closely-related psychiatric disorder. These were tagged with the rhetorical roles [6] the sentences play in the decision. We conducted our experiments on this set of sentences.

4.1.2. Court Decisions Segmentation Dataset (functional parts)

Šavelka and Ashley [7] examined the possibility of automatically segmenting court opinions into high-level functional parts (i.e., Introduction (I), Background (B), Analysis (A), Footnotes (F)) and issue specific parts (i.e., Conclusions(C)). They assembled 316 court decisions from Court Listener and Google Scholar, 143 in the area of cyber crime and 173 involving trade secrets. These were annotated, after which Conditional Random Fields (CRF) models were trained to recognize the boundaries between the sections. We used the cases in the area of cyber crime for our tests. It should be noted that we do not attempt to detect the boundaries, but instead try to classify the annotated text sections.

4.1.3. The Trade Secrets Factors Dataset (factor prediction)

Falakmasir and Ashley [8] assembled a corpus of 172 trade secret misappropriation cases employed in the HYPO, CATO, SMILE+IBP and VJAP programs. Legal experts had labeled the cases by the applicable factors, stereotypical patterns of fact that strengthen or weaken a claim. There are 26 trade secret misappropriation factors. For our experiments, we used the existence of security measures in a case (Factor 6), to deal with a binary classification task.

4.1.4. The Statutory Interpretation Dataset (interpretative value of sentences

Šavelka et al. [9] studied methods for retrieving useful sentences from court opinions that elaborate on the meaning of a vague statutory term. To support their experiments they queried the database of sentences from case law that mentioned three terms from different provisions of the U.S. Code. They manually classified the sentences in terms of four categories with respect to their usefulness for the interpretation of the corresponding statutory term. Here we work with the sentences mentioning 'common business purpose' (149 high value, 88 certain value, 369 potential value, 274 no value). In [9] the goal was to rank the sentences with respect to their usefulness; here, we classify them into the four value categories.

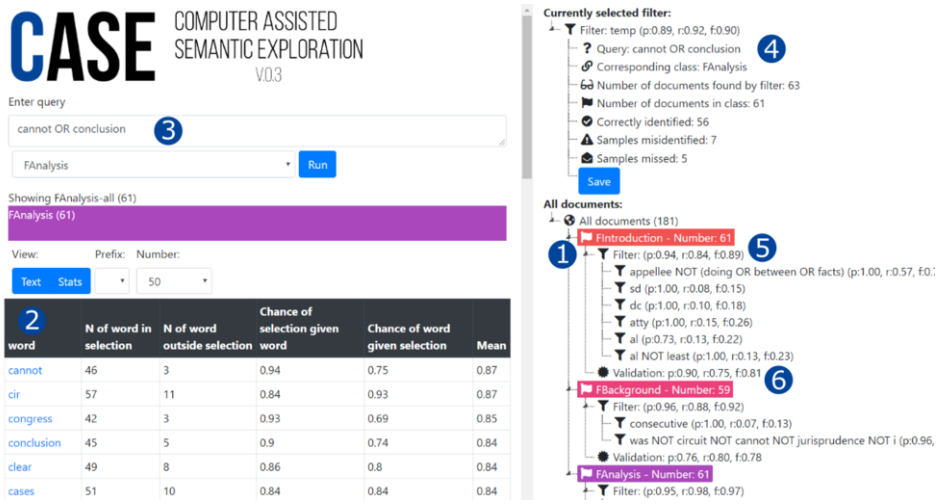


Figure 1. The Computer Assisted Semantic Exploration (CASE) interface.

4.2. Dataset Split

The four datasets are described in section 4.1. For our experiments we split each dataset into three parts: training (20%), validation (10%), and testing (70%). The unusual split (small training set) was used to evaluate the performance of the search queries in situations where very little data is available. This is often the case in the problems of interest in the field of AI & Law. Using the identical dataset splits we created classifiers with the CASE tool and ML methods, as described below.

4.3. CASE - Computer Assisted Semantic Exploration

We developed a tool for Computer Assisted Semantic Exploration (CASE). CASE facilitates seamless creation of Boolean classification rules. Figure 1 shows a screenshot of the interface. The tool supports users in interactively creating Boolean rules using several statistical methods. At (1), CASE displays the possible classes for annotation. By clicking on a class, the user selects texts in that specific class, and is then shown information about the word distribution inside that selection under (2). This list is sortable, and shows several headers, containing metrics useful for the selection of significant words.

Once a user has found a word that is a strong indicator of a specific class, he can create a query in (3), using logical operators such as AND, OR, and NOT. For example, a query to identify the class Analysis could be “cannot OR conclusion.” The query can then be run, and is immediately evaluated, with the results being presented in (4). Here, the user also has the possibility of selecting documents that are misidentified, for example, in order to exclude certain words. The user can thus work on creating queries able to identify classes with high precision and recall in an iterative fashion.

Once the user is content with a query, he can save it and create additional queries. Ideally, in conjunction, the queries will identify documents with high precision and recall. The filters are constantly evaluated against the training data (5), and validation data (6) to prevent over-fitting.

The queries used for the current paper were created by the co-authors of this paper. Using the statistics provided by CASE as well as our previous intuitions about the datasets, we used the tool to add and modify rules until it was difficult to introduce new rules without lowering the validation score. CASE was very helpful in identifying words significant for a class and using these to create the rules.

4.4. Machine Learning

We trained four different types of ML models as benchmarks. We used SKOPE-rules [10] to simulate the situation where the model is forced to construct similar Boolean rules as a human using CASE. The difference is that the rules are learned automatically. We trained random forest classifier, support vector machine (SVM), and fastText [11] models on more sophisticated features. Their predictions are used to investigate how much performance one has to sacrifice in order to benefit from the explanatory power of CASE (computer assisted) and SKOPE-rules (computer generated). We have used the same training sets as those in the CASE experiments to train the models. The validation sets were used to optimize the models' hyperparamters. The same test sets were used for the evaluation.

SKOPE-rules is a Python ML module the aim of which is learning logical, interpretable rules. [10] A decision rule is a logical expression of the form "IF conditions THEN response." The problem of generating such rules has been widely considered in ML, see e.g., RuleFit [12], Slipper [13], LRI [14], MLRules [15]. SKOPE-Rules extracts rules from an ensemble of trees. A weighted combination of these rules is then built by solving an L1-regularized optimization problem over the weights as described in [16]. To force the model to construct the rules that are comparable to those created using CASE, we have used unigram, bigram, and trigram word occurrences as features. The classification model is then a set of rules (possibly overlapping, i.e., OR), where each rule is a conjunction (i.e., AND) of matching or filtering (NOT) on words and phrases. For each data set we have trained a number of binary models, one for each class.

A random forest is an ensemble classifier that fits a number of decision trees on subsamples of the data set. It uses averaging to improve the predictive accuracy and control over-fitting. As an implementation of random forest we used the scikit-learn's Random Forest Classifier module [17]. As features we use TF-IDF weights of (1-4)-grams of lowercase tokens with their POS tags.

An SVM classifier constructs a hyper-plane in a high dimensional space, which is used to separate the classes from each other. As an implementation of SVM we used the scikit-learn's Support Vector Classification module [18]. We used the same features as with the random forest models to train a number of binary classifiers.

FastText is a linear classifier that uses ngram features that are embedded and averaged to form the hidden variable. We worked with the Python wrapper [19] for the original library released by Facebook [20]. As for other classifiers we trained a number of binary classification models using grid search to optimize hyperparameters.

4.4.1. Evaluation

The evaluation of all the models is performed on the test sets (70% of the respective datasets). Note that all the methods were trained on the identical training sets and fine-tuned on the identical validation sets. The performance is measured in terms of precision

Table 1. P, R and F_1 for the different classifiers applied on datasets described in Section 4.1

	CASE			SKOPE			RF			SVM			Fasttext		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
VetClaims															
-sentence	.84	.25	.38	.80	.33	.47	.90	.61	.72	.99	.44	.61	.87	.61	.72
-finding	.71	.38	.50	.63	.40	.49	.77	.26	.39	.83	.57	.67	.68	.59	.63
-evidence	.82	.74	.78	.71	.81	.76	.88	.88	.88	.90	.92	.91	.87	.92	.90
-rule	.71	.48	.57	.60	.65	.63	.95	.55	.70	.90	.78	.83	.87	.79	.82
-citation	.96	.99	.97	.86	.86	.86	.99	.96	.97	.98	.98	.98	.99	.97	.98
-reasoning	.62	.14	.22	.50	.23	.31	.65	.06	.12	.75	.27	.39	.43	.39	.41
-overall	.78	.50	.57	.68	.55	.59	.86	.55	.63	.89	.66	.73	.79	.71	**.74**
-overall-w	.80	.61	.67	.74	.71	.71	.88	.68	.73	.96	.83	**.87**	.83	.80	.81
Section segmentation															
-intro	.90	.75	.81	.83	1.0	.91	1.0	.98	.99	1.0	.99	.99	.98	.95	.97
-backg.	.76	.80	.78	.62	.96	.75	.97	.83	.90	.99	.85	.91	.96	.85	.90
-analysis	.87	.83	.85	.88	.88	.88	.98	.88	.93	.93	.97	.95	.90	.98	.94
-overall	.84	.79	.81	.78	.95	.85	.98	.90	.94	.97	.94	**.95**	.95	.93	.94
-overall-w	.84	.79	.82	.78	.95	.85	.98	.89	.94	.97	.94	**.95**	.95	.93	.94
Trade secrets															
-security	.65	.61	.63	.53	.97	**.69**	.59	.69	.64	.50	1.0	.67	.57	.49	.53
Statutory interpretation															
-high	.72	.45	.55	.66	.39	.49	.91	.10	.17	.96	.22	.36	.61	.46	.52
-certain	.18	.18	.18	.26	.23	.24	.67	.13	.22	.60	.10	.17	.40	.13	.20
-potential	.69	.36	.47	.49	.98	.65	.69	.54	.60	.71	.64	.67	.63	.68	.65
-no	.89	.65	.75	.74	.71	.73	.90	.77	.83	.90	.78	.83	.92	.79	.85
-overall	.62	.41	.49	.54	.58	.53	.79	.39	.45	.79	.44	.51	.64	.52	**.55**
-overall-w	.70	.44	.54	.57	.72	.61	.79	.50	.56	.80	.56	.62	.69	.62	**.64**

(P), recall (R), and F_1-measure (F_1). All the classifiers are evaluated in the one-vs-rest settings where each label within each of the four datasets has its own classifier. We measure aggregate results as well. "Overall" averages the scores for the different classes over the total number of classes. "Overall-w" uses a weighted average, where each class is given a weight according to how often it appears in the test dataset. For each dataset, the highest overall F_1-score is written in bold.

5. Results

The results are presented in Table 1. In general, the rules created by CASE performed similarly to the computer generated SKOPE rules. However, they seem to have a slightly higher precision, with a lower recall. This can have utility for certain use cases. As expected the more complex RF, SVM, and fastText models perform better than the human generated rules. We discuss the trade-offs between explainability and performance below.

Veteran Claims Dataset. Compared to the work in [4], the CASE tool gave us significant flexibility and speed improvements in creating and optimizing the Boolean search rules. For "finding sentences," for example, we confirmed the usefulness of the "finds" and "preponderance" search terms [4], while adding others such as "elements" and "warranted".

Table 2. Comparison between created rules for identifying the analysis section in the segmentation dataset.

CASE (f1: .85) (boolean rules)	SKOPE-rules (f1: .88) (boolean rules)	Random Forest (f1: .93) (important features)
cannot OR apparent OR prohibit OR definition	(cir AND NOT headquarters in AND is) OR (2d AND NOT february 17 AND is not) OR (NOT appeal AND NOT cir AND it is) OR (NOT cir AND is not AND NOT of 18)	is not, that, be, cir, cases, can, is, provides, held, it, statute, in, intended, record, see also, there, 9th cir, subsection, 7th cir, of such, have, may be, congress, when, issue, 3d at, evidence that, if, thus the, as, where, is to, here the, definition of

Court Decisions Segmentation Dataset. The CASE rules achieved higher precision, but lower recall than the SKOPE rules. The created Boolean search rules are quite simple. For identifying the analysis section, for example, the following query was quite successful: "cannot OR apparent OR definition OR prohibition."

Trade Secrets Factors Dataset. This dataset was the most difficult to deal with. There were few cases, and they were long and complex. For training, 33 cases were available. The rules achieved the highest precision among the classifiers. We relied heavily on human intuition, such as the term "non-disclosure" implying the existence of security measures. Building the rules also helped us identify an error in the annotation of a case.

Statutory Interpretation Dataset. This dataset was also very hard to deal with, due to its being unbalanced and the fact that the value of a sentence for statutory interpretation is hard to link to individual terms. Again, we can see the pattern of the CASE rules having higher precision than SKOPE rules, but lower recall.

5.1. Explainability of Rules or Features

Table 2 shows a comparison of rules created using the different systems for classification of the "analysis part" of the court decisions segmentation dataset. For the CASE and SKOPE rules, a document triggering any of the listed queries will result in the document being labeled as "analysis." For the random forest algorithm, we present the most important features, as selected by the algorithm. Overall, the CASE rules are much less complex, while still showing performance that is not much inferior. Further, the CASE rules seem to contain more legally relevant terms, such as "prohibit" and "definition." These properties make the rules easier to explain.

6. Discussion

We have shown that Boolean search rules can be created efficiently with a system such as CASE. In most areas, the performance was weaker than their ML counterparts. However, the CASE rules have advantages that might make their use desirable in some use cases. In this section, we discuss some of the advantages and disadvantages of using such Boolean search rules for classification in legal domains.

6.1. Advantages of using Boolean rules

One advantage of using Boolean rules, developed with the assistance of the CASE platform, is that those rules can *incorporate human intuitions*. Thus, the user can rapidly

formulate and evaluate hypotheses of which terms might prove useful in search rules, assisted by the statistical measures provided by CASE. In doing so, users are able to select words that they know have legal significance in relation to the classifier.

In incorporating this intuition, the user has significantly more *control* over the created model than with ML systems. With ML it is difficult to direct the training of the system, beyond feature selection and hyperparameter optimization. In CASE, on the other hand, the human is always in control of the system. Evaluation occurs continuously, and the human has complete control over how the model develops and how new search terms affect the precision and recall of the search rules. This allows users to fit the rules more exactly to their requirements and use case. Further, the creation process allows the annotator to develop an intuition for the particularities of the dataset in an exploratory fashion. In the trade secrets dataset, the system helped us to discover an error in classification, showcasing this advantage.

The incorporation of human intuition, together with the level of control a human is given over the creation of the search rules, can potentially allow the user to create search rules that are less prone to overfit and therefore *generalize* better. The user can choose to use only phrases that are independent of the specific context of the dataset, thereby creating rules that generalize to other datasets. Since the users decide whether to use a term, even very small datasets could support the creation of the rules with high precision.

A big issue in the practical use of ML in the legal field is the difficulty of *explaining* the created models. This might cause legal professionals not to trust the algorithms. Using Boolean search rules might alleviate this issue. Firstly, the human who makes the decisions in creating models, is fully aware of why a particular word was chosen and used in a certain way. Further, the structure of the created rules, using AND, NOT, and OR, should be easier to grasp than complex ML models. They can thus offer a basis for better explaining why a particular document was chosen, and why not. As can be seen in Table 2, the CASE rules are both simpler and more legally relevant than the ML models.

6.2. Limitations

As can be seen from the results presented in Section 5, ML models often performed better than the human-created rules. This is an interesting result in itself, as it shows the power of well-optimized ML methods even on small datasets. If performance is the most important metric, using ML methods could thus often be preferable. We discuss methods to combine advantages from ML and CASE below (Section 7).

7. Future work

This paper is an initial step in exploring the use of computer-assisted creation of Boolean search rules for text classification. There are many avenues for further research. One is to expand the CASE system. For example, the system could include n-grams beyond simple words. Restructuring the classifiers as multi-label classifiers, and running the best classifiers first, would improve performance. The system should also be expanded to work better with long documents. Another avenue is combining the CASE platform with ML methods in a hybrid approach to harness the advantages of both. For example, CASE could be used to preselect documents from a massive corpus, after which a ML algorithm

could be trained on only those documents. Another approach would be to run a ML model on annotated data, and use CASE subsequently to exclude false positives.

8. Conclusions

In this paper we have proposed and evaluated CASE, a novel approach for computer-assisted text classification using Boolean matching rules. We have shown that in a number of use cases the rules perform surprisingly well using little annotated data while offering superior explanatory power when compared to ML methods.

References

[1] M-L. Antonie, and O. R. Zaiane. "Text document categorization by term association." *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. IEEE, 2002.

[2] V. Korde, and C. N. Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3.2 (2012): 85–99.

[3] L. Yao, C. Mao, and Y. Luo. "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks." *BMC medical informatics and decision making* 19.3 (2019): 71, pp. 31–39.

[4] V. R. Walker, K. Pillaipakkamnatt, A. M. Davidson, M. Linares, and D. J. Pesce "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning." *Proceedings of ASAIL 2019* (2019).

[5] D. R. Hipp, D. Kennedy, J. Mistachkin, SQLite FTS5 Extension (2019, September 20) *SQLite* https://sqlite.org/fts5.html

[6] V. R. Walker, J. H. Han, X. Ni, and K. Yoseda. "Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans' claims dataset." *Proceedings of ICAIL '17*. ACM, 2017.

[7] J. Savelka, and K. D. Ashley. "Using CRF to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection." *ASAIL 2017*.

[8] M. H. Falakmasir, and K. D. Ashley. "Utilizing Vector Space Models for Identifying Legal Factors from Text." *JURIX*. 2017.

[9] J. Savelka, H. Xu, and K. D. Ashley. "Improving Sentence Retrieval from Case Law for Statutory Interpretation." *Proceedings of the ICAIL '19*, pp. 113–122. ACM, 2019.

[10] SKOPE-Rules. (2019, September 11). github.com/scikit-learn-contrib/skope-rules

[11] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov. "Bag of tricks for efficient text classification." *preprint arXiv:1607.01759* (2016).

[12] J. H. Friedman, and B. E. Popescu. "Predictive learning via rule ensembles." *The Annals of Applied Statistics* 2.3 (2008): 916–954.

[13] W. W. Cohen, and Y. Singer. "A simple, fast, and effective rule learner." *AAAI/IAAI* 99 (1999): 335–342.

[14] S. M. Weiss, M. Sholom, and N. Indurkhya. "Lightweight rule induction." (2000).

[15] K. Dembczyński, W. Kotłowski, and R. Słowiński. "Maximum likelihood rule ensembles." *Proceedings of the 25th international conference on Machine learning*, pp. 224–231. ACM, 2008.

[16] J. Friedman, and B. E. Popescu. *Gradient directed regularization for linear regression and classification*. Technical Report, Statistics Department, Stanford University, 2003.

[17] Random Forest Classifier. (2019, September 13). *scikit-learn*. scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[18] Support Vector Classification. (2019, September 13). *scikit-learn*. scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[19] FastText. (2019, September 11). *Github repository*. Retrieved from github.com/facebookresearch/fastText/tree/master/python

[20] FastText. (2019, September 11). https://fasttext.cc/

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190314

Neural Network Based Rhetorical Status Classification for Japanese Judgment Documents

Hiroaki YAMADA [a], Simone TEUFEL [a,b] and Takenobu TOKUNAGA [a]
[a] *School of Computing, Tokyo Institute of Technology, Japan*
[b] *University of Cambridge, Computer laboratory, U.K.*

Abstract. We address the legal text understanding task, and in particular we treat Japanese judgment documents in civil law. Rhetorical status classification (RSC) is the task of classifying sentences according to the rhetorical functions they fulfil; it is an important preprocessing step for our overall goal of legal summarisation. We present several improvements over our previous RSC classifier, which was based on CRF. The first is a BiLSTM-CRF based model which improves performance significantly over previous baselines. The BiLSTM-CRF architecture is able to additionally take the context in terms of neighbouring sentences into account. The second improvement is the inclusion of section heading information, which resulted in the overall best classifier. Explicit structure in the text, such as headings, is an information source which is likely to be important to legal professionals during the reading phase; this makes the automatic exploitation of such information attractive. We also considerably extended the size of our annotated corpus of judgment documents.

Keywords. Japanese NLP, Legal NLP, Argument understanding, Machine learning, Sentence classification, Natural language processing, Neural network, Deep learning, Rhetorical status classification

1. Introduction

Like in all other areas of life, information overload has also become problematic in the legal domain. Legal practitioners, including lawyers and judges, need to find relevant documents for their cases, and efficiently extract case-relevant information from them. In the Japanese legal system, one of the main sources used for this task is the judgment document, an important type of legal document which is the direct output from court trials and contains the judgment, the facts and the grounds[1,2]. They are typically long and linguistically complex, so that it becomes impossible to read all relevant documents carefully. Summaries of judgment documents are a solid solution to the problem, as they would facilitate the decision which documents the legal professionals should read with full attention. Our final goal is to develop methods for automatically generating such summaries.

Our project is based on the observation that the structure of the legal argument can guide summarisation. In the Japanese judgment documents, a common structure exists (Figure 1), which centres around the so-called "Issue Topic," a legal concept corresponding to pre-defined main points which are to be discussed in a particular court case. An example for a legal case about a damage compensation case of a traffic accident in a

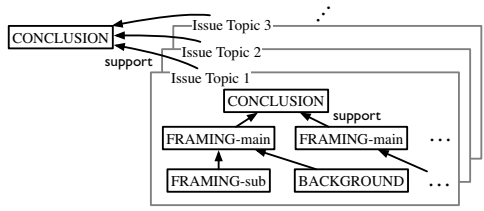


Figure 1. Argument structure of judgment document

bus travel is the question of the degree of plaintiff's own negligence. A case consists of several Issue Topics (three in the figure), and each is associated with a conclusion by the judge, and with supporting arguments for the decision. The task of argument structure extraction can be divided into four subtasks [3]: 1. *Issue Topic Identification*: find sentences that describe an Issue Topic; 2. *Rhetorical Status Classification*: determine the rhetorical status of each sentence; 3. *Issue Topic Linking*: associate each sentence with exactly one Issue Topic; 4. *FRAMING Linking*: link two sentences if one provides argumentative support for the other.

In this paper, we focus on Rhetorical Status Classification (RSC), the task of classifying sentences according to their rhetorical role (e.g. BACKGROUND or CONCLUSION). In the legal domain, this task is often seen as a preprocessing step for later tasks such as legal information extraction, extractive summarisation and argument mining [4,5,6]. We define seven RSC categories as follows; Table 1 lists them and gives an example for each. FACT covers descriptions of the facts giving rise to the case; BACKGROUND is reserved for quotations or citations of law materials (legislation and relevant precedent cases); CONCLUSION marks the decisions of the judge; and IDENTIFYING is a category used for text that states discussion topics. The primary argumentative material is contained in the two categories FRAMING-main and FRAMING-sub. FRAMING-main marks material which directly supports the judge's conclusion, whereas FRAMING-sub is one of the two categories which can support FRAMING-main (the other being BACKGROUND). These categories are crucial for downstream argumentative structure mining (task 4). Material that can not be classified into any of the above classes is covered by the OTHER category.

In our previous work, RSC performance was acceptable overall, but differed across category: in particular, in some of the most important categories for downstream tasks, performance was low. BACKGROUND, which is an important category listing relevant law materials, achieved only F=0.32, and CONCLUSION, which describes the most important argumentative sentences, F=0.39. We were also not fully satisfied with the performance of the two FRAMING categories.

In this paper, we present our improved RSC classifier for Japanese judgement documents, which uses a neural network-based architecture. One of the new information sources for our model is information coming from headings in the text. Our method is motivated by human readers' scanning behaviour during reading. We also present our new, considerably larger annotated corpus of Japanese judgements.

2. Data and Annotation

The corpus we used in previous work [3] consists of 89 Japanese judgment documents of Civil law cases from lower court cases with annotations of argumentative structure.

Table 1. Examples for RSC categories

Label	Example (translated)
IDENTIFYING	*Based on the agreed facts and the gist of the whole argument, we discuss each issue in the following.*
CONCLUSION	*Therefore, the plaintiff's claim is unreasonable since we just found that the officer was not negligent.*
FACT	*The duties of an execution officer are... and officer D properly conducted...*
BACKGROUND	*It is reasonable to find the officer negligent when the officer did not the appropriate... (1997/7/15 ruling of the third Petty Bench of the Supreme Court).*
FRAMING-main	*The measures performed by the officer comply with the normal procedure for inspection.*
FRAMING-sub	*It is considered that officer D entered the estate to confirm the circumstance...*

Table 2. RSC class distribution of our corpora in percent

	FACT	FR-main	FR-sub	CONC	IDEN	BACK	OTH	sent.
Previous (89 doc)	23.1	19.5	11.5	3.9	2.1	0.3	39.7	37,371
New (t&t, 110 doc)	23.5	19.1	10.6	3.8	2.0	0.3	40.6	44,677

They were sourced from website maintained by the supreme court of Japan[1] by a random selection process. Our new corpus extends this set to 120 documents (48,370 sentences, 3.2 million characters) following the same principles, and the same expert annotator (a PhD candidate in a graduate school of Japanese Law, who was paid for this work) was used. The annotation is kept consistent with the preceding paper, i.e, annotations for all four subtasks above are obtained at the same time. Category assignment is exclusive, i.e., only one category can be assigned to each sentence. We reserved ten documents out of 120 documents as development data for hyperparameter tuning. The remaining 110 documents are used for the experiments reported here. Table 2 shows the category distribution and total for our test and training corpus of 110 documents, against the previously used test and training corpus of 89 documents.

3. Conditional Random Field baseline model

Previous work on RSC in legal documents found that RSC is strongly affected by context in terms of other rhetorical roles[3,5]. We therefore use Conditional Random Fields (CRF) [7][2] as a strong baseline model.

As features, we use the seven features from [3]: the **bigram**, **sentence location**, **sentence length** features (calculated in characters). We also use 8 **modality** features based on Masuoka's (2007) modality expression classification, namely the modalities "truth judgment" (4 features; e.g, "*hazu da*" (can be expected to be) or "*beki da*" (should be)), "value judgment" (3 features), and "explanation". The **function expression** feature distinguishes the 199 semantic equivalence classes contained in the function expression dictionary by [10](such as "evidential" and "contradictory conjunction"); this covers 16,801 separate surface expressions. The **cue phrase** feature contains an additional 22 phrases

[1] `http://www.courts.go.jp/`

[2] We used Okazaki's (2011) implementation.

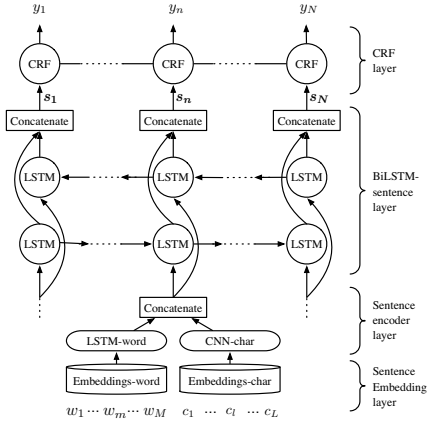


Figure 2. BiLSTM-CRF model for RSC.
$\boldsymbol{w}$: words from an input sentence;
$\boldsymbol{c}$: characters from an input sentence;
$\boldsymbol{s}$: contextualised vectors of sentences;
y: predicted RSC category.

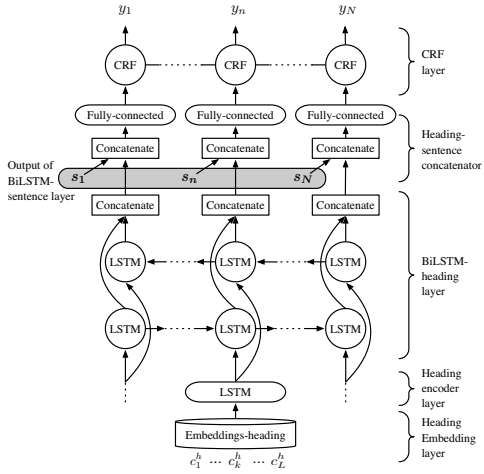


Figure 3. BiLSTM-CRF model with heading.
$\boldsymbol{c^h}$: characters from an input heading;
$\boldsymbol{s}$: contextualised vectors of sentences that are the same in Figure 2.

from a textbook used during the training of judges [11] and from five judgment documents not included in the training and test data, and the **law names** feature, which distinguishes 494 specific law names as features and adds a binary feature indicating the presence of *any* law name in the sentence. A document is then input into the CRF model as a sequence of sentences, where each sentence is represented by the features above.

4. BiLSTM-CRF based model

We tested our BiLSTM-CRF based sentence sequence labelling architecture presented here against the baseline model.

BiLSTM-CRF architectures have recently become the standard deep learning method for modelling sequences is the Bidirectional-LSTM(BiLSTM) [12], which can encode both preceding and succeeding context; they have been used for Named Entity Recognition (NER) and POS-tagging [13]. Context in the form of surrounding text as well as surrounding labels can be taken into account with this architecture: past and future input features can be modelled through BiLSTM layers, whereas sequences of labels can be modelled through a CRF layer. Variants of BiLSTM-CRF differ in how to encode the token vectors which is the input to a sequence level BiLSTM layer: [14] uses a Convolutional neural network (CNN)-based character-level representation in addition to word embeddings, whereas [15] uses a character-level representation which is encoded by another BiLSTM encoder. Our BiLSTM-CRF model has three main components, a sentence encoder layer, a BiLSTM-sentence layer, and a CRF layer (Figure 2).

Sentence encoder layer Our target units are sentences, not words as in the POS-tagging and NER task, so they need to be encoded into vectors before passing them to the BiLSTM layer. The sentence encoder layer consists of two components, LSTM-word and CNN-char. LSTM-word takes word embeddings of sentences as input and outputs

the summarised vector for each sentence. CNN-char is a simple CNN with one layer of convolution [16], which takes character embeddings of sentences as input and generates the summarised vector for each sentence. In pre-experiments, using both LSTM-word and CNN-char showed improved performance over using either of these on their own. While LSTM-word should encode the overall meaning of input sentences, CNN-word should capture the characteristic combinations of characters such as typical combinations of Chinese characters and Hiragana characters. Outputs from LSTM-word and CNN-char are concatenated.

BiLSTM-sentence layer / CRF layer We use the architecture proposed in [13]. The BiLSTM-sentence layer takes a sequence of sentences as input and concatenates the hidden state vectors from two LSTMs run bidirectionally; the output of this step should correspond to a contextualised representation of the input sentence vector, which is then input to a CRF layer that computes the final output.

Dropout For regularisation, we include dropout [17] after the LSTM-word and the BiLSTM-sentence layer.

4.1. Input data and embeddings

The inputs to the sentence encoder layer are vector representation of words and characters. As for word inputs, we use the SentencePiece algorithm [18] to tokenise a sentence into tokens, a step necessitated by the fact that the Japanese script does not use an explicit word separator. SentencePiece is an unsupervised text tokeniser which allows us to tokenise without any pre-defined dictionaries. We trained the tokeniser on 15 thousands Civil and Criminal law judgment documents that are published during 1989—2017, using the same web source as our test and training corpus, but excluding the documents used in it (note that the domains differ slightly as our test and training corpus consists only of Civil Law cases). The tokenised words are then input into the embedding layer.

As for character inputs, we simply split a sentence into characters and input them to the embedding layer. The meaning-bearing part of most open-class Japanese words is due to one or more Chinese characters, which are semi-compositionally combined. The characters themselves might therefore contribute additional meaning components and similarities between words beyond the word identities themselves. Each embedding layer converts the input to embedding vectors, which form the input to the sentence encoder. We initialise the embedding layer for characters with GloVe [19] vectors pre-trained with judgment documents of Civil law cases published in the last 14 years (2004–2017).[3]

4.2. Input and output handling

An input to the model is a sequence of sentences. We restrict the length of the sequence to an odd number w.[4] We obtain a sequence of inputs by sliding the size w window from the beginning to the end of the document, sentence by sentence. The n-th input from document D can be represented as $Q_D^n(w) = \{S_D^{n-(w-1)/2}, ..., S_D^n, ..., S_D^{n+(w-1)/2}\}$, where S_D^i is the i-th sentence in document D. At the beginning and the end of the document,

[3] We found in pre-experiments that halving the corpus we used for the tokenisation experiment (disregarding the older half) lead to better results. The target embedding vector dimension is set to 300.

[4] Preliminary experiments where an entire document was input as a single sequence showed low results. The average length of documents was 403.1 lines, which proved too long even for LSTMs with their ability to store a good amount of long-term context.

we fill padding tokens if necessary. We have w predictions in an output for each sentence according to its relative position in the input. We use the prediction that is located in the middle of output.[5]

5. BiLSTM-CRF based model with Headings

We next present a new model which uses the information contained in the documents' headings. Exploiting explicit structural information from the text, such as headings, could model the reading strategy of legal professionals. In particular, we hypothesise that when a human reader notices a new heading in a document, they might interpret this as a signal of rhetorical status change.

In addition to the components of the BiLSTM-CRF model, a dedicated network for handling heading information is added to the model (Figure 3). The network consists of three parts, heading encoder, BiLSTM-heading, heading-sentence concatenator. The heading encoder is a character-based LSTM encoder which summarises the input character embeddings of a heading and outputs a heading vector. The Heading BiLSTM is similar to the BiLSTM-sentence layer, which generates a contextualised representation of headings per input, but is activated only for headings. It does so by inputting the sentence itself to the the upper network layers; otherwise, a special character "_" is used, which signals the non-existence of a heading. The outputs from the BiLSTM-sentence layer and the heading BiLSTM are concatenated and input to a fully-connected layer. The CRF layer then receives the output from the fully-connected layer.

As headings are not explicitly annotated in our corpus, we detect them automatically using a binary rule-based heading detector based on the presence of sentence-final punctuation and sentence length. The detector's performance of finding headings was $F = 0.89$ ($R = 0.99$, $P = 0.81$), measured on all 2,061 lines in 5 random documents (manually annotated by the first author). 622 lines were headings (lines which only contain headings and nothing else) and 1,439 non-headings (either normal sentences, or lines which erroneously contain both a heading and the beginning of a normal sentence).

6. Experiment

6.1. Experimental setting

We use 110 documents from our corpus for training and testing of our two BiLSTM models described in section 4 and 5. Hyperparameters of BiLSTM-CRF models are empirically tuned using the development data (10 documents). The hyperparameters we use for the experiment are shown in Table 3. We use five-fold cross-validation at the document level.

In order to make sure that any performance improvement over our previous work is not only due to the use of heading information per se but to the architecture, we also make the heading information available to the CRF, in the form of a binary feature expressing heading existence, a variant we call [6]. This means that we report results for a total of four models (**CRF, CRF+H, BiLSTM, BiLSTM+H**). We test significance of macro-

[5] Due to a quirk in the experiments, we only pad at the beginning of documents, not at the end. This leads to some cases in each document where the predicted item is not in the middle of the outputs. In those cases, we use the last prediction of the output.

[6] CRF+H also gets the strings of the headings through bigram feature.

Table 3. Hyperparameters for BiLSTM-CRF models

Hyperparameters	values	Hyperparameters	values	Hyperparameters	values
epochs	1	CNN-char channels	256	heading encoder*	64
word emb dim	300	LSTM-word dropout	0.2	BiLSTM-heading*	64 + 64
char emb dim	300	BiLSTM-sent	128 + 128	final cocat*	128
LSTM-word	64	BiLSTM-sent dropout	0.2	∗ If applicable	
CNN-char window	5	heading emb dim*	64		

averaged F measure using a Monte Carlo paired permutation test randomisation at the sentence level with R=100,000 samples at a significance level of $\alpha = 0.05$ (two-tailed).

6.2. Results

Overall results are shown in Table 4. BiLSTM-CRF+H (F=0.654 with setting $w = 11$) significantly outperforms both CRF (F=0.630) and CRF+H (F=0.632), showing that the Deep Learning architecture with heading information indeed represents an overall improvement. This effect holds also without heading information: BiLSTM-CRF ($w = 21$) (F=0.651) is significantly better than CRF (F=0.630) and CRF+H (F=0.632). The BiLSTM-CRF model family overall outperforms the CRF model family.

Although the macro-averaged F performance difference between BiLSTM and BiLSTM-CRF+H is not significant, several individual categories show significant improvement when heading information is added (see Table 5), namely BACKGROUND (F=0.341), FRAMING-main (F=0.651) and CONCLUSION (F=0.449). These are three of the four categories we care about most, as they carry most information for the legal argumentation and form a basis of our further planned processing in this application.

However, this success is paid for with a significant decrease (from F=0.527 to 0.474) for the FRAMING-sub category. These results notwithstanding, we still promote the heading-enabled BiLSTM as our preferred model, as the three improved categories also include the previously weakest of those 4 categories (BACKGROUND increased from F=0.319 to 0.341). With a roughly equal performance in CONCLUSION and FRAMING-sub of both over F=0.45, this leaves us overall in a better situation than without the heading information.

The confusability between FRAMING-main and FRAMING-sub should be one of the main reasons for the remaining errors. Table 6 shows the confusion matrix of the BiLSTM-CRF+H. 1,990 out of 4,727 FRAMING-sub sentences (42.0%) are wrongly classified as FRAMING-main. According to the agreement study of RSC annotation scheme from the previous study [3], the distinction between those two categories is hard even for human annotators. The problem is that the categories both appear in similar locations and have similar surface characteristics e.g. “therefore” phrase in Japanese.

7. Related Work

Rhetorical Status Classification is a commonly used approach in legal text processing for associating text pieces with their rhetorical status. Our rhetorical annotation scheme of six categories plus the OTHER category is an adaptation of previous schemes for the UK law system [4] and Indian law system [5]. For the automatisation of RSC, CRF and other machine learning models have been employed. For RSC of the UK law system, Hachey and Grover used various supervised machine learning systems, achieving the best

Table 4. Macro-averaged results for models

Models	Precision	Recall	F
CRF	0.681	0.603	0.630
CRF + Heading	0.685	0.605	0.632
BiLSTM-CRF ($w = 11$)	0.663	0.635	0.647
BiLSTM-CRF ($w = 21$)	**0.686**	0.629	0.651
BiLSTM-CRF ($w = 31$)	0.673	0.615	0.638
BiLSTM-CRF + Heading ($w = 11$)	0.679	**0.636**	**0.654**
BiLSTM-CRF + Heading ($w = 21$)	0.657	0.628	0.640
BiLSTM-CRF + Heading ($w = 31$)	0.653	0.620	0.633

Table 5. Results of models by classes (F)

Category	CRF	BiLSTM-CRF	BiLSTM-CRF+H
BACKGROUND	**0.344**	0.319	0.341
CONCLUSION	0.381	0.415	**0.449**
FACT	0.853	**0.890**	0.879
FRAMING-main	0.594	0.642	**0.651**
FRAMING-sub	0.471	**0.527**	0.474
IDENTIFYING	0.792	0.798	**0.806**
OTHER	0.972	0.969	**0.975**

BiLSTM-CRF is $w = 21$ and BiLSTM-CRF+H is $w = 11$.

Table 6. Confusion Matrix of BiLSTM-CRF + Heading ($w = 11$)

		Prediction							
		BGD	CCL	FCT	FRm	FRs	IDT	OTR	Total
Gold	BGD	38	0	19	38	29	0	3	127
	CCL	0	699	42	847	28	6	90	1,712
	FCT	3	36	9,544	500	181	15	235	10,514
	FRm	18	548	745	5,836	1,214	44	132	8,537
	FRs	33	31	628	1,990	1,944	49	52	4,727
	IDT	2	15	30	96	53	710	24	930
	OTR	2	73	191	76	18	7	17,763	18,130
	Total	96	1,402	11,199	9,383	3,467	831	18,299	44,977

results with C4.5 [20] with only the location feature (F=0.65); the second-best (F=0.61) was achieved using a Support Vector Machine [21] with all features (location, thematic words, sentence length, quotation, entities and cue phrases). As for RSC of Indian law system, a CRF classifier with various features similar to our CRF model achieved F=0.82.

Walker et al develop a rule-based RSC classifier from a small amount of labelled data [6]. Their task is to identify rhetorical roles of sentences such as "Finding", which states whether a propositional condition of a legal rule is determined to be true, false or undecided, "Evidence", such as the testimony of a lay witness or a medical record, "Reasoning" which reports reasoning parts underlying the findings of fact (i.e. a premise), "Legal-Rule" which states legal rules, and "Citation" which references legal authorities or other law materials, and "Others". There are close similarities to our categories. 530 sentences were used to develop a rule set set for their classifier, and the paper reports the comparison between their low-cost rule-based classifier (F=0.52).

Some F-measures from previous studies are higher than ours, but this mirrors the difficulty of our task. None of the other schemes makes such fine distinctions as we do, particularly in the lower levels of argumentative support such as those expressed by the FRAMING-main vs FRAMING-sub distinction.

Another piece of work performs *deontic* sentence classification in contract documents [22], using a hierarchical RNN-based architecture. The sentences are classified into "Obligation", "Prohibition", "Obligation List Intro", "Obligation List Item", and "Prohibition List Item". The model is based on a BiLSTM-based sequential sentence classifier which considers both the sequence of words in each sentence and the sequence of sentences like our models, but it does not employ a label sequence optimiser such as our CRF layer.

Outside the legal document processing community, RSC is often used in the area of scientific paper processing for the extraction of relevant material and for summarisation. An RNN-based model similar to ours has been proposed for the RSC of sentences in medical scientific abstracts [23]. Our model shares the basic design (a sentence encoder, a context encoder, and a CRF layer) with this model; however, their model does not consider heading information.

8. Conclusion

In this paper, we proposed to apply a BiLSTM-CRF based model for rhetorical status classification. It performs RSC with sequential labelling by taking inter-sentence level context into account. We also proposed to add a dedicated network which conveys contextualised heading information, after headings have been recognised by a simple automatic heading detector. The model showed significant improvements from the plain BiLSTM-CRF model in BACKGROUND, FRAMING-main and CONCLUSION. We also extended the size of our annotated corpus of Japanese judgment documents. The resulting system showed a significant improvement from our CRF based baseline models.

There are several possible directions for future work. One of these is to train our model with curriculum learning strategy [24]. Curriculum learning is a training approach that exposes a model by giving training examples in a meaningful order, gradually increasing difficulty. RSC seems to fit this training scheme very well, as it shows various patterns of sequences from simple ones such as category repetitions ("FACT, FACT, FACT ...") to more complicated ones such as "FRAMING-sub, FRAMING-sub, FRAMING-main, FRAMING-sub, BACKGROUND ...". Curriculum learning might therefore help our model to learn how to distinguish difficult categories (e.g. FRAMING-sub v.s. FRAMING-main) in an efficient way. Also, we plan to conduct an extrinsic evaluation with a summarisation task by lawyers, which uses the results of the RSC.

Acknowledgments. This work was supported by Tokyo Tech World Research Hub Initiative (WRHI) Program of Institute of Innovative Research, Tokyo Institute of Technology.

References

[1] Ministry of Justice, Japan, "Form of Rendition", Code of Civil Procedure, Article 252.
[2] Ministry of Justice, Japan, "Judgment Document", Code of Civil Procedure, Article 253.
[3] H. Yamada, S. Teufel and T. Tokunaga, Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation, *Artificial Intelligence and Law* **27**(2) (2019), 141–170.

[4] B. Hachey and C. Grover, Extractive summarisation of legal texts, *Artificial Intelligence and Law* **14**(4) (2006), 305–345.
[5] M. Saravanan and B. Ravindran, Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment, *Artificial Intelligence and Law* **18**(1) (2010), 45–76.
[6] V.R. Walker, K. Pillaipakkamnatt, A.M. Davidson, M. Linares and D.J. Pesce, Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning (2019).
[7] J.D. Lafferty, A. McCallum and F.C.N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. ISBN ISBN 1-55860-778-1.
[8] N. Okazaki, CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007.
[9] T. Masuoka, *Nihongo Modariti Tankyu (Japanese Modality Investigations)*, Kuroshio shuppan, 2007.
[10] S. Matsuyoshi, S. Sati and T. Utsuro, A Dictionary of Japanese Functional Expressions with Hierarchical Organization, *Journal of Natural Language Processing* **14**(5) (2007), 123–146.
[11] Judicial Research and Training Institute of Japan, *The guide to write civil judgements (in Japanese)*, 10th edn, Housou-kai, 2006.
[12] A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural networks* **18**(5–6) (2005), 602–610.
[13] Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, *CoRR* **abs/1508.01991** (2015).
[14] X. Ma and E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2016, pp. 1064–1074.
[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural Architectures for Named Entity Recognition, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270.
[16] Y. Kim, Convolutional Neural Networks for Sentence Classification, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751.
[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research* **15**(1) (2014), 1929–1958.
[18] T. Kudo and J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71.
[19] J. Pennington, R. Socher and C. Manning, Glove: Global Vectors for Word Representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543.
[20] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN ISBN 1-55860-238-0.
[21] C. Cortes and V. Vapnik, Support-vector networks, *Machine learning* **20**(3) (1995), 273–297.
[22] I. Chalkidis, I. Androutsopoulos and A. Michos, Obligation and Prohibition Extraction Using Hierarchical RNNs, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 254–259.
[23] D. Jin and P. Szolovits, Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3100–3109.
[24] Y. Bengio, J. Louradour, R. Collobert and J. Weston, Curriculum Learning, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, ACM, New York, NY, USA, 2009, pp. 41–48. ISBN ISBN 978-1-60558-516-1.

Short Papers

This page intentionally left blank

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190316

Privacy and Monopoly Concerns in Data-Driven Transactions

Duygu AKŞİT KARAÇAM[1]
Catholic University of Leuven (KU Leuven) Faculty of Law (LL.M) Alumna
Attorney at Law registered at Istanbul Bar Association

Abstract. The increase of data-driven M&A transactions have raised apprehension over potential violations of data privacy rights. The economic significance attributed to Big Data has also called into question whether data privacy could be a parameter in merger control proceedings. Our purpose is to address the privacy and monopoly concerns arising from data-driven transactions within the scope of both the EC Regulation and the purpose limitation principle under the GDPR.

Keywords. Big Data Analysis, GDPR, EC Merger Regulation, Artificial Intelligence, Targeted Advertising, Behavioral Marketing, Merger Control

1. Introduction

Consider a subscriber to an online video-on-demand ("**VOD**") service provider, who enjoys watching crime documentaries. The artificial intelligence ("**AI**") operated by this platform may suggest entertainment content related to legal enforcement, which in our instance could be *Elementary* to this individual. To make the content more interesting, the algorithm would need to display a more tailored banner, based on this individual's preferences. If he or she likes dramas but never pays attention to action shows, the banner as recommended by the algorithm may contain a fragment of the relationship between the characters, instead of violent scenes. In this assumption, the AI renders an automated decision based on the data retrieved from the online VOD service provider and uses it to the benefit of the same. In this study our focus is on what would happen if a third party such as a search engine operator, merges with this online VOD service provider and, eventually, transfers search profile of this individual to the latter? For example, an online content provider can create a library of product types as well as mentions of specific product types by applying machine learning to the scripts of the shows. It is then possible to correlate this library with an individual's viewing habits through data mining where for example shows that contain specific products are viewed more often than others. With this knowledge it is then possible to recommend the individual other content that features similar products as well as to bring commercials with such products to the individual while they are watching the actual content. While this kind of targeted marketing may be less intimidating for some; what happens if, an individual is not granted a loan by the banks due to his/her interest in mortgage crisis documentaries? This may seem a bit exaggerated but it does demonstrate the prolific outcome that may be achieved through the processing of consolidated data. The question

[1] E-mail: duyguaksit@gmail.com

is then, where should the lawmaker draw the line in data-driven mergers and acquisitions ("**Concentrations**")?

2. The Purpose Limitation Principle in a Nut Shell

Processing of personal data is prohibited unless one of the legitimate grounds in Article 6 of the General Data Protection Regulation [1] ("**GDPR**") is applicable, including but not limited to, consent of data subjects, contractual necessity, or, legitimate interests pursued by the data controller [2]. Article 5(1)(b) of the GDPR states that personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. Articles 13 and 14 of the GDPR lays out notification duties for data controllers with regards to processing of personal data which was either collected from the data subject or third parties [3]. As per Articles 13(2)f and 14(2)g of the GDPR, data subjects must be provided with information on the existence of automated decision-making if any; and, meaningful information regarding the logic involved, as well as the significance and the envisaged consequences of such processing. This provision should be interpreted flexible enough to enable the data subjects exercise their rights under the GDPR [4].

Remarkably, as per Article 13(3) and 14(4) of the GDPR, where the controller intends to process the personal data for a purpose other than for which the personal data were collected, the controller shall provide the data subject prior to that further processing, with information on that other purpose. This is of great relevance to data-driven Concentrations, since the purposes for the personal data processing may change or cease to exist after the transaction. In that case, the Concentration should re-fulfil the notification duties in accordance with the GDPR; furthermore, may even be required to obtain consent from the data subjects, in the absence of legitimate grounds for data processing post-transaction. This should be evaluated on a case-by-case basis. It may also be that the transaction has no effect on the data processing activities concerned; or, the purposes for the processing of personal data may not have changed at all post-transaction. However, one should always remember that, pursuant to Article 5(2) of the GDPR, data controllers shall be responsible for, and be able to demonstrate compliance with the purpose limitation principle.

3. Big Data[2]: A Source of Monopoly Power?

Big Data is a powerful form of data mining that relies on huge volumes of data, faster computers, and new analytic techniques to discover hidden correlations [5]. Big Data analysis promises a much higher success rate than traditional selection methods as the former may be used to acquire customers, analyze competitors/pricing, or, optimize distribution, marketing and branding [6].

Big Data has especially gained significance for companies with two-sided business models that simultaneously serve different groups of consumers. The most common example would be social networking websites, where users pay no subscription fees, or the like; but, in exchange, either willingly or unconsciously, make their personal data available to the service providers. For instance, in Facebook's case, whereas one

[2] We refer to Big Data, which contains personal data as defined under Article 4(1) of the GDPR.

customer group is offered social networking services free of charge; the second customer group pays monetary remuneration in anticipation of attracting the first group's attention with advertisements. Some argue that the first group is also paying for Facebook's services with their personal data and the time they spend on Facebook which exposes them to the advertisements of the other group [7]. This "dual dignitary economic role for personal data" have led scholars and regulators to examine what role data protection should play in the assessment of merger control [8].

The EC has held in Google Search (Shopping) case that, the fact that a product or service is provided free of charge does not prevent the goods or services from constituting an economic activity within the meaning of the EU competition legislation [9]. This decision was considered to be an evolutionary recognition of data as new currency in digital markets by some scholars [10]. On the other hand, some commentators argue that, users utilize multiple online services even for the same type of task (multi-homing); and, collection of data by one provider does not detract others from collecting and using the same data [11]. It is also indicated that personal data cannot function like currency given that its value to consumers or businesses varies based on a number of factors; whereas, actual currency derives its usefulness from its common value agreed by all parties to the transaction [12]. From our perspective, while Big Data may not literally be the new currency, the competitive advantage gained therefrom is so crucial that one may be able to discern market trends well before the other players in the market [13].

4. Data-Related Concerns in Merger Control

By virtue of Articles 2(3) and 8(3) of the Council Regulation on the Control of Concentrations between Undertakings [14] ("**EC Merger Regulation**"), the EC has the authority to declare a proposed Concentration incompatible with the EU common market, if it finds that the Concentration would significantly impede effective competition as a result of creation or strengthening of a dominant position [15].

The EC has held on several occasions that, privacy-related concerns arising from the use of data under the Concentration's control, falls within the scope of the EU data protection rules, and not the EU competition law rules [16, 17]. The Court of Justice of the European Union ("**CJEU**") has also ruled in Asnef-Equifax decision that personal data-related issues are not a matter for competition law and may be resolved by data protection laws [18]. In this context, the *Monopolkommission* stressed that: "The importance of data for the commercial success of companies should be taken into account more prominently in competition proceedings. This is particularly important in merger control proceedings – frequently relatively new internet service providers, characterized by low turnover, but potentially highly valuable data inventories, are acquisition targets. In contrast, aspects entirely related to data protection should be addressed outside competition law proceedings [19]." Without prejudice to the foregoing, Big Data may be a source of monopoly power, if the data-driven transaction leads to the exclusion of competition, since the merged entity may refuse third parties from accessing the data or provide differentiated access to the data. This may create a barrier to market entry, also strengthening the market power of the Concentration [8]. However, the possession of Big Data *per se* does not necessarily raise competition concerns.

In Google/DoubleClick decision, the EC evaluated whether the combination of DoubleClick's assets with Google, in particular the consumer-provided-data generated

by internet use, would allow the merged entity to achieve a position in the market that could not be replicated by its competitors. It was noted that, even if Google and DoubleClick's data collections were available as an input for DoubleClick, it was unlikely that this competitiveness would confer on the merged entity a competitive advantage that could not be matched by its competitors. In fact, the combination of data on web surfing behavior is already available to Google's many competitors and the competitors could also purchase data or targeting services from third parties whose commercial activities cover third party cookies, deep packet inspection *etc*. [20]. The EC concluded that the combination of two datasets was unlikely to bring more traffic to AdSense and squeeze out competitors, which would eventually enable the Concentration to charge higher prices for its intermediation services [20].

In Facebook/WhatsApp decision, the EC investigated whether Facebook's acquisition of WhatsApp would materially strengthen Facebook's position in the online advertising services market, as a result of the increased amount of data to be controlled by Facebook post-transaction [17]. First off, the EC has not conducted an investigation with specific focus on data or data analysis services, due to its finding that neither of the parties were active in said markets at the time [17]. The EC considered in its assessment that there are so many alternative market participants that collect user data alongside Facebook with 6.39% estimated share of data; including Google (33%) which accounts for the signification portion of internet user data, or Adobe (1.27%), Yahoo! (0,65%), Microsoft (0.02%), and others (58.67%). Accordingly, there would remain a large amount of internet user data valuable for advertising purposes, which are not within Facebook's exclusive control [17]. The EC indicated that the transaction would raise competition concerns, only if the combined datasets were to allow Facebook to strengthen its position in advertising [17]. At this point, it is worth noting that the EC was criticized in a Working Paper, for not having assessed whether the combination of datasets would entail consumer harm in the consumer communication markets [21].

The role of the GDPR in data-driven Concentrations was officially recognized in the Microsoft/LinkedIn decision where the EC concluded that data combination could be implemented by the merged entity only to the extent allowed by applicable data protection laws, such as the GDPR or national laws [22]. The EC explained that, provided that the combination of two datasets is legitimate under applicable data protection laws, there are two main ways in which a merger could raise horizontal issues due to the combination of datasets which were previously in the possession of two independent companies. Either (a) data combination may increase the merged entity's market power in a hypothetical market for the supply of this data, or increase barriers to entry or expansion in the market for competitors in need of such data to operate in the same market; or, (b) even if there is no intention or technical possibility of combining two datasets, it may be that before the merger, the two companies were competitors on the basis of the data they controlled and that this competition is eliminated by the merger in question [22]. The EC further emphasized other factors such as (a) parties do not make available their data to third parties for advertising purposes, (b) there are other large amounts of data valuable in terms of advertising purposes which are not within Microsoft's exclusive control, and (c) parties of the dispute are small market players and compete with each other only to a limited extent in online advertising market. In light of this, the EC held that the transaction did not raise serious doubts in its compatibility with the internal market in the online advertising market [22]. However, with respect to the customer relationship management (CRM) software solutions market, the EC stated that machine learning (ML) and CRM software solutions require access to multiple data

sources in order to provide useful insights and that LinkedIn is only one data source among many others [22]. Hence, data-related competition concerns have different aspects to be considered in each market, which requires case by case analysis for each subsector concerned.

Last but not least, the approval of a transaction within the scope of the EC Merger Regulation does not prevent national data protection authorities from conducting their own investigations in parallel with or after the merger control investigation by the EC [23]. As per Article 21(4) of the EC Merger Regulation, Member States of the EU may take appropriate measures against Concentrations to protect legitimate interests other than those taken into consideration by the EC Merger Regulation. The legitimate interests specified in under the same Article are (a) public security, (b) the plurality of the media and (c) prudential rules. This is not an exclusive list as Member States are also entitled to communicate other public interests to the EC, who is to assess whether the proposed legitimate interest is compatible with the general principles and other principles of Community Law. As per the same Article, if the EC acknowledges grounds related to data protection as a legitimate basis to adopt measures against a Concentration, the Member State concerned may, under its national laws, subject the Concentration to additional conditions or block the Concentration altogether if prohibiting the transaction is proportionate for the purposes of protection of public interest [23]. As of yet, no Concentration has been blocked on the basis of this ground, thus, only time will tell whether non-compliance with the GDPR will be considered as a legitimate interest to take measures against a Concentration under the EC Merger Regulation.

5. Conclusion

Although the CJEU and the EC has previously held that the issues relating to personal data are not to be dealt within the scope of competition law, this is not to be interpreted as the EC is prohibited from scrutinizing data-related competition concerns in merger control proceedings. While the possession of large sets of personal data *per se* does not necessarily raise competition concerns, the EC is entitled to block a data-driven Concentration, if it finds that the transaction would significantly impede effective competition by means of creation and/or strengthening of a dominant position. Regardless of whether a Concentration is cleared or not by the competition authorities, the data controllers may be required to fulfill the notification duties set forth under the GDPR once more after the transaction, if the purposes which rendered the processing activity lawful in the first place have changed or ceased to exist. In fact, the Concentration may even be required to obtain consent from the data subjects, if no other legitimate grounds set forth under the GDPR are applicable post-transaction. This is the current state of art, which may be challenged through Article 21(4) of the EC Merger Regulation, or a more consumer-welfare focused approach in merger control proceedings.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119/1, 04.05.2016.

[2] M. Botta and K. Widememann. EU Competition Law Enforcement Vis-À-Vis Exploitative Conducts in the Data Economy Exploring the Terro Incognita. Max Planck Institute for Innovation & Competition Research Paper No. 18-08, 2018
[3] S. Wachter, B. Mittelstadt and L. Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7 (2017), 76-99..
[4] A. D. Selbst and J. Powles. Meaningful information and the right to explanation. *International Data Privacy Law* 7(2017), 233–242.
[5] I. Rubinstein. Big Data: The End of Privacy or a New Beginning. NYU School of Law, Public Law Research Paper No. 12-26, 2013. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2157659 Date of access: 03.11.2019.
[6] P. Vanini. Asset Management, 2019. Available at: https://ssrn.com/abstract=2710123. Date of access: 03.11.2019.
[7] T. Körber. Is Knowledge (Market) Power? - On the Relationship Between Data Protection, 'Data Power' and Competition Law, 2018. Available at https://ssrn.com/abstract=3112232, Date of access: 03.11.2019.
[8] O. Lynskey. DAF/COMP/WD(2018)70, Directorate for Financial and Enterprise Affairs Competition Committee, Non-Price Effects of Mergers, 2018. Available at: https://one.oecd.org/document/DAF/COMP/WD(2018)70/en/pdf Date of access: 03.11.2019.
[9] European Commission, Case No. AT.39740, Google Search (Shopping) decision, 27.06.2017.
[10] A. D. Chirita. Data-Driven Mergers Under EU Competition Law, 2018. Available at: https://ssrn.com/abstract=3199912 Date of access: 03.11.2019.
[11] A. V. Lerner. The Role of Big Data in Online Platform Competition, 2014. Available at: https://ssrn.com/abstract=2482780 Date of Access: 03.11.2019.
[12] D. A. Balto and M. C. Lane. Monopolizing Water in a Tsunamı: Finding Sensible Antitrust Rules for Big Data, 2016. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2753249 Date of Access: 21.10.2019.
[13] M. E. Stucke and A. P. Grunes. Debunking the Myths Over Big Data and Antitrust, CPI Antitrust Chronicle, 2015. University of Tennessee Legal Studies Research Paper No. 276. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2612562 Date of access: 03.11.2019.
[14] Council Regulation (EC) No 139/2004 of 20 January 2004 on the Control of Concentrations between Undertakings, OJ 024, 29.01.2004 P. 0001 – 0022.
[15] I. Graef. Blurring Boundaries of Consumer Welfare: How to Create Synergies between Competition, Consumer and Data Protection Law in Digital Markets, 2016. Max Planck Institute Post-Doc conference on Personal Data in Competition, Consumer Protection and IP Law: Towards a Holistic Approach. Available at: https://ssrn.com/abstract=2881969 Date of access: 03.11.2019.
[16] European Commission, Case No. M.7813, Sanofi/ Google/ DMI JV decision, 23.02.2016.
[17] European Commission, Case No. COMP/M.7217, Facebook/Whatsapp decision, 03.10.2014.
[18] Case C-238/05, Asnef Equifax v. Ausbank, ECLI:EU:C:2006:734.
[19] The Monopolies Commission, Special Report 68 on Competition Policy: The Challenge of Digital Markets, 2015. Available at https://www.monopolkommission.de/images/PDF/SG/SG68/S68_summary.pdf Date of access: 03.11.2019.
[20] European Commission, Case No. COMP/M.4731, Google/DoubleClick decision, 11.03.2008.
[21] E. Deutscher. How to Measure Privacy-Related Consumer Harm in Merger Analysis? A Critical Reassessment of the EU Commission's Merger Control in Data-Driven Markets, 2018. EUI Working Paper LAW 2018/13. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3075200 Date of access: 03.11.2019.
[22] European Commission, Case No. M. 8124, Microsoft/LinkedIn decision, 06.12.2016.
[23] I. Graef, D. Clifford and P. Valcke. Fairness and Enforcement: Bridging Competition, Data Protection and Consumer Law. International Data Privacy Law 8 (2018) 200-223.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190317

Realising ANGELIC Designs Using Logiak

Katie ATKINSON [a], Trevor BENCH-CAPON [a], Tom ROUTEN [b,1], Alejandro SÁNCHEZ [b], Stuart WHITTLE [c], Rob WILLIAMS [c] and Catriona WOLFENDEN [c]

[a] *Department of Computer Science, The University of Liverpool, UK*
[b] *Things Prime GmbH, Basel, Switzerland.*
[c] *Weightmans LLP, Liverpool, UK*

Abstract. ANGELIC is a methodology for encapsulating knowledge of a body of case law. Logiak is a system intended to support the development of logic programs by domain experts, and provides an excellent environment for the rapid realisation of ANGELIC designs. We report our use of Logiak to realise ANGELIC designs, using both Boolean factors and factors with magnitude.

1. Introduction

The ANGELIC methodology for designing systems intended to encapsulate case law was described in [1]. In partnership with the UK law firm, Weightmans, this methodology was used to build a substantial application designed to support decisions as to whether or not claims for compensation for Noise Induced Hearing Loss (NIHL) should be contested [2]. Realising the design required considerable effort in [2] and a custom built interface had to be produced from scratch. To address these problems we explored the use of a target implementation platform, Logiak, to enable rapid and convenient realisation of the design, and to supply a user interface as part of the package.

Logiak has been used to implement two ANGELIC designs. First we re-implemented the design for NIHL of [2]. Then as part of the exploration of extending ANGELIC to handle factors with magnitude we re-implemented the design described in [3] which added magnitudes to the well known US Trade Secrets program CATO [4]. A longer version of this paper [5] which supplies additional details can be found at https://cgi.csc.liv.ac.uk/research/techreports/.

2. ANGELIC and Logiak

The ANGELIC methodology described in [1] is designed to encapsulate case law knowledge to be used in factor based reasoning systems [6]. The methodology is

[1]Corresponding Author: Tom Routen; Email: routen@mangologic.com

based on traditional knowledge elicitation techniques, drawing the information from a variety documents under the guidance of a domain expert. The knowledge is presented in a form similar to the abstract factor hierarchy of CATO [4]. However, each node also has a set of *acceptance conditions* which state precisely how children relate to their parent node and enable the structure to be interpreted as an Abstract Dialectical Framework (ADF) [7]. Thus the design document provides both the advantages of a hierarchical structure, and a fine grained, domain relevant, partitioning of the knowledge base, while having the formal properties of the ADF. A fuller description of the stages of the methodology can be found in [2], and several examples of design structures in [1]. Some adaptations were made to provide what was needed for Logiak, most notably the association of questions with each of the base level factors to support the provision of an interface. More changes were made to enable ANGELIC to accommodate factors with magnitude. The most significant of these was the use of a limited number of design patterns as the acceptance conditions.

Logiak, produced by Things Prime GmbH, is a system with two main aspects:

- Firstly, it is a "no code" environment within which it is possible to create systems, including mobile systems, by configuration only.
- Secondly, as its name suggests, it is a system concerned to facilitate the representation of complex decision logic.

The design of Logiak has been influenced by its deployment in projects which use mobile technology to support often poorly-trained health workers in under-resourced settings to nevertheless follow best practice in diagnostic and treatment logic. Users include *Médicines sans Frontières*. Diagnosis and treatment logic can be very complex logic indeed and, while the WHO's Digital Health Guideline (www.who.int/reproductivehealth/publications/digital-interventions-health-system-strengthening/en/) affirms the use of decision support software to improve the quality of care, it remarks on "the importance of ensuring the validity of the underlying information, such as the algorithms and decision-logic".

Logiak permits explicit representation of both procedural and declarative logic and clearly separates the two. In Logiak, one defines "processes" in two parts: "nodes" and "conditions". The "nodes" are sequential and each represents either an interaction with the user (e.g. obtaining input) or a background action (e.g. updating a variable). A Process is therefore a sequence of such nodes executed one after the other. However, the execution can be affected by the specification of "preconditions" for nodes or groups of nodes (if a precondition is not true, the node is not executed). Such conditions are defined purely declaratively, either in terms of values of variables or responses from the user. Additionally, and importantly, one can define "meta-conditions" – i.e. conditions can be logical combinations of other conditions.

Experience has shown that the clean separation of the declarative from the procedural means that it is straightforward for domain experts to become fluent in specifying the fundamental logic of a Process.

3. Noise Induced Hearing Loss (NIHL)

Hearing loss induced by noise to which workers are subjected as part of their employment is widespread and it is possible for workers to make claims for compensation against negligent employers. Weightmans act for employers and their insurance companies by advising on whether claims should be settled or contested. The NIHL application was implemented in Logiak as a proof of concept of the compatibility of ANGELIC and Logiak, and to demonstrate the interface produced from Logiak.

Table 1. Selected nodes from NIHL design document

ID	Factor	Children	Conditions	Description
20	Breach of Duty	26 Employee told of Risks 27 Methods to reduce noise 28 Protection zone 29 Health surveillance 30 Risk assessment	REJECT IF Employee told of Risks AND Methods to reduce noise AND Protection zone AND Health Surveillance AND Risk assessment ACCEPT otherwise	The employer did not follow the code of practice in some respect.
28	Protection zone	Base Level	Q6 Yes	Employer provides methods to identify areas where noise level are high

3.1. Design

The design document used for the NIHL application was essentially the same as that produced in [2]. The only difference was that the base level factors were now associated with a question to be posed to the user. A set of questions and possible responses, taken from the check-list document used in the elicitation and the interface designed for [2], were supplied so that the interface can also be generated from the document. The rows for the node *BreachOfDuty* and one of its base level children are shown in Table 1. Question 6, used to give a value to *Protection zone*, was *Did the employer fix protection zones? Yes/No.* This design was then realised using Logiak as described in the next section.

3.2. Realisation

Using Logiak to create a functioning interactive system from the ANGELIC specification of NIHLwas largely a matter of (simply) transcribing the design document elements. The first kind of transcription is to take the questions associated with base level factors and enter them into a Logiak Process, to create a user dialogue (shown as Figure 1 of [5]).

The second (and more interesting) "transcription" relates to the logic: one defines the conditions in Logiak, in a way which closely mirrors the acceptance

conditions defined in the ANGELIC specification. In Logiak, one can define conditions of various types. The simplest are those defined on the basis of user responses to questions, and so correspond directly to ANGELIC "base level factors". For example, for the yes/no question "Did the employer fix protection zones?" we define a condition named "Protection Zone" which is true if and only if the user responded affirmatively to said question (Figure 2 of [5]).

Using these conditions (ANGELIC base level factors), in Logiak we can define "meta conditions" which are logical combinations of other conditions. For example, for the ANGELIC "Breach of Duty" conditions, we defined a Logiak meta-condition "No breach of duty" (shown in Figure 3 of [5]), which is true if all base conditions relating to employer duties are satisfied and false otherwise. We then defined a meta-condition "Breach of Duty" which is true if the "No breach of duty" is false.

This indicates that the only aspect of implementing a system in Logiak based on ANGELIC which is not effectively transcribing the ANGELIC methodology output in a one-to-one manner, is in mapping the accept-reject logic of ANGELIC into declarative logic. ANGELIC makes use of defaults, for example, whereas in Logiak all conditions must be explicit. In practice, this poses little difficulty.

After these two kinds of "transcription" from ANGELIC, one has defined an interactive process in Logiak which can be delivered either on the web or as a mobile app without any further programming. Users can respond to the questions and Logiak will compute the logic dictated by the conditions. Within the Logiak environment, one can interact with a process defined to check and debug the logic as portrayed in Figure 1. The interface that will be seen by end users is shown in the left hand pane. If desired, the question shown to users can be accompanied by explanatory text and pictures.

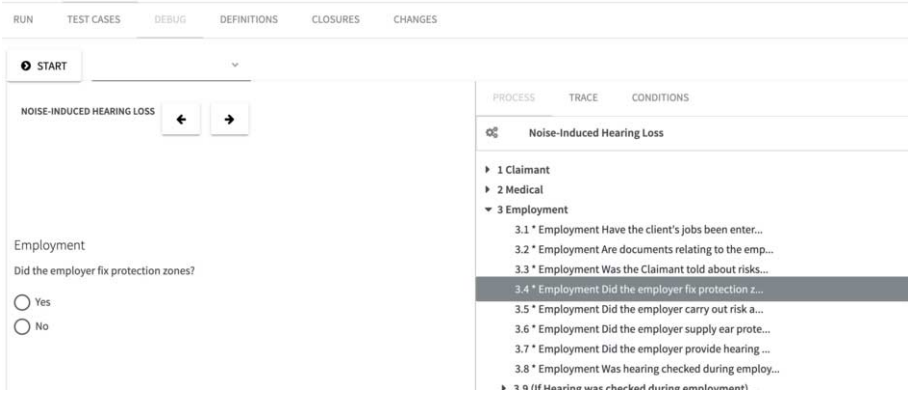


Figure 1. Executing the Process in the debugger

4. CATO with Magnitudes

The CATO application used factors with magnitude as well as Boolean factors. The need for some factors to have magnitude has become widely recognised in AI and Law: in particular the need for magnitudes in CATO was discussed in [3].

Table 2. Abstract Factors in CATO. LM is Legitimate Means and QM is Questionable Means

Parent	Type	Value	Child 1	Child 2	Condition Pattern
Known	Boolean	LM	Limitations	KnownOutside	ThresholdException2
IllegalMethods	Boolean	QM	Criminal	Dubious	Or

Table 3. Base Level Factors in CATO application

Factor	Type	Question	Pattern
AgreedNotToDisclose	Boolean	Did the defendant agree not to disclose?	QueryTheUser1
SecurityMeasures	Magnitude	On a scale of 0-10, how strong were the security measures taken by the plaintiff?	QueryTheUser3 (10)

4.1. Design

To adapt to factors with magnitude, the Boolean design used in [1] was rewritten with some of the base level factors given magnitudes. This involved some changes to the original design, in order that the acceptance conditions could be written so as to accommodate non-Boolean factors. One change was to rewrite the original ADF of [1] as a 2-regular ADF (shown in full in [3]) in which every parent node has exactly two children. This facilitates implementation by making the treatment of nodes more uniform. This design was implemented in Prolog [3], but the code was extremely procedural and rather laborious to construct because a fine grained level of control had to be imposed. What this exercise did achieve, however, was the identification of a limited number of patterns for acceptance conditions. Building on [3], twelve patterns were identified for the current exercise. The twelve patterns were *And*, *Or*, *Weighted Sum*, *Weighted Difference*, 2 kinds of *exception*, 3 uses of *thresholds* and 3 varieties of *Query the User*. For details of the patterns see [5].

Instead of acceptance conditions, each node is now associated with one of these twelve patterns, showing how the parent relates to its children. The base level factors are associated with a question and one of the *Query the User* patterns. Note that the patterns require the specification of *weights* and *thresholds*. These were specified for *values* rather than individual factors, and so each factor needs to be associated with a value. We used the five values identified for CATO in [8]. The weights and thresholds can be set to reflect the relative importance of the values but we used equal weights and thresholds. Effects of varying the weights and thresholds are discussed in [3]. Example nodes for abstract factors are shown in Table 2, and example base level factors are shown in Table 3.

4.2. Realisation

Within Logiak, a Process can contain not only interactions, such as the questions to the user as described above, but also actions. Actions are Process steps which happen in the background without user interaction and can include, for example, the creation and updating of variables. Conditions can be defined on the values of such variables, just as they can be defined on user responses. Actions which update numeric variables can use an expression language and the task of

reflecting ANGELIC's use of factors became effectively the inclusion of variable update actions using expressions which implement the patterns described above, making it a quasi-mechanical process. The implementation in Logiak could be simplified by a direct association of a magnitude with each Condition representing an ANGELIC factor, and implementing the "patterns" above, not as explicitly constructed expressions but as system operators. This would also hide the detailed expressions from the implementer, which would be more in line with the "no code" ethos of the system.

5. Discussion and Concluding Remarks

Both implementations were evaluated against the applications described in [2] and [3]. They were run using the same test data and produced fully correct results. The close structural correspondence between Logiak and ANGELIC greatly facilitated the verification of the implementation against the design. Moreover the discipline imposed by the implementation meant that any imperfections and unclarities could be detected and resolved. The CATO exercise threw up 15, mostly minor, queries, leading to a better design. Moreover, the immediate availability of a user interface meant that end users could be involved in evaluation. Weightmans provided positive feedback on the NIHL application.

The ability to rapidly turn the design into a useable application greatly enhances the development process, by identifying problems at early stage so that the design can be refined, and by enabling end users and domain experts to participate in the process using the interface which is part of the Logiak package. Further, implementation in Logiak means that it is unnecessary to develop a separate user interface, which required a substantial additional effort for NIHL [2]. Providing a straightforward way of implementing ANGELIC designs is an important addition to the methodology, greatly increasing its practical usability.

References

[1] L. Al-Abdulkarim, K. Atkinson and T. Bench-Capon, A methodology for designing systems to reason with legal cases using ADFs, *AI and Law* **24**(1) (2016), 1–49.
[2] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, S. Whittle, R. Williams and C. Wolfenden, Noise induced hearing loss: Building an application using the angelic methodology, *Argument & Computation* **10**(1) (2019), 5–22.
[3] T. Bench-Capon and K. Atkinson, Lessons from Implementing Factors with Magnitude, in: *Proceedings of JURIX 2018*, 2018, pp. 11–20.
[4] V. Aleven, Teaching case-based argumentation through a model and examples, PhD thesis, University of Pittsburgh, 1997.
[5] K. Atkinson, T. Bench-Capon, R. T, A. Sánchez, S. Whittle, R. Williams and C. Wolfenden, Implementing ANGELIC Designs Using Logiak, Technical Report, ULCS-19-002, University of Liverpool, 2019.
[6] T. Bench-Capon, HYPO'S legacy: introduction to the virtual special issue, *Artificial Intelligence and Law* **25**(2) (2017), 205–250.
[7] G. Brewka and S. Woltran, Abstract Dialectical Frameworks, in: *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, 2010.
[8] A. Chorley and T. Bench-Capon, An empirical investigation of reasoning with legal cases through theory construction and application, *AI and Law* **13**(3) (2005), 323–371.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190318

Renvoi in Private International Law: A Formalization with Modal Contexts

Matteo BALDONI [a], Laura GIORDANO [b] and Ken SATOH [c]
[a] *Università di Torino, Dipartimento di Informatica, Italy*
[b] *Università del Piemonte Orientale, DISIT, Italy*
[c] *Principles of Informatics Research Division, NII, Japan*

Abstract. The paper deals with the problem of formalizing the renvoi in private international law. A rule based (first-order) fragment of a multimodal logic including context modalities as well as a (simplified) notion of common knowledge is introduced. It allows context variables to occur within modalities and context names to be used as predicate arguments, providing a simple combination of meta-predicates and modal constructs. The nesting of contexts in queries is exploited in the formalization of the renvoi problem.

1. Introduction

Given an international matter (is Taro a heir of John?), one wants to decide whether the matter is valid in a given country (such as in Japan) or not. In some cases, such as when Taro's parents do not have the same nationality, this matter cannot be answered only considering the legislation of one country, and requires the determination of the jurisdiction of the matter. For instance, if there is a legal child-parent relationship between Taro and John in John's home country, the application of the law in Japan, means the application of the law in force in that country.

Private international law "enables the coexistence of multiple normative systems, having distinct and often contradictory rules" [4]. Deciding the jurisdiction over a certain case, i.e. establishing which country has the jurisdiction over that case, is only one of the different tasks which have to be considered for modeling private international law, and Dung and Sartor in [4] also consider the issue of deciding the court having competence as well as the issue of establishing the legal system according to which the court has to decide. Dung and Sartor provide an analysis of private international law and propose a formal model based on modular argumentation.

In this paper, we specifically consider the so-called *renvoi*: determining the jurisdiction in one country may require for the determination of the jurisdiction in another country, a situation which may generate a sequence of references to different countries. Renvoi is not considered in [4]. Our work is not intended to deal with normative conflicts, as done in the belief revision approaches, starting with the seminal work in [3], and in the defeasible reasoning approaches to normative conflicts [8,6,7], which usually require some kind of priority among norms to be taken into account. In particular, [7] exploit defeasible logic to deal with the problem of interpreting the foreign law in a domestic legal system, dealing with normative an interpretative gaps.

As observed by Dung and Sartor, private international law enables the coexistence of multiple normative systems having contradictory rules without the necessity of defining priorities among the rules or systems: "conflicts between competences and between rules are avoided by distributing the cases between authorities of the different normative systems (jurisdiction) and by establishing what set of norms these authorities have to apply to each given case (choice of law)". There are only limited exceptions to this principle. This motivates our choice of dealing with scenarios, as the one introduced below, using a monotonic modal formalism, although, in the general case, a nonmonotonic formalism might be needed, such as modular argumentation in [4] and defeasible logic in [7].

Let us consider the following scenario. For simplicity, we do not consider the competence issue and assume the legal system of the country of jurisdiction is always applied.

Example 1.1 (Renvoi) Suppose the following laws hold in *every country*:

1. Inheritance matter, such as a property of heir, will be determined in jurisdiction of the home country of Descendant.
2. A legitimate child-parent relationship between Child and Parent will be determined in jurisdiction of the home country of Parent, or determined in jurisdiction of the home country of Spouse of Parent if there is a biological child-parent relationship between Child and Parent.
3. Marriage will be determined in jurisdiction of the home country of either spouse.
4. The home country is Person's nationality, if Person has only one nationality.
5. The home country is decided by the most related country for a Person, if Person has multiple nationality.

Domestic Rules that hold in *Japan*:

1. A marriage relationship holds between Spouse 1 and Spouse 2 if there is an agreement on marriage between Spouse1 and Spouse 2 and they register their marriage in Japan.
2. Child is a heir of a Parent if there is a child-parent relationship between them.
3. Child and Parent have a child-parent relationship if there is a legitimate child-parent relationship between them, or if there is a non-legitimate child-parent relationship between them.

Furthermore, we have the following facts:

- John has multiple nationalities of Country1 and Country2.
- Yoko has a single nationality of Japan.
- John usually lives in Coutry1.
- John and Yoko agreed to get married and registered their marriage at Coutry1.
- John and Yoko had a son named Taro.

Consider the following questions:

- 'John is married with Yoko' is valid in Japan?
- 'Taro is a heir of John' is valid in Japan?

Motivated by the scenario above, in this paper we introduce a formalism which is the rule-based (first order) fragment of a multimodal logic including context modalities as well as a (simplified) notion of common knowledge. For instance, in the example above, legislation of Japan can be represented by a modal context while general laws (such as the jurisdiction laws), which hold in any context, exploit context variables and global facts

are captured as (common) knowledge. In the simplified example we are considering, we assume a single set of jurisdiction rules rather than one for each country. The formalism is a rule based fragment of the modal language in [1], extended with context variables, and allows the interactions among contexts to be captured, context variables to occur within modalities and context names to be used as predicate arguments, thus supporting a simple combination of meta-predicates and modal constructs.

2. A modal formalization

We consider the rule-based fragment of the language in [1], extended by allowing variables to occur within modalities in rule definitions. Let $\mathscr{L}_k^{\Box}$ be a first order multimodal language containing: countably many variables, constants, function and predicate symbols; a finite set $Ctx = \{c_1, \ldots, c_n\}$ of constant symbols, called *contexts*; the logical connectives $\neg$, $\wedge$, $\supset$, and quantifiers $\forall$ and $\exists$, as in the predicate calculus, and the modalities $\Box$ and $[C]$, where C can be a variable or a context constant c_i in Ctx.

As the variables X occurring in a modality $[X]$ are intended to be instantiated only with constants in Ctx (as we will see later), the ground formulas of the language may contain two kinds of modalities: the modalities $[c_1], \ldots, [c_k]$, which represent k different contexts and the modality $\Box$, which can be regarded as a sort of (weak) "common knowledge" operator. A modal formula $[c_i]\alpha$ can be read as "α belongs to context c_i" or "agent c_i believes α". A modal formula $\Box\alpha$ can be read as "α holds in all contexts" or "all agents believe α".

Let A represent atomic formulas of the form $p(t_1, \ldots, t_s)$, where p a predicate symbol and $t_1, \ldots, t_s$ are terms of $\mathscr{L}$, and let $\top$ be a distinguished proposition (*true*). The syntax of the *clausal fragment* of $\mathscr{L}_k^{\Box}$ is the following:

$$G ::= \top \mid A \mid G_1 \wedge G_2 \mid \exists x G \mid [a_i]G \mid [X]D \mid \Box G$$
$$D ::= H \leftarrow G \mid D_1 \wedge D_2 \mid [c_i]D \mid [X]D \mid \Box D \mid \forall x D$$
$$H ::= A \mid [c_i]H \mid [X]D \mid \Box H$$

where G stands for a *goal*, D for a *clause* or *rule*, H for a *clause head*. Sequences of modalities may occur in front of goals, in front of rule heads and in front of rules. In the following D will interchangeably be regarded as a conjunction or a set of clauses (rules). A program P consists of a closed set of rules D. Also, we will adopt the convention that all the variables free in a rule D are implicitly universally quantified in front of it.

We say that a program P is *context safe* if each variable X occurring in a modality $[X]$ in a rule D of P, also occurs in an atom $context(X)$ in the body of D. We assume the predicate *context* has a built_in definition as $\forall X(context(X) \leftrightarrow (X = c_1 \vee \ldots \vee X = c_k))$, so that the context safeness condition guarantees that each context variable will be bounded to some context constant in all the possible groundings of the program P. In essence, this corresponds to a typing condition.

Referring to the example above, we can introduce the context `japan` containing the domestic rules specific to `japan`, using a Prolog-like notation, as follows:

```
□[japan] {
    heir(Child,Parent) :−child_parent_rel(Child,Parent).
    child_parent_rel(Child,Parent) :−
        legitimate_child_parent_rel(Child,Parent).
    child_parent_rel(Child,Parent) :−
        non_legitimate_child_parent_rel(Child,Parent).
```

marriage(Spouse1,Spouse2) :− agreement(marriage,Spouse1,Spouse2),
registered(marriage,Spouse1,Spouse2,japan). }

The modality □ in front of the context modality [japan] is needed to make each context definition globally visible from all the other contexts (so that a goal [japan]G can occur in the body of any, local or global, rule in the program). Observe that non-modal atoms in the body of rules in a context can be proved either locally to the same context or using other rule definitions as those introduced below.

The following rules establish the validity of a property in some country, based on properties which may hold in the same or other countries (or globally). They are intended to capture laws (1) and (2). The modalities [CountryA] and [CountryB] can only be instantiated with the constants japan, country1 and country2:

(*A*) □[CountryA](heir(Child,Parent) :−
context(CountryA),context(CountryB),
home_country(Parent,CountryB)), [CountryB]heir(Child,Parent)).

(*B*) □[CountryA](legitimate_child_parent_rel(Child,Parent) :−
context(CountryA),context(CountryB), home_country(Parent,CountryB),
[CountryB]legitimate_child_parent_rel(Child,Parent)).

(*C*) □[CountryA](legitimate_child_parent_rel(Child,Parent) :−
[CountryA]marriage(Parent,Spouse), home_country(Parent,CountryB),
[CountryB]legitimate_child_parent_rel(Child,Parent),
biological_child_parent_rel(Child,Parent)).

For instance, the second rule states that a legitimate child-parent relationship holds in CountryA if it holds in CountryB, where CountryB is the home country of the parent.

Global rules and facts can be encoded prefixing them with the □ operator, to mean that they are visible anywhere in the program (including contexts japan and country1):

□(marriage(Spouse1,Spouse2) :− marriage(Spouse2,Spouse1)).
□(home_country(Person,Country) :− single_nationality(Person,Country)).
□(home_country(Person,Country) :−
multi_nationality(Person,Country), most_related(Person,List,Country)).
□ multi_nationality(john,[country1,country2])).
□ habitual_residence(john,country1)).
□ single_nationality(yoko,japan)).
□ bilogical_child_parent_relation(taro,john)).
□ bilogical_child_parent_relation(taro,yoko)). ...

We refer to [2] for a description of the Kripke semantics and of the goal directed proof procedure for this rule based language.

Let us consider, as an example, the query "is Taro a heir of John valid in Japan?", which is captured by the goal [japan]heir(taro,john). This goal succeeds from the program above, using the following instance of rule (B):

□([japan]legitimate_child_parent_rel(taro,john) :−
context(japan),context(country1), home_country(john,country1),
[country1]legitimate_child_parent_rel(taro,john)).

and exploiting the definition of heir and child_parent_rel from the context japan, the definition of legitimate_child_parent_rel and marriage from the context country1, and the definition of biological_child_parent_rel, etc. from the global facts.

3. A formalization of renvoi in private international law

The formalization of the running example given in Section 2 establishes the validity of a property in some country, based on properties which may hold in the same or other countries. For instance, in rule (A), the validity of proposition heir(Child,Parent) in the context CountryA, depends on the validity of the same property in context CountryB. However, the rules in the program do not make any distinction among the validity of a property in a context and the jurisdiction of the same property in that context. Introducing such a distinction is essential to capture renvoi.

In particular, to check property heir(taro, john) in Japan, we need first to determine the jurisdiction of the property heir, with Japan as applying country, using rule (A), rather than using rule for heir in the context japan. Indeed, according to law (1), an inheritance matter, such as a property of heir, is to be determined in the jurisdiction of the home country of the parent. In this example, heir(taro, john) is to be determined in "country1", as "country1" is the home country of John.

We then reformulate our query as holds(heir(taro, john), japan), and we can introduce for heir, as for every property whose jurisdiction is to be determined, a rule:

$$\Box(\texttt{holds}(\texttt{heir}(\texttt{Child},\texttt{Parent}),\texttt{CountryA}) :- \\ [\texttt{CountryA}]\texttt{jurisd}(\texttt{heir}(\texttt{Child},\texttt{Parent}),\texttt{CountryB}), \\ [\texttt{CountryB}]\texttt{heir}(\texttt{Child},\texttt{Parent})).$$

where the goal [CountryA]jurisd(Matter,CountryB) is used to determine the jurisdiction CountryB of the Matter in CountryA i.e., the country in which the property heir(Child,Parent) is to be proven.

In general, to decide the jurisdiction of a matter, we first have to determine the property involved (for instance, the matter *hair* is concerned with the property *inheritance*). The jurisdiction of a matter is then given by the jurisdiction of the corresponding property. For simplicity, we will not exemplify this aspect here. We reformulate rule (A) to determine the jurisdiction of *heir* as follows:

$$(A)\ \Box[\texttt{CountryA}](\texttt{jurisd}(\texttt{heir}(\texttt{Child},\texttt{Parent}),\texttt{CountryC}) :- \\ \texttt{context}(\texttt{CountryA}),\texttt{context}(\texttt{CountryB}),\ \texttt{home\_country}(\texttt{Parent},\texttt{CountryB}), \\ [\texttt{CountryB}]\texttt{jurisd}(\texttt{heir}(\texttt{A},\texttt{B}),\texttt{CountryC})).$$

The determination tool may point out that we have to decide the validity of the matter in a different jurisdiction with respect to the current one. In rule (A) the jurisdiction for the matter heir(Child,Parent) is determined as the country of the parent (CountryB), which may be different from the current jurisdiction (CountryA). In such a case, we need again to decide the jurisdiction according to the private international law in the new country (i.e., CountryB). This is called a "renvoi". If a loop in the "renvoi" is detected, the jurisdiction is set to the starting country of the loop. For example, if the private international laws determines the following sequence of jurisdictions A, B, C, D, B, then we can decide the jurisdiction for the matter to be country B.

In order to deal with such a kind of loop in renvoi, we introduce the following general rule: $(R)\ \Box[\texttt{CountryA}]\Box[\texttt{CountryA}](\texttt{jurisd}(\texttt{Matter},\texttt{CountryA}) :- \top.$

For instance, when applying rule (A) in case $\texttt{home\_country}(\texttt{Parent}, \texttt{CountryA})$ holds, the second subgoal in the body of (A), i.e., $[\texttt{CountryB}]\texttt{jurisd}(\texttt{heir}(\texttt{A}, \texttt{B}), \texttt{CountryC})$, immediately succeeds with $\texttt{CountryB} = \texttt{CountryA}$, letting $\texttt{CountryC} = \texttt{CountryA}$, as the home country of the Parent is precisely CountryA, the country in which the determination of jurisdiction was issued.

To avoid other, spurious jurisdictions to be found, a "cut" should be added in the body of rule (R), although, of course, this is a feature which cannot be captured by rule-based language above. In [9] an encoding of cut by means of and announce predicate and an integrity constraint is exemplified, based on a notion of *global abduction*. To capture the correct behavior of renvoi, avoiding spurious solutions, an extension of the formalism with abduction or with some form of default negation would be needed. This will be subject of further work.

4. Conclusions and related work

Dung and Sartor in [4] provide a logical model of private international law, based on modular argumentation, as a way of coordinating the different normative systems without imposing a hierarchical order on them. They do not consider the issue of modeling chains of references. In this paper we exploit a rule based fragment of a modal logic with agent (or context) modalities, a simplified notion of common knowledge and context variables to capture renvoi (i.e., chains of references). As we have already mentioned above, our language is monotonic. Modeling private international law in its full generality might require a combination of both nonmonotonicity and modularity (see [4] and [7]). This motivates a nonmonotonic extension of the proposed rule-based formalism, that will be considered for future work.

The formalism we have considered is clearly related with other formalisms for dealing with multi-agent systems in computational logic and in Answer Set Programming (we refer to [5] for a survey).

References

[1] M. Baldoni, L. Giordano, and A. Martelli. A modal extension of logic programming: Modularity, beliefs and hypothetical reasoning. *J. Log. Comput.*, 8(5):597–635, 1998.

[2] Satoh K. Baldoni M., Giordano L. Renvoi in Private International Law: a Formalization with Modal Contexts. In *Technical Report Univ. Piemonte Orientale, www.di.unipmn.it/TechnicalReports/TR-INF-2019-10-05-UNIPMN.pdf*, October 2019.

[3] Alchourrón C.E. and Makinson D. Hierarchies of regulations and their logic. In *Hilpinen R. (eds) New Studies in Deontic Logic. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol 152*. Springer, Dordrecht, 1981.

[4] P. M. Dung and G. Sartor. The modular logic of private international law. *Artif. Intell. Law*, 19(2-3):233–261, 2011.

[5] A Dyoub, S. Costantini, and G. De Gasperis. Answer set programming and agents. *Knowledge Eng. Review*, 33:e19, 2018.

[6] B. Johnston and G. Governatori. Induction of defeasible logic theories in the legal domain. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, Edinburgh, Scotland, UK, June 24-28, 2003*, pages 204–213, 2003.

[7] A. Malerba, A. Rotolo, and G. Governatori. Interpretation across legal systems. In *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Annual Conference*, pages 83–92, 2016.

[8] H. Prakken and G. Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Artif. Intell. Law*, 4(3-4):331–368, 1996.

[9] K. Satoh. "all's well that ends well" - a proposal of global abduction. In *Proc. 10th Int. Workshop on Non-Monotonic Reasoning (NMR 2004), Whistler, Canada, June 6-8, 2004*, pages 360–367, 2004.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190319

A Dialogical Model of Case Law Dynamics

Trevor BENCH-CAPON and John HENDERSON
Dept. of Computer Science, University of Liverpool, UK

Abstract. We describe a set of dialogue moves which give a procedure to model the development of case law over a sequence of cases.

Keywords. legal case based reasoning, evolution of case law, argumentation.

1. Introduction

In [1] we discussed how case law develops as new cases arise. Our account was based on [2]. In this paper we take the ideas further, and provide a more precise account in the form of a set of dialogue moves. Current understanding of the domain is expressed as one or more rules, based on previous decisions. The existing rules will provide a reason to decide in accordance with them, but the other party can propose a counter argument based on a modification of the rules. This should be, as far as possible, consistent with previous decisions. If the counter argument is accepted, a refined understanding of the law will be expressed using the modification. In this way the theory may be reconstructed in the light of each new case to express an improved understanding.

We will first set out the machinery of our model, and the set of dialogue moves we have developed. These moves have been applied to an example based on the fictional area of law described in [1]. The example with a sequence of sixteen cases is given in full in the longer version of this paper [3] available at https://cgi.csc.liv.ac.uk/research/techreports/. The applicability to real cases was illustrated by the discussions in [1] of the thread of cases from [2] and some US 4th Amendment cases.

2. Elements of the Model

Throughout this paper we will give illustrations based on the example of [1], a fictitious welfare benefit, called *Independence Allowance* (IA). IA is paid to enable a measure of financial independence to those who are not expected to work.

As with HYPO [4], we represent cases as a set of facts. Facts are predicates of arity 1, and the domains may be boolean, an enumerated set of values, or a specified numeric range. The six facts used in the example are shown in Table 1.

Table 1. Factual Predicates. Sentence is a prison sentence: for non-prisoners it will be 0. If not yet entered workforce, value for entered workforce is age +1

Predicate	Domain	Predicate	Domain
Age	0-130	Apprentice	Yes,No
Sentence	0-30	Absence	0-130
Current Education	Primary, Secondary, College, University, No	Entered Workforce	12-130

Table 2. Factors for Independence Allowance. Vague factors have an upper and lower bound.

Factor	Rule
Infant	Age < 5
Child	Age < 16(low)/19(high)
PrimarySchoolchild	Current Education is Primary
Schoolchild	Current Education in {Primary, Secondary, College}
AgeofConsent	Age ≥ 16
Minor	Age ≥ 18
BelowSchoolLeavingAge	Age < 19
Young Adult	Age ≥ 18 AND Age < 30(low)/35(high)
Elderly	Age ≥ 60(low)/80(high)
Pensionable	Age ≥ 66
DeemedRetired	Age ≥ 72
Prisoner	Sentence > 0
Short Stay Prisoner	Sentence < 1
Full Time Education (FTE)	Current Education in {Primary, Secondary, College, University}
Continuing FTE	Current Education in {Primary, Secondary, College, University} AND Entered Workforce $>$ Age
Apprenticed	Apprentice = True
AbsenceDegree	Moderate if Absence/Age > 0.5: Substantial if Absence/Age > 0.8

Like [4] these facts can be mapped to factors using simple rules as shown in Table 2. These factors are intended to pick out potentially legally significant patterns of fact. For non-boolean facts we follow [5], so that where we have a dimension such as age or education, the factors identify points or ranges on that dimension. Some factors, like *child*, may lack precise bounds.

A rule will comprise a set of factors as antecedent, a *positive* outcome (to reflect the burden of proof) as consequent, and sets of positive exceptions and negative exceptions. Positive exceptions have a positive outcome despite the antecedent not being satisfied, and negative exceptions have a negative outcome despite the antecedent being satisfied. Each exception will be a set of factors.

3. Procedure

When the first case has been decided, the *ratio* of that case will offer a reason (as in the *reason model* of [5]) why the case was so decided. From this reason a rule can be derived, to be applied to future cases. This reason will be more general than the particular facts of the case and the terms used as the reason factor might be vague like *child*, or precise (at any given time) like *minor*.

Given a rule, a new case will either satisfy the rule, fall under a positive exception, fall under a negative exception, or be inapplicable. If it satisfies the rule or a positive exception that will be an argument for the positive side; if it satisfies a negative exception that will be an argument for the negative side. If no rule is applicable there is a "negation as failure" argument for the negative side. Although following the rule would apply the existing theory, the theory must be reconsidered in the light of the new case. There will therefore be a number of ways to respond by proposing modifications to the theory. We will now describe the responses and the rebuttals of these responses for each of the four situations. This gives a three-ply argumentation structure, which is commonly used in legal reasoning with cases, e.g. HYPO and its descendents [6].

3.1. First Ply

There are four possible moves here, two for the claimant and two against:

- ApplyRule(R). This can be played if there is a rule R for which the antecedent is satisfied by the new case. It argues for a positive outcome.
- ApplyPosException(R,E,V). This can be played if there is a rule R with a positive exception E which is satisfied by the new case. V is the value promoted by the exception. It argues for a positive outcome.
- ApplyNegException(R,N,V). This can be played if there is a rule R with a negative exception N which is satisfied by the new case. Again V is the value promoted by the exception. It argues for a negative outcome.
- NoRule. This can be played if there is no rule for which the antecedent is satisfied by the new case. It argues for a negative outcome.

3.2. Second Ply

The responses here will depend on the move made in the first ply.

3.2.1. ApplyRule, ApplyPosException and ApplyNegExcaption

There are a number of possible replies. The same replies can be used for all three of these first ply moves.

- DoesNotApply(R/PosEx/NegEx,Factor,NewFactor,V). This can be used if a factor in the rule, positive exception or negative exception is vague, and the case falls within the "penumbra of doubt". The respondent will propose a replacement *NewFactor* falling within the range of *Factor*, but such that the rule/positive exception/negative exception no longer applies (e.g. replace *child* with *ageOfConsent* for a 17 year old). V is the social value that would be promoted by adopting the new factor.
- ProposeException(R, Factor,V). This is used if there is a factor in the new case not present in the previous cases to which the rule applied. It proposes that factor as a negative exception for ApplyRule and ApplyPosException and as a new positive exception for ApplyNegExcaption. V suggests a social purpose which would be advanced by adopting the exception.

- Narrow(R/PosEx/NegEx, Factor, NewFactor, V). This prevents the rule, positive exception or negative exception from applying by proposing to replace *Factor* with *NewFactor* in the antecedent/positive exception/negative exception. *NewFactor* may be a smaller range of the same dimension as *Factor*, or require an additional fact to hold, (e.g replacing *FTE* with *Continuing FTE* from Table 2). It is argued that the narrowing would serve some social purpose, V.
- Broaden(NegEx,Factor,NewFactor,V). For ApplyRule this enables a negative exception to apply by broadening a factor in that negative exception. *NewFactor* may be a larger range of the same dimension as *Factor*, or remove a fact from the definition of *Factor*. It is argued that the broadening would serve some social purpose, V.

3.2.2. NoRule

ProposeException and *Broaden* can also be used here, by enabling an existing rule to apply, and there are two new moves.

- NewRule(R,V): This argues that a new rule is required for cases of this type. V suggests a social purpose which would be advanced by recognising the new type. As for all rules the outcome is positive.
- ProposeException(R,Factor,V). This can be used if there a factor in the new case which was not present in the previous cases, to enable a positive exception to the rule to apply. It proposes *Factor* as a positive exception. It differs from *NewRule*, in that the case is seen as an exception, rather than as a new, distinguished, group of cases.
- Broaden(R/PosEx, Factor,NewFactor): This enables a rule or positive exception to apply by broadening a factor in the antecedent/positive exception. *NewFactor* may be a larger range of the same dimension as *Factor*, or remove a fact from the definition of *Factor*.
- Analogy(R,Factor1,Factor2,Similarities): This contends that, on the basis of some similarities, a new factor, *Factor1*, is sufficiently analogous to an existing factor in the rule, *Factor2*, that they should be treated the same.

3.3. Third Ply

Each of these responses can be met with rebuttals. To rebut *DoesNotApply* the rebutter needs to include the case in the range of the factor in the antecedent.

- RuleDoesApply(Factor,NewFactor2,V2). Where *NewFactor2* is an alternative replacement for *Factor*, which does include the new case (e.g. *minor* rather than *ageOfConsent* fot a 17 year old). *V2* is the social value promoted by adopting the proposed new factor, and it is argued to be preferred to the value promoted by the factor proposed in the response.

For the moves depending on a value, *ProposeException*, *Narrow*, *Broaden* and *NewRule*, the rebuttal will turn on the desirability of promoting the value. A rebuttal can therefore deny that it does promote this value, or put forward a preferred value which the proposal would demote.

- NoPromotion(Factor,V): The proposed exception would not promote the desired value.
- Demotion(Factor,V2): The proposed exception would demote *Value2*, which is preferred to the value promoted according to the response.

For *DoesNotApply*, *NewRule* and *ProposedException* a rebuttal based on precedents can be used. If existing negative instances satisfy the new factor or the proposed rule, or positive instances contain the proposed exception, precedential constraint [7] excludes the proposed exception.

- Precedent(R/Exception,C): The proposed rule or exception was not applied in a precedent case, C.

When the response involves broadening or narrowing, an alternative rebuttal will contend that the proposed movement is too great to be acceptable.

- TooGreat(Factor,NewFactor). *NewFactor* would entail too great a movement and so *Factor* should continue to be used.

The final response is *Analogy*. To rebut this move, it is necessary to cite differences which make the proposed analogy unacceptable.

- NoAnalogy(Factor1,Factor2,Differences). *Differences* are the differences between the proposed new factor and the existing factor.

For example, if *father* was proposed as an analogy to *mother*, gender would be a difference, and might or might not be considered significant,

3.4. Resolution

After three plies a decision has to be made whether to stay with the original rule or to accept the modification. This will be a matter for argument, as in the Justices' Conference in the Supreme Court. Modelling these arguments is, however, outside the scope of this paper, which is intended to describe the public proceedings. The nature of the decision will depend on the type of the rebuttal. *RuleDoesApply*, *NoPromotion*, and *Demotion*, all turn on a value judgement (see e.g. [8]). Here the judges much choose which purpose or value they wish to promote. The preferred values are intended to reflect what [2] called the "common ideas of society", and may change over time, to adapt the law to changing social attitudes.

Precedent is a powerful rebuttal and should, given a strict interpretation of *stare decisis*, normally succeed. Sometimes, however, a precedent is not followed or even explicitly overruled: either it is too old and no longer represents the "common ideas of society", or it may be anomalous and conflict with other precedents, or perhaps a new value, not considered in the precedent, has subsequently emerged. In such cases the judge must decide whether there are sufficient grounds to disregard the precedent (see the discussion of *Robbins v California* in [1]).

TooGreat requires the judge to consider whether the proposed broadening or narrowing is too great a step to be acceptable, even if permitted by precedents [5]. Here the judge must come to a view on what seems appropriate.

Finally, *NoAnalogy* requires the judge to decide whether the similarities or the differences are more persuasive in the context of the case. A discussion of these matters can be found in [9].

A fully worked example stepping through a sequence of sixteen cases concerning Independence Allowance and an extended discussion is given in [3].

4. Concluding Remarks

Is automating the procedure feasible?. The first ply is straightforward: checking where a rule or exception applies is simple. The second ply is a little less straightforward. If a rule is applied, identifying a factor with questionable bounds, or factors that would represent a narrowing or broadening to exclude or include the case is easy, but identifying the rationale for these modifications is not. Similarly identifying a factor that *could* serve as an exception is easy, but whether the proposal would be sensible or not requires genuine understanding of the domain. Exceptions, broadenings and factors that would provide useful analogies or antecedents to new rules can be identified, but some semantic understanding is required to judge whether it would be worth advancing them. In the third ply, identifying whether there is a factor that would include the new case to allow *RuleDoesApply* is easy. Similarly discovering a precedent is not a problem. However, identifying differences for *NoAnalogy*, or that a value is not promoted or demoted requires a proper understanding of the terms [9]. That a broadening or narrowing is too great can always be argued, but judgement is required to form a view as to whether the claim is likely to be successful.

Thus two kinds of knowledge are required: knowledge about the rules, cases and background factors is precise and can be used to automatically suggest legally possible moves. Selecting the best move and assessing its worth, however, requires a far deeper understanding of the domain, of a sort that would require a comprehensive ontology. Fortunately such an ontology already exists in the heads of lawyers. This suggests that the proposed system should be designed as a support system, making suggestions as to the possible moves, which then require selection and justification with values by the user.

References

[1] J. Henderson and T. Bench-Capon, Describing the Development of Case Law, in: *Proceedings of the 17th ICAIL*, ACM, 2019, pp. 32–41.
[2] E.H. Levi, *An introduction to legal reasoning*, University of Chicago Press, 2013.
[3] T. Bench-Capon and J. Henderson, Modelling Case Law Dynamics with Dialogue Moves, Technical Report, ULCS-19-003, University of Liverpool, 2019.
[4] K. Ashley, *Modeling legal arguments: Reasoning with cases and hypotheticals*, MIT press, Cambridge, Mass., 1990.
[5] A. Rigoni, Representing dimensions within the reason model of precedent, *Artificial Intelligence and Law* **26**(1) (2018), 1–22.
[6] T. Bench-Capon, HYPO'S legacy: introduction to the virtual special issue, *Artificial Intelligence and Law* **25**(2) (2017), 205–250.
[7] J. Horty and T. Bench-Capon, A factor-based definition of precedential constraint, *Artificial Intelligence and Law* **20**(2) (2012), 181–214.
[8] M. Grabmair, Modeling Purposive Legal Argumentation and Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism, PhD thesis, U of Pitt, 2016.
[9] K. Atkinson and T. Bench-Capon, Reasoning with Legal Cases: Analogy or Rule Application?, in: *Proceedings of the 17th ICAIL*, 2019, pp. 12–21.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190320

Defeasible Systems in Legal Reasoning: A Comparative Assessment

Roberta CALEGARI [a,1], Giuseppe CONTISSA [b], Francesca LAGIOIA [b,c], Andrea OMICINI [a], and Giovanni SARTOR [b,c]

[a] *Dipartimento di Informatica – Scienza e Ingegneria (DISI),* ALMA MATER STUDIORUM–*Università di Bologna, Italy*
[b] *CIRSFID-Faculty of Law,* ALMA MATER STUDIORUM–*Università di Bologna, Italy*
[c] *European University Institute, Law Department, Italy*

Abstract. Different formalisms for defeasible reasoning have been used to represent legal knowledge and to reason with it. In this work, we provide an overview of the following logic-based approaches to defeasible reasoning: Defeasible Logic, Answer Set Programming, ABA+, ASPIC+, and DeLP. We compare features of these approaches from three perspectives: the logical model (knowledge representation), the method (computational mechanisms), and the technology (available software).On this basis, we identify and apply criteria for assessing their suitability for legal applications. We discuss the different approaches through a legal running example.

Keywords. AI and Law, legal reasoning, defeasible reasoning, argumentation

1. Introduction

Different approaches have been adopted to deal with defeasibility in law, including argumentation frameworks, which capture defeasibility through the interaction of conflicting arguments [1]. Even though defeasibility is a key aspect of legal reasoning, no comparative analysis of existing approaches has been carried out so far including their features, the available software tools, and more generally the advantages and disadvantages offered by different legal applications.

The present work aims to make a first step in this direction by pursuing two main goals. The first is to provide an assessment of some existing formalisms which may be useful in supporting informed choices by developers. The second is to identify some general methodological guidelines and criteria for determining which which formalisms for defeasible reasoning are more suitable for intended applications. We hope that by providing a framework for the analysis, comparison, and selection of the appropriate computable models of defeasible reasoning, we will contribute to strengthening the link between theory and application, and fostering successful integration. In our contribution, we have taken into account previous works dealing with the comparison of formalisms for defeasible reasoning [2,3], considering a more diverse set of formalisms, and focusing

[1]Corresponding Author: Roberta Calegari E-mail: roberta.calegari@unibo.it.

not only on expressiveness, but also on inference methods and the availability of software tools.

2. Running Example

To highlight the differences and similarities between the selected approaches, let us consider a hypothetical but realistic legal case concerning medical malpractice.

Patient John seeks compensation against Doctor Mary, claiming that Mary caused harm to him, and appeals to a legal rule stating that if a doctor causes harm to a patient, then the doctor has an obligation to pay damages, unless it is proven that the doctor was not negligent. This rule establishes a presumption of negligence against the doctor and a conditioned presumption of non-negligence favouring the doctor —the doctor was careful if he followed medical guidelines. Let us assume that expert evidence is provided by the two parties. In the following we consider different combinations of claims and see the conclusions generated by different approaches.

2.1. Nonmononic reasoners

Defeasible Logic DL is a well-know formalism for defeasible reasoning, originally proposed by Nute [4], and later extended in various directions, including deontic logic. Let us assume that expert witness Mark claims that there was harm, while expert witness Edward claims that the knowledge (the guidelines) was correctly followed.

```
patient('John'). doctor('Mary'). expert('Mark'). expert('Edward').
say('Mark',harmed(doctor('Mary'), patient('John'))). say('Edward',careful(doctor('Mary'))).
liable(doctor(D)) :- harmed(doctor(D), patient(P)).
neg liable(doctor(D)) :- used_correctly(knowledge, doctor(D)).
harmed(doctor(D), patient(P)) :- say(X,harmed(doctor(D), patient(P))), expert(X).
used_correctly(knowledge, doctor(D)) :- say(X,careful(doctor(D))), expert(X).
Answer for @liable(doctor('Mary')): no.
```

Running the query *@liable(doctor('Mary'))*, as well as *@neg liable('Mary')*, we obtain *false*, since the inferences for liable and not liable defeat one another. Adding a priority for the rule against liablility over the rule for liablilty, we obtain *yes* for Mary's non-liability. Assume now that Marks also intervenes on the issue of compliance with the guidelines, claiming that Mary did not follow the guidelines. The outcome is surprisingly that Mary is liable. In fact, the rule on the exclusion of liability would not be triggered, given that the antecedent *used_correctly(knowledge, doctor('Mary'))* could not be established, given the contradictory claims of the two experts. This aspect of the functioning of DL is called ambiguity blocking: when two conflicting inferences clash and there is no priority, the inferences cancel each other out.

ASP Answer set programming (ASP) is an approach to logic programming oriented towards difficult (primarily NP-hard) search problems. This input yields no results because of the unsolved contradiction between rules one and two. Note that the standard ASP format, used by systems such as Clingo and DLV2, does not support the use of preferences over rules. To express that the rule with conclusion *harmed(doctor(D), patient(P))* applies unless the doctor uses the knowledge correctly, we have to introduce a negation by failure *not used_correctly(knowledge, doctor(D))* in the body of that rule. If the input is so modified, Clingo provides a stable model according to which there is no liability.

```
liable(doctor(D)) :- harmed(doctor(D), patient(P)).
not liable(doctor(D)) :- used_correctly(knowledge, doctor(D)).
harmed(doctor(D), patient(P)) :- say(X,harmed(doctor(D), patient(P))), expert(X).
used_correctly(knowledge, doctor(D)) :- say(X,careful(doctor(D))), expert(X).
patient(john). doctor(mary). expert(mark). expert(edward).
say(mark,harmed(doctor(mary), patient(john))). say(edward,careful(doctor(mary))).
Answer: UNSATISFIABLE
```

2.2. Structured Argumentation

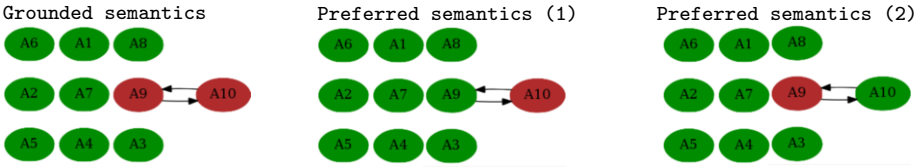
DeLP DeLP is a formalisation of defeasible reasoning in which the results of Defeasible Logic and Argumentation are combined [5]. The behavior is the same as DL, but it allows for ambiguity propagation, i.e., it may develop inferences based on conflicting propositions (as in ASPIC's preferred semantics).

```
Patient(john). Doctor(mary). Expert(mark). Expert(edward).
Say_harmed(mark, mary, john). Say_careful(edward, mary).
Liable(D) -< Harmed(D, P). ~Liable(D) -< Used_correctly(knowledge, D).
Harmed(D,P) -< Say_harmed(X,D, P), Doctor(D), Patient(P), Expert(X).
Used_correctly(knowledge, D) -< Say_careful(X,D), Doctor(D), Expert(X).
```

ASPIC+ ASPIC+ is a popular framework for structured argumentation, exploiting Dung's abstract semantics [6]. ASPIC allows users to choose from different semantics: grounded, preferred, semi-stable, and stable. The preferred semantic is particularly significant for the law, since it shows alternative extensions for unsolved conflicts. The use case is encoded as in the following listing with its corresponding argumentation graph under the grounded semantics, where both arguments A9 and A10 are rejected, since they defeat each other. The assessment changes if we add rule priorities. If we add a preference for rule r2 over rule r1 we find that A9 is now justified, while A10 is rejected. This shows an interesting difference between DL and ASPIC. In DL an unsolved conflict between two inferences means that such inferences (and the inferences expanding them) are irrelevant. In ASPIC the conflicting arguments can still defeat other arguments, and prevent the defeated arguments from being included in all preferred extensions.

```
Premises:patient(john);doc(mary);exp(mark);exp(edw)
Assumption:say_harm(mark,mary,john);say_careful(edw,mary)
Rules: [r1] harm(D,P) => liable(D);
[r2]used(K,D),doc(D) => ~liable(D);
[r3]say_harm(X,D,P),doc(D),patient(P),exp(X)=>harm(D,P);
[r4]say_careful(X,D),doc(D),exp(X)=>used(kb,D);
```

```
A1:say_careful(edw,mary)
A2:exp(edw) A6:exp(mark)
A3:say_harm(mark,mary,john)
A4: patient(john) A5: doc(mary)
A7: A1,A2,A5=>used(kb,mary)
A8: A3,A4,A5,A6=>harm(mary,john)
A9: A7,A5=>~liable(mary)
A10: A8=>liable(mary)
```



ABA+ In ABA+ arguments are sets of assumptions used to infer conclusions. Each rule has to be ground (i.e., no variables allowed). This is due to the fact that the tool uses a semantics-preserving mapping from ABA+ to abstract argumentation and uses ASPARTIX, for determining extensions. Moreover, ABA+ does not deal with preferences over rules, it only supports preferences over assumptions. The contraries of each assumption must be explicitly declared.

Table 1. Comparison under the *modelling* perspective: what aspects of legal argument can be captured.

	Defeasible Logic	**Argumentation**			**ASP**
		ABA+	ASPIC+	DeLP	
Model	DL Nute	AA Dung	AA Dung	DL Nute	ASP
Rules & Presumption	no argument notion	✓	✓	✓	no argument notion
Defeaters	✓	contraries	undercut, rebut, undermine	✓	✓
Preferences	✓	encoded	✓	✓	✓
Deontic Logic	✓	no	on strict rules	no	no
Argumentation Schemes	no	no	meta-ASPIC	no	no

3. Guidelines for Comparison and Evaluation

3.1. Model Perspective

Logical model and argument structure Even though different approaches to defeasible reasoning share a common background, they often adopt different logical models. DL is based on an inferential semantics, while ASP is based on the stable-set semantics of logic programming. On these approaches understanding an argument means exploring the inference tree derived by the application of the rules. On the other side, argumentation approaches explain their outcome through the attack and defeat relations between the applicable arguments. An advantage of Dung's abstract argumentation systems is the possibility of dealing with different semantics: the alternative between grounded and preferred semantics offers a choice between focalising on "sure" outcomes, or exploring alternatives that depend on possible solutions to rule conflicts.

Strict rules, defeasible rules, and presumptions. DL, as well as DeLP, ASPIC+ and ABA+, provides for the use of both strict and defeasible rules even thought in ABA+ defeasible rules are strict rules plus assumptions. While in some frameworks (ASPIC+ or ABA) assumptions are explicitly introduced, in other frameworks, such as DL, they can be modelled as rules with an empty antecedent.

Defeaters and attacks. Contrarily to other approaches, defeaters in DL and DeLP can be expressed via explicit rules (in the latter in the form of explicit undercutting defeaters, too). In DL and in argumentation-based approaches different types of attacks are distinguished (e.g., undercutting, rebutting, or undermining, while DeLP defines a single general notion of attack.

Preferences. Preferences among rules are supported by all approaches considered (in ABA preferences concern assumptions).

Deontic logic. Deontic modalities have been introduced in various logics to make them more suitable for legal reasoning. DL has been extended to support deontic modalities [7]. Such modalities are not supported natively by any ASP representation. With respect to ASPIC+ extensions to deontic logic have been defined, but have not yet been implemented in any reasoning tool.

Argumentation schemes. Patterns of informal argumentation often occur in real-world decision-making and in discussions between humans. These consideration lead to their formalization into argumentation frameworks, such as the meta-ASPIC model [8]. To the best of our knowledge no implementation is provided.

Table 2. Comparison of legal reasoning approaches from the *method* perspective, i.e., focusing on the reasoning/computational method used for legal inference and argumentation.

	Defeasible Logic	**Argumentation**			**ASP**
		ABA+	ASPIC+	DeLP	
Complexity	Polinomial/ linear	Polinomial			NP
Inconsistency handling (conflicting rules)	ambiguity blocking vs ambiguity propagation	ambiguity propagation →undecided			ambiguity propagation →unsatisfiability
Inconsistency handling (conflicting facts)	derive results despite them	derive results despite them			unsatisfiability
Credoulous/Skeptical	skeptical	✓	✓	skeptical	✓

3.2. Method Perspective

Complexity. All approaches are efficient in terms of reasoning time. However, acceptability of a proposition in an argumentation framework under grounded semantics can be computed in polynomial time, while defeasible logic, if restricted to propositional logic, has linear complexity. Finally ASP is NP.

Credoulous/skeptical and inconsistency handling: conflicting rules. An important difference among the three approaches is the way they handle inconsistency. DL, originally based on ambiguity blocking, has been "tuned" to obtain ambiguity propagation, i.e., the inferences based on conflicting claims. Outcomes similar to those obtained in ambiguity blocking and ambiguity propagation in DL can be obtained by a grounded or preferred Dung's semantics in ABA+ and ASPIC+, as discussed above. Conflicting facts in ASP lead to unsatisfiability because of the standard definition of consistency. DL as well as argumentation frameworks, on the other hand, handle inconsistencies to deliver defeasible outcomes according to their semantics.

3.3. Technology Perspective

Table 3. Comparison of legal reasoning approaches from the *technology* perspective (availability, accessibility and usability of software resources).

	Defeasible Logic	**Argumentation**			**ASP**
		ABA+	ASPIC+	DeLP	
Technology	d-Prolog/SPINdle	ABAPlus	TOAST	TweetyProject	Clingo/DLV
Open source	✓	✓	✓	✓	✓
IDE KB Support	no	no			no
Contradiction warning	no	no			✓

Technology. In terms of tools for reasoning support, there is at least one stable open source reasoner available for each approach. Sometimes no complete documentation manual is provided, leading to some difficulties in rule transcription.

IDE and Contradiction warning. While a number of reasoning tools have been developed, no tool is currently available to support knowledge encoding, to the best of our knowledge. This means that any legislation has to be manually written in the language supported by the reasoners. All tools lack a form of inconsistency highlighting.

4. Conclusion

Our analysis has shown that there is a strong convergence between different systems for defeasible reasoning. However, some differences exist, which may be relevant to different application domains. The possibility of using open (non-ground) rules in knowledge, and of using different instances of the same predicates in different rules, could be a key advantage, especially when the same rule has to be applied to different instances within a single argument. All the described systems, except for ABA+ and SPINDLE, have this feature. When a system has to deal with a high number of uncertain conflicts, the ability to rely not only on skeptical, but also on credulous reasoning may be important. Argumentation approaches (such as DeLP, ASPIC+, and ABA+) have this ability natively (though also ambiguity propagation in DL can also lead to similar results). When a system has to address complex issues of legal reasoning, and full explainability is required; the ability to provide a picture of existing arguments and of the relations between them, and an explanation on what arguments should or could be finally endorsed, may become a decisive feature. This is a feature we could find in ASPIC+ and ABA. From a technological perspective, many improvements need to be made in order to make existing tools really usable and effective in a distributed environment, as well as, documented and easily downloadable/deployable. The results presented here represent just a preliminary exploration of the logic-based approaches to defeasible reasoning, but it can provide starting guidelines for a methodological comparison of the various approaches.

Acknowledgments. This work has been partially supported by the European Union's Justice programme under Grant Agreement No. 800839 for the project "InterLex: Advisory and Training System for Internet-related private International Law".

References

[1] H. Prakken and G. Sartor, Law and logic: A review from an argumentation perspective, *Artificial Intelligence* **227** (2015), 214–245.

[2] S. Batsakis, G. Baryannis, G. Governatori, I. Tachmazidis and G. Antoniou, Legal Representation and Reasoning in Practice: A Critical Comparison, in: *Legal Knowledge and Information - JURIX 2018*, M. Palmirani, ed., Frontiers in Artificial Intelligence and Applications, IOS Press, United States, 2018, pp. 31–40.

[3] G. Charwat, W. Dvorak, S. Gaggl, J. Wallner and S. Woltran, Methods for solving reasoning problems in abstract argumentation – A survey, *Artificial Intelligence* (2014).

[4] D. Nute, *Defeasible Reasoning: A Philosophical Analysis in Prolog*, in: *Aspects of Artificial Intelligence*, J.H. Fetzer, ed., Springer Netherlands, Dordrecht, 1988, pp. 251–288. ISBN ISBN 978-94-009-2699-8.

[5] A. Garcia and G. Simari, Defeasible logic programming: an argumentative approach, *Theory and Practice of Logic Programming* **4**(2) (2004), 95–138.

[6] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* **77**(2) (1995), 321–357.

[7] G. Governatori, A. Rotolo and E. Calardo, Possible World Semantics for Defeasible Deontic Logic, in: *Deontic Logic in Computer Science*, T. Agotnes, J. Broersen and D. Elgesem, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 46–60. ISBN ISBN 978-3-642-31570-1.

[8] J. Muller, A. Hunter and P. Taylor, Meta-level argumentation with argument schemes, in: *International Conference on Scalable Uncertainty Management*, Springer, 2013, pp. 92–105.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190321

Legal Compliance in a Linked Open Data Framework

Enrico FRANCESCONI [a,1], Guido GOVERNATORI [b]
[a] *Publications Office of the European Union (Luxembourg); Istituto di Informatica Giuridica e Sistemi Giudiziari – IGSG-CNR (Italy)*
[b] *Data61, CSIRO, Australia*

Abstract An approach for legal compliance representation and checking within a Linked Open Data framework is presented. It is based on modeling deontic norms in terms of ontology and ontology property restrictions. It is also shown how the approach can handle norm defeasibility. Such methodology is implemented by decidable fragments of OWL 2, while legal reasoning is implemented by available decidable reasoners.

Keywords. Legal reasoning, Norm compliance, Semantic Web, OWL 2

1. Introduction

Machine readable rules represent a precondition for developing legal information systems with automatic reasoning facilities. Approaches were proposed to formalize reasoning on deontic notions [1], norm compliance [2] or legal argumentation [3]. In the Semantic Web, languages as OWL/RDF(S) for modeling real world scenarios, and mainly SWRL or RIF for legal rules are typically used. Recently LegalRuleML for legal rules modeling and defeasible reasoning has been proposed [5]. The Linked Open Data (LOD) approach to the Semantic Web is producing a growing amount of RDF triples for concepts, rules and facts. LOD principles recommend OWL/RDF(S), while implementing OWL 2 decidable profiles[2] allows to use available reasoners [6]. In this paper we discuss a legal reasoning framework [7] based on the distinction between the concepts of *Provision* and *Norm*. In particular, an approach for norm compliance in the LOD framework, based on decidable OWL 2 profiles, is here presented and tested. In Section 2 the distinction between *Provision* and *Norm* is discussed [8]; in Sections 3 and 4 norms modeling by ontologies able to implement defeasible norm compliance reasoning within a decidable framework is described and tested; in Section 5 some conclusions are reported.

2. Provisions and Norms

The legal order can be seen as a legal discourse composed by linguistic entities or *speech acts* [9] with descriptive or prescriptive functions. Every linguistic entity can be seen in

[1] Corresponding Author: E. Francesconi, e-mail: enrico.francesconi@igsg.cnr.it
[2] like OWL 2 DL, OWL 2 EL, OWL 2 QL and OWL 2 RL

a twofold perspective: as a set of signs, organized in words and sentences representing a normative statement, typically called *Provision* [10] [11], as well as the meaning for application of such normative statement, typically called *Norm* [8] [13]. Provisions and related norms have different roles and properties pertaining to the different domain they operate at. A provision represents the building block of the legal order. A norm can either modify the text of other provisions (as for the amendments) or can introduce restrictions on the real world (in case of obligations, for example).

Let's consider two examples of rules:

R1: *The supplier shall communicate to the consumer all the contractual terms and conditions*

R2: *According to a [country] law one cannot drive over 90 km/h*

Both rules are speech acts, namely *Provisions* in specific regulations. The *Provision Model* [11] [12] describes provisions in terms of *types* (as *Definition*, *Duty/Obligation*, *Right/Permission*) and *attributes* (as the *Bearer* or the *Counterpart* of a *Right/Permission*). According to the Provision Model, R1 can be classified as an *Obligation* of a *Supplier* towards *Consumer*, while R2 as an *Obligation* for any *Driver* in the related specific country. A Provision Model annotation can support advanced legal information retrieval (see [12]). When we consider the application of R1 and R2 on specific facts, we actually talk about *Norms*. Real world scenarios and facts can be effectively represented by ontologies and related individuals, respectively. Norms, providing constraints on the reality, can be modeled as restrictions on ontology properties. Legal compliance checking is a process aiming to verify if a fact, occurring in the real world, complies with a legal norm (namely the related restrictions).

Hereinafter we illustrate our approach for modeling norms with the aim of implementing legal compliance checking in a decidable framework.

3. Modeling norms for legal compliance checking

The scenario of R1 can be modeled by an ontology including a class Supplier, having a boolean property hasCommunicatedConditions (see Fig. 1, myo: is a namespace for a fictitious ontology "MyOntology"). Norm R1, expressing an obligation, states that suppliers must communicate purchasing conditions to the consumers. In our approach norm R1 is represented as a restriction on the property hasCommunicatedConditions able to identify the class SupplierR1CompliantIndividuals of individuals for which the value of the property under consideration is "true" (Fig. 1). The individuals of the class Supplier complying with this norm are all and only those belonging to the subclass SupplierR1Compliant. Such a representation results in the OWL 2 DL profile, allowing us to use an OWL 2 DL decidable reasoner, as for example Pellet[3]. The inferred model produced by Pellet establishes the rdfs:subClassOf relation between SupplierR1Compliant and Supplier. Therefore, compliance checking according to R1 is a problem of checking if an individual of type Supplier belongs also to the class SupplierR1Compliant. As a concrete example let's consider the two individuals myo:s1 and myo:s2 (Fig. 1) of Supplier. myo:s1 is an individual not compliant with R1, while myo:s2 is complaint with R1. The following SPARQL query:

[3]https://github.com/stardog-union/pellet

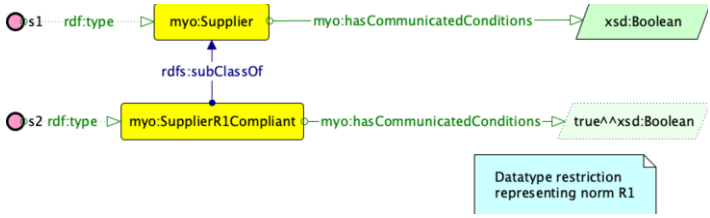
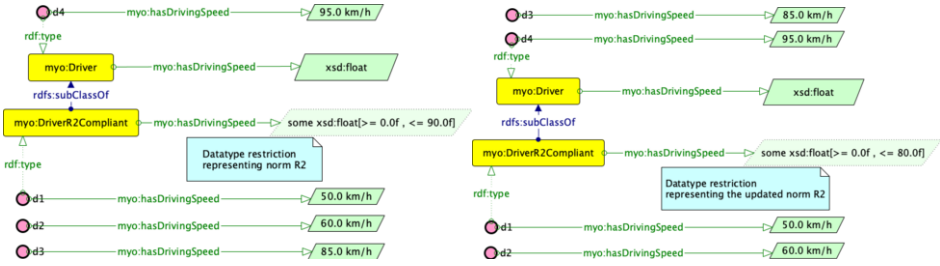


Figure 1. Norm R1 as restriction on the property hasCommunicatedConditions and examples of compliant and non-compliant individuals (the subclass relation between SupplierR1Compliant and Supplier is inferred).

```
SELECT ?x WHERE { ?x rdf:type myo:SupplierR1Compliant }
```

is able to select the individuals which are complaint with R1 (in our case s2).

In case of R2, the vehicles circulation scenario can be modeled in terms of an ontology including a class Driver, having a datatype property hasDrivingSpeed with range in the xsd:float datatype (Fig. 2a). Norm R2, expressing an obligation, states that, ac-



(a) Original norm R2 (b) New version of norm R2

Figure 2. Norm R2 represented as restriction on property hasDrivingSpeed and examples of compliant and non-compliant individuals.

cording to the related country law, one cannot drive over 90 km/h. In our approach norm R2 is represented as a restriction on the property myo:hasDrivingSpeed able to identify the class DriverR2CompliantIndividuals of individuals for which the values of the property under consideration are in the range [0.0, 90.0] km/h (Fig. 2a). In OWL 2 this can be expressed by the xsd:minInclusive and xsd:maxInclusive datatype bound properties. The individuals of the class Driver complying with this norm are all and only those belonging to the subclass DriverR2Compliant. Also such R2 modeling results in the OWL 2 DL decidable profile. The inferred model establishes a rdfs:subClassOf relation between DriverR2Compliant and Driver. Therefore, compliance checking for norm R2 is a problem of checking if an individual of type Driver belongs also to the class DriverR2Compliant. As a concrete example, let's consider the four individuals of the class Driver shown in Fig. 2a. The individual myo:d4 is not compliant with R2 having speed 95.0 Km/h $\geq$ 90.0 Km/h (Fig. 2a). The following query:

```
SELECT ?x WHERE { ?x rdf:type myo:DriverR2Compliant }
```

is able to select the individuals complaint with R2 (here myo:d1, myo:d2 and myo:d3).

In both the previous examples norm compliance checking is performed in a LOD framework within a decidable profile.

4. Modeling norms defeasibility for legal compliance checking

Defeasibility is a broad concept in the legal domain including "contrary to evidence" reasoning in argumentation systems [4], as well as reasoning with norm conflicts or norm exceptions [14] in normative systems. Two examples, one dealing with norm conflict and one with norm exception, are here modeled within a description logic framework able to provide support for defeasible reasoning, for example in norm compliance checking.

As first example let's consider rule R2, modeled in Section 3, and the following new version of rule R2, introducing a more strict driving speed limit at 80.0 Km/h:

R2 : *According to a [country] law one cannot drive over 80 km/h*

The *new version of R2* can defeat the previous compliance conclusions, in the sense that individuals, compliant with the *previous version of R2*, might not be compliant with it anymore. To cope with this norm change, the same model can be updated (without any change in class or property names) by changing the original datatype property restriction on myo:hasDrivingSpeed with a new one expressed by the *new version of R2*, as shown in Fig. 2b. Without changing anything on the individuals, their membership to the class myo:DriverR2Compliant changes so that, for example, the individual myo:d3, compliant with the previous version of R2 (Fig. 2a), is no more compliant with the new version of it (Fig. 2b). Therefore, the query able to select compliant individuals remains the same:

```
SELECT ?x WHERE { ?x rdf:type myo:DriverR2Compliant }
```

which retrieves the only now compliant individuals d1 and d2.

As second example, let's consider the following rule R3[4] which establishes the limits for engaging credit activities in Australia, composed by the following 3 statements ($R3 = R3a \cup R3b \cup R3c$):

R3a) *It is forbidden to engage in a credit activity without a credit license.*

R3b) *It is permitted to engage in a credit activity if acting on behalf of a principal and the principal holds a credit activity provided the principal has not been elected to the parliament.*

R3c) *It is permitted to engage in a credit activity if acting on behalf of a body corporate and the person has been appointed as representative of the body corporate.*

The defeasibility of norm R3 consists in an exception (R3a) which can defeat the previous compliance conclusions about the engagement of an agent in a credit activity, and in the exceptions of exception to it (R3b and R3c) which can defeat the conclusions about the prohibition established by R3a. The whole scenario addressed by norm R3 can be modeled through an ontology (Fig. 3) describing a class Agent and a specific subclass AgentEngagingCreditActivity of those agents who engage in a credit activity. Also in this case the deontic concepts Prohibition and Permission, expressed in R3a, R3b and R3c, are represented as restrictions on the datatype properties having domain AgentEngagingCreditActivity and expressing the conditions which the norm operates on. The individuals of the class Agent can engage a credit activity, thus belonging to the subclass AgentEngagingCreditActivity. According to the constraints expressed in R3, individuals can:

p1. have a credit license (hasCreditLicence)

p2. act on behalf of a principal (isActingOnBehalfOfPrincipal)

p3. have principal holding a credit activity (isPrincipleHoldingCreditActivity)

[4] section 29 of the Australian Consumer Credit Protection Act

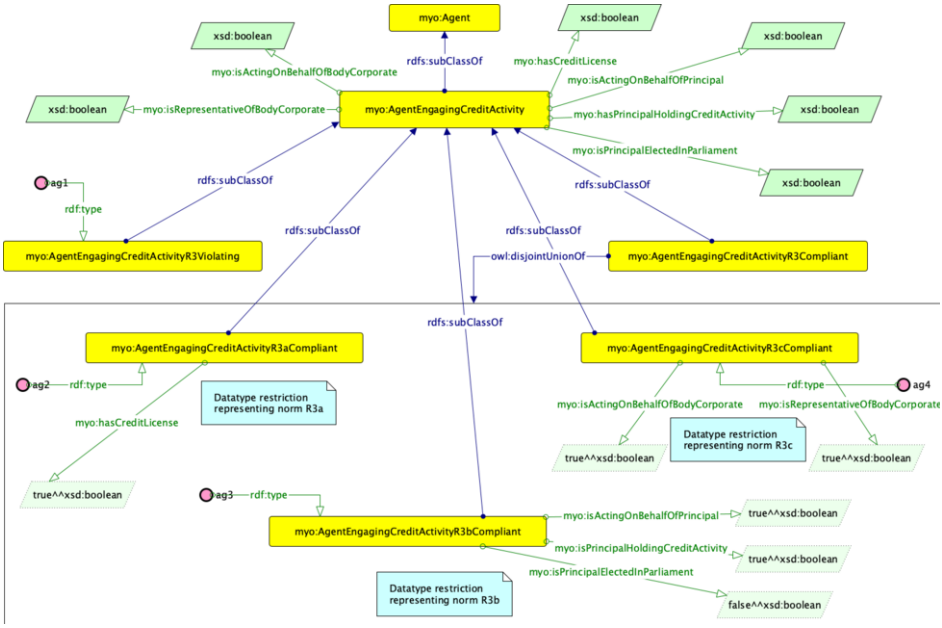


Figure 3. Norm R3 as restriction on AgentEngagingCreditActivity's properties (subclass relations between classes of compliant or violating individuals and AgentEngagingCreditActivity are inferred)

p4. have principal elected in Parliament (isPrincipleElectedInParliament)
p5. act on behalf of a body corporate (isActingOnBehalfOfBodyCorporate)
p6. act as representative of a body corporate (isRepresentativeOfBodyCorporate)

Norm R3a states that engaging in a credit activity "is forbidden without a credit license". Therefore, if an individual of the class AgentEngagingCreditActivity has a credit license (hasCreditLicence = "true") the activity is permitted. This is modeled as restriction on the property hasCreditLicence so to create a subclass AgentEngagingCreditActivityR3aCompliant of individuals having "true" as value of the property hasCreditLicence. In Fig. 3 the individual ag2 is compliant to the norm R3a.

Norm R3b states that the activity is permitted also when the individual "is acting on behalf of a principle" and "the principle holds a credit activity" and "the principle is not elected in Parliament". This is modeled through a multiple restriction on the properties isActingOnBehalfOfPrincipal = true, isPrincipleHoldingCreditActivity = true and isPrincipleElectedInParliament = false, to create a subclass of individuals for which the previous three restrictions contemporarily hold. In Fig.3, ag3 is compliant with norm R3b. Very similar considerations can be made for R3c modeling, concerning restrictions on the properties expressed by the conditions for individuals compliant with R3c.

The individuals compliant with the whole R3, composed by R3a, R3b and R3c, are therefore those belonging to the class AgentEngagingCreditActivityR3Compliant, obtained as disjoint union of the classes AgentEngagingCreditActivityR3aCompliant, AgentEngagingCreditActivityR3bCompliant, AgentEngagingCreditActivityR3cCompliant.

In all the other cases, engaging in a credit activity is forbidden. Therefore, the individuals which do not respect a combination of restrictions on properties of compliant individuals, violate norm R3, namely they belong to the class AgentEngagingCreditAc-

tivityR3Violating. In Fig. 3, ag1 violates norm R3. The combination of the property restrictions p1, ..., p6 able to identify individuals violating norm R3 can be obtained by the negation of the combination of properties of compliant individuals. In the case of R3, and applying the De Morgan laws, we obtain:

$$\neg[p1 \vee (p2 \wedge p3 \wedge \neg p4) \vee (p5 \wedge p6)] = \neg p1 \wedge (\neg p2 \vee \neg p3 \vee p4) \wedge (\neg p5 \vee \neg p6)$$

In order to verify which individuals are compliant or are violating R3, the following queries on the inferred model are respectively sufficient:

```
SELECT ?x WHERE {?x rdf:type myo:AgentEngagingCreditActivityR3Compliant}
SELECT ?x WHERE {?x rdf:type myo:AgentEngagingCreditActivityR3Violating}
```

5. Conclusions and future developments

In this paper we have presented an approach for legal compliance checking within a LOD framework. It is based on the representation of deontic norms in terms of domain ontology and ontology properties restrictions. The approach is implemented by decidable fragments of OWL 2, able to guarantee computational tractability and the possibility of using available reasoners. We have also shown how this approach can handle norm defeasibility. A development of this work will be the identification of specific knowledge modeling patterns able to represent defeasible deontic norms for legal compliance.

References

[1] J. M. Broersen, C. Condoravdi, N. Shyam, and G. Pigozzi, eds., *Deontic Logic and Normative Systems - 14th Int. Conference, DEON 2018*, (Utrecht, The Netherlands), College Publications, July 3-6 2018.

[2] R. Muthuri, G. Boella, J. Hulstijn, S. Capecchi, and L. Humphreys, "Compliance patterns: harnessing value modeling and legal interpretation to manage regulatory conversations," in *Proc. of ICAIL 2017*, (London, United Kingdom,), pp. 139–148, ACM, June 12-16 2017.

[3] H. Prakken and G. Sartor, "Law and logic: A review from an argumentation perspective," *Artificial Intelligence*, no. 227, pp. 214–245, 2015.

[4] T. Athan, G. Governatori, M. Palmirani, A. Paschke, and A. Wyner, "LegalRuleML: Design principles and foundations," in *The 11th Reasoning Web Summer School*, 2015.

[5] G. Governatori, M. Hashmi, H. Lam, S. Villata, and M. Palmirani, "Semantic business process regulatory compliance checking using LegalRuleML," in *Knowledge Engineering and Knowledge Management*, no. 10024 in LNAI, pp. 746–761, Springer International, 2016.

[6] F. Gandon, G. Governatori, and S. Villata, "Normative requirements as linked data," in *Proceeding of the JURIX Conference* (A. Wyner and G. Casini, eds.), vol. 302, pp. 1–10, IOS Press, 2017.

[7] E. Francesconi, "Reasoning with deontic notions in a decidable framework," in *Knowledge of the Law in the Big Data Age* (G. Peruginelli and S. Faro, eds.), vol. 317 of *Frontiers in Artificial Intelligence and Applications*, pp. 63–77, IOS Press, 2019.

[8] A. Marmor, *The Language of Law*. No. 978-0-19-871453-8, Oxford University Press, 2014.

[9] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

[10] J. Raz, *The Concept of a Legal System*. Oxford University Press, 1980.

[11] C. Biagioli, *Modelli Funzionali delle Leggi. Verso testi legislativi autoesplicativi*, vol. 6 of *Legal Information and Communications Technologies Series*. European Press Academic Publishing, 2009.

[12] E. Francesconi, "A description logic framework for advanced accessing and reasoning over normative provisions," *Int. Journal on Artificial Intelligence and Law*, vol. 22, no. 3, pp. 291–311, 2014.

[13] H. Kelsen, *General Theory of Norms*. Clarendon Press, Oxford, 1991.

[14] G. Casini, T. A. Meyer, K. Moodley, and U. S. I. Varzinczak, "Introducing defeasibility into owl ontologies," in *Proceedings of the International Semantic Web Conference*, pp. 409–426, 2015.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190322

Deontic Closure and Conflict in Legal Reasoning

Guido GOVERNATORI [a,1] Robert MULLINS [b]
[a] *Data61, CSIRO, Australia*
[b] *School of Law, The University of Queensland, Australia*

Abstract. We identify some legal reasoning patterns concerning deontic closure and conflicts in defeasible deontic logics. First, whether the logic allows the derivation of permissions from conflicting norms. Second, whether the logic treats norms as closed under logical implication. We suggest appropriate approaches for legal settings.

1. Introduction

Normative systems can be understood as sets of norms, with each norm represented as an"**IF** conditions **THEN** conclusion" structure [10, 4]. Rule-based systems of this sort provide an adequate framework for the representation of norms, normative systems and legal knowledge (see, for example, [5, 9] for some rule-based frameworks for legal reasoning). It has been argued that for the successful representation of norms and legal reasoning rule-based systems should account for both defeasible reasoning [8], and reasoning with deontic concepts [7]. We refer to a system combing both aspects as a defeasible deontic logic. The use of defeasible deontic logics is a well-established aspect of research on legal reasoning and argumentation. Here we introduce and discuss a number of complex reasoning patterns that arise when using defeasible deontic logics to represent legal reasoning. The patterns concern the logics' approach to deontic closure and conflicts. In each case we provide examples and suggest the most appropriate approach for legal settings.

2. Defeasible Deontic Logic

We do not make use of any specific defeasible deontic logic. Rather, we outline general abstract characteristic of these logics before considering a variety of reasoning patterns any such logic will need to accommodate. We assume formulas from a logical language that includes the deontic operators O and P for obligation and permission, and the implication operator $\rightarrow$. As usual in deontic logic we assume that a prohibition is a negative obligation, i.e., $\mathsf{F}a \equiv \mathsf{O}\neg a$. Two additional operators P_w and P_s denote *weak* and *strong* permission, respectively. Eventually, a 'generic' permission will be understood as the disjunction of the corresponding strong and weak permissions, namely: $\mathsf{P}a \equiv (\mathsf{P}_w a \vee \mathsf{P}_s a)$. We also assume that a normative system is consistent. Namely, we assume: $\mathsf{O}a \wedge \mathsf{O}\neg a \rightarrow \bot$ and $\mathsf{O}\neg a \wedge \mathsf{P}a \rightarrow \bot$. Norms are represented by rules, where a rule is an expression with the following form:

[1]Corresponding author: Guido Governatori, e-mail: guido.governatori@data61.csiro.au

$$r\colon A_1, \ldots A_n \Rightarrow C \tag{1}$$

where r is the name/id of the rules, $A_1,\ldots A_n$, the antecedents of the rule, are formulas in the language (including deontic formulas), and C the conclusion of the rule is a formula in the language (including a deontic formula). Notice that $\Rightarrow$ is not an operator of the object language.

We stipulate that all facts are given as formulas not containing deontic operators. Obligations, prohibitions, and permissions are are derived from rules (corresponding to norms) using an all-things-considered argumentation structure as follows:

To derive conclusion C

- there should be a rule (argument) for C such that the rule is applicable and
- all possible rules conflicting with the conclusion C (counterargument for C) are either
  - ∗ rebutted (meaning that the rule is not applicable)
  - ∗ defeated (meaning the rule is weaker than an applicable rule for the conclusion we want to prove)

The exception to the structure above is the derivation of weak permissions. As we state below, for our purposes we derive that something is weakly permitted just in case we fail to derive that the opposite is obligatory.

In the rest of the paper, we use the notation $r\colon \cdots \Rightarrow C$ to represent a rule r when the antecedent of the rule is assumed to hold and the content of the antecedent is not relevant for the discussion.

Note that the abstract formulation we provide in this section is compatible with several existing defeasible deontic logics, for example [5, 9, 11].

3. Permission from conflicting norms

Permission is sometimes defined as the absence of a prohibition or the absence of an obligation to the opposite. *Weak permission* arises where there are no norms against the permitted behaviour and no norms expressly permitting the behaviour. In other words, p is weakly permitted if $\not\vdash \mathsf{O}p$. *Strong permission* arises where norms explicitly permit an action in derogation from inconsistent norms [12, 1].

We first consider the case where there are two norms that are directly in conflict, the first of which makes a obligatory and the second forbidding a, where there is no mechanism for resolving the conflict. Accordingly, suppose we have the following norms/rules:

$$n_1\colon \ldots \Rightarrow \mathsf{O}a \qquad\qquad n_2\colon \ldots \Rightarrow \mathsf{O}\neg a \tag{2}$$

Under the sceptical reading we have that $\not\vdash \mathsf{O}a$ and $\not\vdash \mathsf{O}\neg a$. For a sceptical reasoner, however, it may still be necessary to determine whether a is permissible. This is not because the reasoner needs to determine whether a is legal, but because whether a is permitted may trigger other normative requirements. So, for example, if we have a third norm n_3 with the following form:

$$n_3\colon \mathsf{P}a \Rightarrow X. \tag{3}$$

a sceptical reasoner needs to determine whether a is permitted in order to determine whether X holds.

There are three options for this case. Option 1 is to argue that both a and $\neg a$ are permitted. There are rules that mandate, respectively, a and $\neg a$. On the assumption that the normative system is consistent, if we derive the obligation that a, we should also

derive that a is permissible; otherwise we would have an obligation that a that was impermissible to discharge. Thus, n_2 is a norm that "strongly" derogates the obligation that a (strongly, in the sense that it would make a forbidden), and n_1 is a norm explicitly derogating the prohibition that a. Option 2 is to assert that neither a nor $\neg a$ is permitted: whatever one does the resultant state of affairs will be illegal. We believe however that this is not an acceptable option, assuming the consistency of the set of obligations. Option 3 is to assume that there is a gap (for defeasible deontic logics admitting such an option). We believe that only the first option is appropriate for legal reasoning. If it is possible to have conflicting norms (as n_1, n_2), then it is reasonable to assume, as we hinted in Section 2, that the normative system remains consistent, since real-life legal systems provide principles to resolve conflicts (against option 2). In adjudication of a case, moreover, a judge has to decide whether X holds, and in general cannot refrain from taking a decision (against option 3). A decision has to be taken systematically (taking into account that n_3 could itself derogate from another another norm with $\neg X$ as its consequence).

Moreover, there are situations where sceptical reasoners will be required to determine whether the permission in the antecedent of the norm is strong or weak. Thus, suppose instead of n_3 we have

$$n_3' : \mathsf{P}_w a \Rightarrow Y \tag{4}$$

where a is weakly permitted, does Y hold? Or

$$n_3'' : \mathsf{P}_s a \Rightarrow Z \tag{5}$$

where a is strongly permitted, does Z hold?

Norms with a strong permission in the antecedent are not uncommon. They can be used to formulate norms expressing rights where the right confers a permission on one party that impliedly confers an obligation on another party. Strong permissions also appear in the antecedent of a norm when a party exercises an entitlement (or permissions). Another case of a norm corresponding to n_3 (or more precisely, n_3') would be when the norm recites:

> If in derogation to Section X y does Z, then it is obligatory that ...

For an example with an explicit weak permission consider the following three norms.

Section 1 If a person lives in Italy for more than 183 consecutive days over a 12-month period, then the person is obliged to pay taxes in Italy on the person's worldwide income.

Section 2 A citizen of a country that signed a mutual tax treaty with Italy is exempt from paying her taxes in Italy, provided the citizen maintains fiscal residence in the country that signed the tax treaty with Italy.

Section 3 If a person is exempt from paying taxes on her worldwide income in Italy for reasons not listed in Section 2 and elects not to pay such taxes in Italy, then the person has to declare the countries where the person pays such taxes.

The first norm (Section 1) sets an obligation based on some factual condition for the obligation to be in force. Section 2 provides an explicit derogation to the obligation set in Section 1. Thus, when Section 2 applies, the obligation in Section 1 is not in force and we have a strong permission. If the applicability condition for Section 1 does not hold, then the corresponding obligation is not in force, and the opposite activity (i.e., not paying income taxes in Italy) is permitted. This permission is weak: assuming there are

no other norms, there is no norm that explicitly exempts the payment of taxes in Italy. Finally, the clause in Section 3 takes exemption from Section 1 as part of the condition of applicability of another legal requirement (the obligation to declare where the income taxes are going to be paid). In this case, the permission invoked is weak; the provision explicitly excludes the explicit derogations provided in Section 2.

4. Closure under logical implication

In this section we address the closure of obligations under logical implication in cases of normative conflict. We introduce several cases distinguished by the nature of the conflict in question. Consider, again, the norm: $n_1 : \ldots \Rightarrow \mathsf{O}a$. Suppose that the norm is applicable, and therefore the obligation $\mathsf{O}a$ is in force. Is the permission $\mathsf{P}a$ in force as well? What about the strong and weak versions of the permissions (i.e., $\mathsf{P}_s a$ and $\mathsf{P}_w a$)? Suppose that, in addition to n_1, we have the implication $a \rightarrow b$. The issues are now:

1. Are we allowed to conclude $\mathsf{O}b$? If so, under what conditions (e.g., the norms that either make b forbidden or $\neg b$ permitted are not applicable or defeated)?
2. Are we allowed to conclude $\mathsf{P}b$? If so, under what conditions (e.g., that the norms forbidding b are either not applicable or defeated)?

The appropriateness of logical closure in the context of legal requirements has been debated by Lou Goble [3] and John Broome [2]. Broome argues that closure is not a feature of positive requirements (such as law). In response, Goble offers the example of law that says 'there shall be no camping at any time on public streets', 'it does not seem much of a defense for a camper to plead that the law never said that there should be no camping on the streets on Thursday night'. Broome's reply is that the law you have breached does not forbid camping on Thursday, it forbids camping *at any time*. Here we introduce some observations in support of Broome's position.

Continuing with Goble's example, suppose we had two norms, one creating an obligation not to camp at any time and another permitting camping on a Tuesday:

$$n_6 : \ldots \Rightarrow \mathsf{O}\neg camping \qquad\qquad n_7 : \ldots \Rightarrow \mathsf{P}camping\ tuesday$$

where the implication $\neg camping \rightarrow \neg camping\ tuesday$ holds, and there are are no further rules. Here the safest conclusion seems to be $\neg\mathsf{O}\neg camping\ tuesday$, which supports the conclusion that we cannot derive $\mathsf{O}b$ from $\mathsf{O}a$ through closure unless there are no applicable or undefeated norms that make b forbidden or $\neg b$ permitted.

However, even where legal obligation is not closed under logical implication, legal systems probably feature defeasible closure rules as part of their interpretive canon, with something like the following form: $OX \wedge (X \rightarrow Y) \Rightarrow OY$

In the camping example, this defeasible closure rule would be defeated by the more specific permissive norm n_7.

Issue (2) seems to be related to the question of whether we can derive permission from conflicting norms, though the conflict in this case is indirect. Supposing we have two norms, one of which imposes an obligation on all campers in the forest not to light a fire of any sort, and a second which imposes an obligation on all park rangers to light their fire with a gas burner:

$$n_8 \ldots \Rightarrow \mathsf{O}\neg fire \qquad\qquad n_9 \ldots \Rightarrow \mathsf{O}gas\ burner$$

Where the implication $\neg \mathit{fire} \rightarrow \neg \mathit{gas\ burner}$ holds. It does not seem like the appropriate conclusion, in this case, is $\mathsf{P}\neg \mathit{gas\ burner}$. This seems to suggest that in these sorts of cases the condition for concluding $\mathsf{P}b$ through closure is that there are no applicable or undefeated norms forbidding b.

Conflicts and Closure: partially direct conflict Here, we discuss cases where logical closure created conflicts between norms. Consider norm n_1 and the norm

$$n_4 \colon \ldots \Rightarrow \mathsf{O}\neg b \tag{6}$$

and either the implication $a \rightarrow b$ or the weaker $a \Rightarrow b$. What can we conclude: namely, $\mathsf{O}a$, $\mathsf{O}b$, or $\mathsf{O}\neg b$? And, more importantly, under what conditions are these conclusions correct?

To begin with, there is an intuitive conflict between n_1 and n_4. If we accept that obligation is not closed under logical consequence, then we need some sort of defeasible closure rule in order to explain the apparent conflict between the two norms. Suppose that we have our two norms n_8 and n_9, mentioned above, where, of course, $\mathit{gas\ burner} \rightarrow \mathit{fire}$. If there is no closure under logical consequence, then there is no conflict (obviously), and we have $\mathsf{O}\mathit{gas\ burner}$ and $\mathsf{O}\neg \mathit{fire}$. This seems right to us *prima facie*. If we were looking to *describe* the law, we would say that there is both an obligation not to light a fire and an obligation to light any fire using a gas burner. However, we also want to be able to describe why there is an intuitive conflict between the two laws. Thus we need some sort of defeasible meta-norm, like

$$n_{10} \colon \mathsf{O}\neg \mathit{fire} \wedge (\neg \mathit{fire} \rightarrow \neg \mathit{gas\ burner}) \Rightarrow \mathsf{O}\neg \mathit{gas\ burner}.$$

This meta-norm, when combined with n_8, would then be in conflict with n_9. Our legal intuition is that the resolution of this conflict would then depend on the relative priorities of n_8 and n_9. So if n_8 is higher in priority than n_9, then the argument chain involving n_8 and n_{10} will prevail, and there will be an obligation not to light a fire.

Conflict and Closure: fully indirect conflict The previous case involved a direct conflict. In other cases conflict is not direct, but induced by logical implications or other (constitutive) norms. For example, the norms

$$n_1 \colon \ldots \Rightarrow \mathsf{O}a \qquad\qquad n_5 \colon \ldots \Rightarrow \mathsf{O}b$$

paired with the implications $a \rightarrow c$ and $b \rightarrow \neg c$ are in indirect conflict. As before, we must address the correct sceptical response. What can we conclude: $\mathsf{O}a$, $\mathsf{O}b$, $\mathsf{O}c$, $\mathsf{O}\neg c$? More importantly, under what conditions are these conclusions correct?

The need to accommodate indirect conflicts is clearest if the prescriptions in n_1 and n_5 are "compensable", namely, instead of n_1 and n_5 we have[2]

$$n_1''' \colon \ldots \Rightarrow \mathsf{O}a \otimes \mathsf{O}d \qquad\qquad n_5' \colon \ldots \Rightarrow \mathsf{O}b \otimes \mathsf{O}e$$

Our view is that indirect conflicts between obligations should not block their derivation. Occasionally, these sorts of indirect conflict arise in private law without affecting the validity of the obligations in question. For example, in *J Lauritzen A.S v Wijsmuller B.V*[3] the defendants operated two ships, the Super Servant 1 and the Super Servant 2,

[2]Here we use the notation proposed by [6] to model compensatory obligations, $\mathsf{O}a \otimes \mathsf{O}b$ means that the a is the primary obligation, and the obligation of b is in force when the obligation of a is violated (namely, $\neg a$ holds, and the fulfilment of b compensates for the violations of the obligation of a).

[3][1990] 1 Lloyd's Rep 1.

which they planned to use to complete two different contracts. After Super Servant 2 sunk off the coast of Zaire, the defendants could not fulfil both contracts using only Super Servant 1. The defendants decided it was impossible to fulfil the plaintiff's contract. The court nonetheless held that the defendants were in breach of contract, and ordered the defendant to pay damages to the plaintiff.

Suppose we have two basic contractual rules:

$$n_1\colon \ldots \Rightarrow \mathsf{O}Contract1 \qquad n_5\colon \ldots \Rightarrow \mathsf{O}Contract2$$

Along with the implications:

$$Contract1 \rightarrow Super\ Servant1 \qquad Contract2 \rightarrow \neg Super\ Servant1$$

Descriptively, in cases like this where there is indirect conflict, it is best to say that the defendant has both obligations (i.e. $\mathsf{O}Contract1$ and $\mathsf{O}Contract2$).

If we introduce the idea that both obligations are compensable into the logic, then the sense in which the two obligations conflict is particularly clear. If the defendant uses the Super Servant 1 to perform $Contract1$, then they owe the party to $Contract2$ compensation, or *vice versa*. So we now have:

$$n_1'''\colon \ldots \Rightarrow \mathsf{O}Contract1 \otimes \mathsf{O}Damages1 \tag{7}$$

$$n_5'\colon \ldots \Rightarrow \mathsf{O}Contract2 \otimes \mathsf{O}Damages2 \tag{8}$$

We have both obligations, and whichever one is not fulfilled will be compensable. Indirect conflict between the two obligations does not block their derivation.

5. Conclusions

In this contribution we discussed some reasoning patterns that may arise in the use of defeasible deontic logics for the representation of legal knowledge. In each case, we argued that certain reasoning patterns must be preserved in order to ensure that defeasible deontic logics are appropriate for representation of legal reasoning. The approach we favour generally tolerates forms of direct and indirect conflict between norms while rejecting the strict closure of deontic operators under logical consequence.

References

[1] C.E. Alchourrón and E. Bulygin. Permission and permissive norms. *Theorie der Normen*:349–371, 1984.
[2] J. Broome. *Rationality Through Reasoning*. Wiley-Blackwell, 2013.
[3] L. Goble. Normative Conflicts and the Logic of 'Ought'. *Noûs*, 43:450–489, 2009.
[4] T.F. Gordon, G. Governatori, and A. Rotolo. Rules and Norms: Requirements for Rule Interchange Languages in the Legal Domain. In *RuleML 2009*, pages 282–296. Springer, 2009.
[5] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco. Computing Strong and Weak Permissions in Defeasible Logic. *Journal of Philosophical Logic*, 42:799–829, 2013.
[6] G. Governatori and A. Rotolo. Logic of Violations: A Gentzen System for Reasoning with Contrary-To-Duty Obligations. *Australasian Journal of Logic*, 4:193–215, 2006.
[7] A.J.I. Jones and M.J. Sergot. Deontic logic in the representation of law: Towards a methodology. *Artif. Intell. Law*, 1:45–64, 1992.
[8] H. Prakken and G. Sartor. Law and logic: A review from an argumentation perspective. *Artif. Intell.*, 227:214–245, 2015.
[9] H. Prakken, A.Z. Wyner, T.J.M. Bench-Capon, and K. Atkinson. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25:1141–1166, 2015.
[10] G. Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, 2005.
[11] L.W.N. van der Torre and S. Villata. An ASPIC-based legal argumentation framework for deontic reasoning. In *COMMA 2014*, pages 421–432. IOS Press, 2014.
[12] G.H. von Wright. *Norm and Action: A Logical Inquiry*. Routledge & Kegan Paul London, 1963.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190323

A Computational Model for Pragmatic Oddity

Guido GOVERNATORI [a,1] Antonino ROTOLO [b]
[a] *Data61, CSIRO, Australia*
[b] *University of Bologna, Italy*

Abstract. We introduce a computational model based on Deontic Defeasible Logic to handle the issue of Pragmatic Oddity. The key idea is that a conjunctive obligation is allowed only when each individual obligation is independent of the violation of the other obligations. The solution makes essential use of the constructive proof theory of the logic.

Keywords. Pragmatic Oddity, Defeasible Deontic Logic

1. Introduction

The problem of Pragmatic Oddity, one of the issues related to the formal treatment of the so called contrary-to-duty obligations, introduced by Prakken and Sergot [10], is illustrated by the scenario that when you make a promise, you have to keep. But if you do not, then you have to apologise. The oddity is that when you fail to keep your promise, you have the obligation to keep the promise and the obligation to apologise. In our view, what is odd, is not that the two obligations are in force at the same time, but that if one admits for form a conjunctive obligation from the two individual obligations then we get an obligation that is impossible to comply with. In the scenario, when the promise is broken, we have the conjunctive obligation of keeping the promise and to apologise from not having kept the promise. The Pragmatic Oddity problem arises when we have a conjunctive obligation, i.e., $\mathsf{O}(a \wedge b)$ derived from the two individual obligations ($\mathsf{O}a$ and $\mathsf{O}b$) where one of the conjuncts is contrary-to-duty obligations triggered by the violation of the other individual obligation, for example when $\neg a$ entails that $\mathsf{O}b$ is in force.

Most of the work on Pragmatic Oddity (e.g., [10, 3]) focuses on the issue of how to distinguish the mechanisms leading to the derivation of the two individual obligations, and create different classes of obligations. Consequently, the solution to the Pragmatic Oddity problem is to prevent the conjunction when the obligations are from different classes. Accordingly, if the problem is to prevent to have a conjunctive obligation in force when the individual obligations are in force themselves, the simplest solution is to have a deontic logic that does not support the aggregation axiom[2]:

$$(\mathsf{O}a \wedge \mathsf{O}b) \rightarrow \mathsf{O}(a \wedge b)$$

[1]Corresponding author: Guido Governatori, e-mail: guido.governatori@data61.csiro.au
[2]See, among others, [4].

However, a less drastic solution, advocated by Parent and van der Torre [8, 9], is to restrict the aggregation axiom to independent obligations (meaning that one obligation should not depend on the violation of the other obligation).

We are going to take Parent and van der Torre's suggestion and propose a simple mechanism in Defeasible Deontic Logic to guard the derivation of conjunctive obligations. The mechanism guarantees that the obligations in a conjunctive obligation are independent of the violations of the individual obligations. The mechanism is founded on the proof theory of the logic.

2. Defeasible Deontic Logic

Defeasible Deontic Logic [5] (DDL) is a sceptical computationally oriented rule-based formalism designed for the representation of norms. The logic extends Defeasible Logic [1] with deontic operators to model obligations and (different types of) permissions and provides an integration with the logic of violation proposed in [7]. The logic is based on a constructive proof theory that allows for full traceability of the conclusions In the rest of this section we are going to show how the proof theory can be use to propose a simple and (arguably) elegant treatment of the issue of Pragmatic Oddity. To this aim, here, we restrict ourselves to the fragment of DDL that excludes permission and permissive rules, since they do not affect the way we handle Pragmatic Oddity: Definition 10 describing the mechanisms for Pragmatic Oddity, is independent from any issue related to permission, and can be used directly in the full version of the logic. Accordingly, We consider a logic whose language is defined as follows.

Definition 1. Let PROP be a set of propositional atoms, O the modal operator for obligation. The set Lit = PROP $\cup \{\neg p \mid p \in \text{PROP}\}$ denotes the set of *literals*. The *complement* of a literal q is denoted by $\sim q$; if q is a positive literal p, then $\sim q$ is $\neg p$, and if q is a negative literal $\neg p$, then $\sim q$ is p. The set of *deontic literals* is DLit = $\{\mathsf{O}l, \neg\mathsf{O}l \mid l \in \text{Lit}\}$. If $c_1, \ldots, c_n \in$ Lit, then $\mathsf{O}(c_1 \wedge \cdots \wedge c_n)$ is a *conjunctive obligation*.

We introduce the compensation operator $\otimes$. This operator builds chains of compensation called $\otimes$-expressions, where an $\otimes$-expression is a sequence of one or more literals concatenated by the $\otimes$ operator. In addition we stipulate that $\otimes$ obeys the following property (duplication and contraction on the right):

$$\bigotimes_{i=1}^{n} a_i = \left(\bigotimes_{i=1}^{k-1} a_i\right) \otimes \left(\bigotimes_{i=k+1}^{n} a_i\right)$$

where there exists j such that $a_j = a_k$ and $j < k$.

Given an $\otimes$-expression A, the *length* of A is the number of literals in it. Given an $\otimes$-expression $A \otimes b \otimes C$ (where A and C can be empty), the *index* of b is the length of $A \otimes b$. We also say that b appears at index n in $A \otimes b$ if the length of $A \otimes b$ is n.

The meaning of a compensation chain $c_1 \otimes c_2 \otimes \cdots \otimes c_n$ is that $\mathsf{O}c_1$ is the primary obligation, and when violated (i.e., $\neg c_1$ holds), then $\mathsf{O}c_2$ is in force and it compensates for the violation of the obligation of c_1. Moreover, when $\mathsf{O}c_2$ is violated, then $\mathsf{O}c_3$ is in force, and so on until we reach the end of the chain when a violation of the last element is a non-compensable violation where the norm corresponding to the rule in which the chain appears is not complied with.

We adopt the standard DL definitions of *strict rules*, *defeasible rules*, and *defeaters* [1]. However, for the sake of simplicity, and to better focus on the non-monotonic aspects that DDL offers, in the remainder, we use only defeasible rules and defeaters.

Definition 2. Let Lab be a set of arbitrary labels. Every rule is of the type $r\colon A(r) \hookrightarrow C(r)$ where $r \in$ Lab is the name of the rule; $A(r) = \{a_1, \ldots, a_n\}$, the *antecedent* (or *body*) of the rule, is the set of the premises of the rule (alternatively, it can be understood as the conjunction of all the elements in it). Each a_i is either a literal, a deontic literal or a conjunctive obligation; $\hookrightarrow \in \{\Rightarrow, \Rightarrow_{\mathsf{O}}, \rightsquigarrow, \rightsquigarrow_{\mathsf{O}}\}$ denotes the type of the rule. If $\hookrightarrow$ is $\Rightarrow$, the rule is a *defeasible rule*, while if $\hookrightarrow$ is $\rightsquigarrow$, the rule is a *defeater*. Rules without the subscript O are constitutive rules, while rules with such a subscript are prescriptive rules, and in case the rule is defeasible, the conclusion derived from the rule is an obligation. $C(r)$ is the *consequent* (or *head*) of the rule is a single literal for defeaters and constitutive rules, and an $\otimes$-expressions for prescriptive defeasible rules.

Given a set of rules R, we use the following abbreviations for specific subsets of rules: R_{def} denotes the set of all defeaters in the set R; $R[q, n]$ is the set of rules where q appears at index n in the consequent. The set of (defeasible) rules where q appears at any index n is denoted by $R[q]$; R^{O} denotes the set of all rules in R with O as their subscript. $R^{\mathsf{O}}[q, n]$ is the set of (defeasible) prescriptive rules where q appears at index n. The set of (defeasible) prescriptive rules where q appears at any index n is denoted by $R^{\mathsf{O}}[q]$;

Definition 3. A *Defeasible Theory* is a structure $D = (F, R, >)$, where F, the set of facts, is a set of literals and modal literals, R is a set of rules and $>$, the superiority relation, is a binary relation over R.

A theory corresponds to a normative system, i.e., a set of norms, where every norm is modelled by some rules. The superiority relation is used for conflicting rules, i.e., rules whose conclusions are complementary literals, in case both rules fire. Namely, the superiority just determines the relative strength between two rules.

Definition 4. A *proof* P in a defeasible theory D is a linear sequence $P(1) \ldots P(n)$ of *tagged literals* in the form of $+\partial$, $-\partial$, $+\partial_{\mathsf{O}}q$ and $-\partial_{\mathsf{O}}q$, where $P(1) \ldots P(n)$ satisfy the proof conditions given in Definitions 8–10.

The tagged literal $+\partial q$ means that q is *defeasibly provable* as an institutional statement, or in other terms, that q holds in the normative system encoded by the theory. The tagged literal $-\partial q$ means that q is *defeasibly refuted* by the normative system. Similarly, the tagged literal $+\partial_{\mathsf{O}}q$ means that q is *defeasibly provable* in D as an obligation, while $-\partial_{\mathsf{O}}q$ means that q is *defeasibly refuted* as an obligation. The initial part of length i of a proof P is denoted by $P(1..i)$.

A rule is *applicable* for a literal q if q occurs in the head of the rule have already been proved with the appropriate mode. On the other hand, a rule is *discarded* if at least one of the literals in the antecedent has not been proved. However, as literal q might not appear as the first element in an $\otimes$-expression in the head of the rule, some additional conditions on the consequent of rules must be satisfied. Defining when a rule is applicable or discarded is essential to characterise the notion of provability for constitutive rules and then for obligations ($\pm\partial_{\mathsf{O}}$).

Definition 5. A rule $r \in R[q, j]$ is *body-applicable* iff for all $a_i \in A(r)$:

1. if $a_i = \mathsf{O}l$ then $+\partial_{\mathsf{O}}l \in P(1..n)$;
2. if $a_i = \neg\mathsf{O}l$ then $-\partial_{\mathsf{O}}l \in P(1..n)$;
3. if $a_i = \mathsf{O}(c_1 \wedge \cdots \wedge c_m)$ then $+\partial_{\mathsf{O}}c_1 \wedge \cdots \wedge c_m \in P(1..n)$;

4. if $a_i = l \in \text{Lit}$ then $+\partial l \in P(1..n)$.

A rule $r \in R[q, j]$ is *body-discarded* iff $\exists a_i \in A(r)$ such that

1. if $a_i = \mathsf{O}l$ then $-\partial_{\mathsf{O}} l \in P(1..n)$;
2. if $a_i = \neg\mathsf{O}l$ then $+\partial_{\mathsf{O}} l \in P(1..n)$;
3. if $a_i = \mathsf{O}(c_1 \wedge \cdots \wedge c_m)$ then $-\partial_{\mathsf{O}} c_1 \wedge \cdots \wedge c_m \in P(1..n)$;
4. if $a_i = l \in \text{Lit}$ then $-\partial l \in P(1..n)$.

Definition 6. A rule $r \in R^{\mathsf{O}}[q, j]$ such that $C(r) = c_1 \otimes \cdots \otimes c_n$ is *applicable* for literal q at index j, with $1 \leq j < n$, in the condition for $\pm\partial_{\mathsf{O}}$ iff

1. r is body-applicable; and
2. for all $c_k \in C(r)$, $1 \leq k < j$, $+\partial_{\mathsf{O}} c_k \in P(1..n)$ and $+\partial{\sim}c_k \in P(1..n)$.

Conditions (1) represents the requirements on the antecedent stated in Definition 5; condition (2) on the head of the rule states that each element c_k prior to q must be derived as an obligation, and a violation of such obligation has occurred.

Definition 7. A rule $r \in R[q, j]$ such that $C(r) = c_1 \otimes \cdots \otimes c_n$ is *discarded* for literal q at index j, with $1 \leq j \leq n$ in the condition for $\pm\partial_{\mathsf{O}}$

1. r is body–discarded; or
2. there exists $c_k \in C(r)$, $1 \leq k < l$, such that either $-\partial_{\mathsf{O}} c_k \in P(1..n)$ or $+\partial c_k \in P(1..n)$.

In this case, condition (2) ensures that an obligation prior to q in the chain is not in force or has already been fulfilled (thus, no reparation is required).

For space reasons we only provide the proof conditions for the positive tags. The definitions of the negative tags can be obtained from the definition of the corresponding positive tag by apply the principle of strong negation (that transform the Boolean operators and quantifiers in their dual, and swapping "applicable" and "discarded" [2, 6]. We now introduce the proof conditions for ∂ and ∂_{O}.

Definition 8 (Defeasible provability for an institutional statement)**.**
$+\partial$: If $P(n+1) = +\partial q$ then
(1) $q \in F$ or
(2.1) ${\sim}q \notin F$ and
(2.2) $\exists r \in R[q]$ such that r is applicable, and
(2.3) $\forall s \in R[{\sim}q]$, either
(2.3.1) s is discarded, or either
(2.3.2) $\exists t \in R[q]$ such that t is applicable and $t > s$.

The proof conditions for $\pm\partial$ are the standard conditions in defeasible logic, see [1] for the full explanations.

Definition 9 (Defeasible provability for an obligation)**.**
$+\partial_{\mathsf{O}}$: If $P(n+1) = +\partial_{\mathsf{O}} q$ then
(1) $\mathsf{O}q \in F$ or
(2.1) $\mathsf{O}{\sim}q \notin F$ and $\neg\mathsf{O}q \notin F$ and
(2.2) $\exists r \in R^{\mathsf{O}}[q, i]$ such that r is applicable for q, and
(2.3) $\forall s \in R^{\mathsf{O}}[{\sim}q, j]$, either
(2.3.1) s is discarded, or either
(2.3.2) $s \in R^{\mathsf{O}}$ and $\exists t \in R^{\mathsf{O}}[q, k]$ such that t is applicable for q and $t > s$.

To show that q is defeasibly provable as an obligation, there are two ways: (1) the obligation of q is a fact, or (2) q must be derived by the rules of the theory. In the second case, three conditions must hold: (2.1) q does not appear as not obligatory as a fact, and $\sim q$ is not provable as an obligation using the set of deontic facts at hand; (2.2) there must be a rule introducing the obligation for q which can apply; (2.3) every rule s for $\sim q$ is either discarded or defeated by a stronger rule for q.

We are now ready to provide the proof condition under which a conjunctive obligation can be derived. The condition essentially combines two aspects: the first that a conjunction holds when all the conjuncts hold (individually). The second aspect is to ensure that the derivation of one of the individual obligations does not depend on the violation of the other conjunct. To achieve this, we determine the line of the proof when the obligation appears, and then we check that the negation of the other elements of the conjunction does not occur in the previous derivation steps.

Definition 10 (Defeasible provability for a conjunctive obligation)**.**
If $P(n+1) = +\partial_{\mathsf{O}} c_1 \wedge \cdots \wedge c_m$, then
$\forall c_i, 1 \leq i \leq m$,
(1) $+\partial_{\mathsf{O}} c_i \in P(1..n)$ and
(2) if $P(k) = +\partial_{\mathsf{O}} c_1 \wedge \cdots \wedge c_m, k \leq n$, then $\forall c_j, 1 \leq j \leq m, c_j \neq c_i, +\partial \sim c_j \notin P(1..k)$.

Again, the proof condition to refute a conjunctive obligation is obtained by strong negation from the condition to defeasibly derive a conjunctive obligation.

In what follows we use $\cdots \Rightarrow c$ to refer to an applicable rule for c where we assume that the elements are not related (directly or indirectly) to the other literals used in the examples.

Compensatory Obligations The first case we want to discuss is when the conjunctive obligation corresponding to the Pragmatic Oddity has as conjuncts an obligation and its compensation. This scenario is illustrated by the rule:

$$\cdots \Rightarrow_{\mathsf{O}} a \otimes b$$

In this case, it is clear that we cannot derive the conjunctive obligation of a and b, since the proof condition that allows us to derive $+\partial_{\mathsf{O}} b$ explicitly requires that $+\partial \sim a$ has been already derived (condition 2 of Definition 6). In this case, it is impossible to have the obligation of b without the violation of the obligation of a.

Contrary-to-duty The second case is when we have a CTD. The classical representation of a CTD is given by the following two rules:

$$\cdots \Rightarrow_{\mathsf{O}} a \qquad\qquad \neg a \Rightarrow_{\mathsf{O}} b$$

In this case, it is possible to have situations when the obligation of b is in force without having a violation of the obligation of a, namely, when a is not obligatory. However, as soon as we have $\mathsf{O}a$, we need to derive $\neg a$ to trigger the derivation of $\mathsf{O}b$ (Definition 5).

Pragmatic Oddity via Intermediate Concepts The situations in the previous two cases can be easily detected by a simple inspection of the rules involved; there could be more complicated cases. Specifically, when the second conjunct does not immediately depends on the first conjunct, but it depends through a reasoning chain. The simplest structure for this case is illustrated by the following three rules:

$$\cdots \Rightarrow_{\mathsf{O}} a \qquad \neg a \Rightarrow b \qquad b \Rightarrow_{\mathsf{O}} c$$

To derive $\mathsf{O}c$, we need to prove b. To prove b we require that $\neg a$ has already been proved.

Pragmatic Un-pragmatic Oddity What about when there are multiple norms both prescribing the contrary-to-duty obligation, and at least one of the norms is not related to the violation of the primary norm?

$$r_1 \colon \cdots \Rightarrow_{\mathsf{O}} a \otimes b \qquad r_2 \colon \cdots \Rightarrow_{\mathsf{O}} b \qquad \neg a$$

In this situation you can have the following two proofs:

(1)	$+\partial\neg a$	fact	(1)	$+\partial_{\mathsf{O}} a$	from r_1
(2)	$+\partial_{\mathsf{O}} a$	from r_1	(2)	$+\partial_{\mathsf{O}} b$	from r_2
(2)	$+\partial_{\mathsf{O}} b$	from r_1 and (1) and (2)	(3)	$+\partial\neg a$	fact
			(4)	$+\partial_{\mathsf{O}} a \wedge b$	from (1) and (2)

In the proof on the left $\mathsf{O}b$ ($+\partial_{\mathsf{O}} b$) depends on the violation of the primary obligation of r_1. In this case, we cannot derive the conjunctive obligation of a and b. However, in the other proof, that demonstrates the independence of $\mathsf{O}b$ from $\neg a$, given that the derivation of $\neg a$ occurs in a line after the line where $+\partial_{\mathsf{O}} b$ is derived.

3. Summary

We have proposed an extension of Defeasible Deontic Logic able to handle the so called Pragmatic Oddity paradox. The mechanism we used to achieve this result was to provide a schema that allows us to give a guard to the derivation of conjunctive obligations ensuring that each individual obligation does not depend on the violation of the other obligation. The mechanism is given by the proof theory of defeasible logic. The next steps are (1) to study the complexity of the approach and to verify that the logic obtained is still computationally feasible (a prima facie analysis, based on the structure of the proof conditions for conjunctive obligations, seems to suggest the complexity to be quadratic and then still feasible, and mostly practical for real life applications, where it is unlikely to have many conjunction obligations, and they have a small number of conjuncts); (2) to devise efficient algorithms to implement the novel proof conditions.

References

[1] G. Antoniou, D. Billington, G. Governatori, and M.J. Maher. Representation Results for Defeasible Logic. *ACM Transactions on Computational Logic*, 2:255–287, 2001.

[2] G. Antoniou, D. Billington, G. Governatori, M.J. Maher, and A. Rock. A Family of Defeasible Reasoning Logics and its Implementation. In W. Horn, editor. *ECAI 2000*, pages 459–463. IOS Press, 2000.

[3] J. Carmo and A.J. Jones. Deontic logic and contrary-to-duties. In, *Handbook of philosophical logic*, pages 265–343. Springer, 2002.

[4] L. Goble. A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483, 2005.

[5] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco. Computing Strong and Weak Permissions in Defeasible Logic. *Journal of Philosophical Logic*, 42:799–829, 2013.

[6] G. Governatori, V. Padmanabhan, A. Rotolo, and A. Sattar. A Defeasible Logic for Modelling Policy-based Intentions and Motivational Attitudes. *Logic Journal of the IGPL*, 17:227–265, 2009.

[7] G. Governatori and A. Rotolo. Logic of Violations: A Gentzen System for Reasoning with Contrary-To-Duty Obligations. *Australasian Journal of Logic*, 4:193–215, 2006.

[8] X. Parent and L. van der Torre. "Sing and Dance!" In F. Cariani, D. Grossi, J. Meheus, and X. Parent, editors. *Deontic Logic and Normative Systems*, pages 149–165. Springer International Publishing, 2014.

[9] X. Parent and L. van der Torre. The pragmatic oddity in norm-based deontic logics. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pages 169–178.

[10] H. Prakken and M.J. Sergot. Contrary-to-Duty Obligations. *Studia Logica*, 57:91–115, 1996.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190324

Frequent Use Cases Extraction from Legal Texts in the Data Protection Domain

Valentina LEONE [1] and Luigi DI CARO
Computer Science Department, University of Turin, Italy

Abstract. Because of the recent entry into force of the General Data Protection Regulation (GDPR), a growing of documents issued by the European Union institutions and authorities often mention and discuss various use cases to be handled to comply with GDPR principles. This contribution addresses the problem of extracting recurrent use cases from legal documents belonging to the data protection domain by exploiting existing Ontology Design Patterns (ODPs). An analysis of ODPs that could be looked for inside data protection related documents is provided. Moreover, a first insight on how Natural Language Processing techniques could be exploited to identify recurrent ODPs from legal texts is presented. Thus, the proposed approach aims to identify standard use cases in the data protection field at EU level to promote the reuse of existing formalisations of knowledge.

Keywords. legal ontologies, ontology design patterns, NLP for legal texts

1. Introduction

Written documents are produced in every legal domain in order to spread the law. In the data protection domain, because of the entry into force of the General Data Protection Regulation (GDPR) on May 25th 2018, the debate about how to guarantee the protection of personal data has acquired a pivotal focus. The GDPR sets several measures and practises that different stakeholders dealing with the processing of personal data should adopt to protect data subject's rights and achieve a full compliance with the Regulation. These obligations and rules represent a set of use cases to be properly handled.

The need for the involved actors to comply with the new principles prescribed by the GDPR encouraged the modelling of computational models to support the automatic compliance checking. GDPRov [1], GDPRtEXT [2] and PrOnto [3] ontologies are the main examples of this effort. However, despite these resources model similar use cases, each of them adopts its own ontological commitment, i.e. its own perspective about the data protection domain. These different perspectives bring to ontological representations that, despite being characterised by some distinctive representational choices, share some similarities in the way in which they model the knowledge related to the field of interest.

The problem of redundant representations of knowledge clashes with the principles of reuse and economy of information promoted by the Linked Data [4] in the Semantic

[1]Corresponding Author: Valentina Leone. E-mail: leone@di.unito.it.
The work was partially supported by EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 (MIREL).

Web context. Following this trend, Ontology Design Patterns (ODPs) were proposed as modelling solutions to solve recurrent ontology design problems [5].

In light of those considerations, this contribution addresses the problem of identifying, inside legal texts related to the data protection domain, the use cases for which a standardised modelling solution is already provided by an existing ODP. The approach relies on Natural Language Processing (NLP) techniques to automatically extract evidences of those patterns inside a corpus legal documents.

The paper is organised as follows: Section 2 presents some related works, Section 3 provides an overview of the ODPs that were selected to represent the data protection domain, Section 4 describes a preliminary experiment aimed at extracting one of the selected ODPs from legal documents through NLP, Section 5 ends the paper with the conclusion and the future work.

2. Related work

Legal ontologies in the data protection field. The Data Protection ontology[2] [6] was the first effort to provide a representation of the data protection domain including GDPR related concepts. More recently, GDPRov[3] [1] described the provenance of consent and the data life-cycle modelling abstract workflows to depict how consent and data are collected, used, stored, deleted and shared. GDPRtEXT[4] (GDPR text EXTension) [2] represents the relevant concepts expressed by the GDPR linking them to the parts of the Regulation containing the corresponding definitions. Finally, PrOnto (Privacy Ontology) [3,7] groups the concepts it represents in six macro-classes (i.e., personal data, rights and obligations, processing operations, roles, legal bases, purposes) and aims to provide a model on which approaches of legal reasoning and compliance checking can be applied.

Ontology Design Patterns. Ontology Design Patterns (ODPs) are small ontologies modelled as reusable components that provide a standardised representation of recurrent ontology design problems [5]. This definition implies the presence of use cases which occur frequently inside the domain of interest to be formally represented. A use case is usually expressed by formulating some competency questions for which the proposed ODP should be able to provide a modelling solution, making clear which are the involved entities and the interactions among them. Over the years, the Ontology Design Patterns Portal[5] [8] collected several contributions aimed to provide standardised solutions to different use cases, thus becoming the main reference on the Web for disclosing new ODPs.

Open Information Extraction. Open Information Extraction (OIE) [9] focuses on the extraction of <subject, predicate, object> triples from unstructured texts. Reverb [10] and DefIE [11] are some of the main contributions to OIE, the former adopting syntactical constraints, the latter applying a Word Sense Disambiguation step in order to filter out uninformative relations. Other approaches to OIE, such as KrankeN [12] and ClausIE [13] focus on the extraction of N-ary relations to address the loss of information resulting from limiting the extraction of triples to those identifying binary relations.

[2] http://bit.ly/2uhumDv
[3] https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology
[4] http://bit.ly/2xwjTZJ
[5] http://ontologydesignpatterns.org

Table 1. The list of CPs that were selected from the Ontology Design Pattern Portal and that model use cases of interest in the data protection domain.

Acting For	Action	Activity Specification
Agent Role	Complaint Design Pattern	Communication Event
Information Realization	Object Role	Part Of
Participation	Periodic Interval	Privacy Policy Personal Data
Task Execution	Time Indexed Participation	Time Indexed Person Role
Time Indexed Part Of	Time Indexed Situation	Time Interval
Time Period		

3. ODPs for the legal domain

A preliminary analysis of the Ontology Design Patterns Portal was performed in order to select candidate ODPs modelling use cases that could be possibly find in the data protection domain. In particular, the analysis focused on content design patterns (CPs) listed in the dedicated Web page[6]. CPs differ form other ODPs because the solutions they propose focus on the modelling of classes and properties of a domain, instead of providing domain-independent solutions more focused on solving design expressivity problems [14,15] .

The portal does not set constraints to the type of CPs that can be submitted, allowing to insert both patterns referring to a specific domain as well as patterns modelling general cross-domains use cases. A list of domains that can be associated to the CPs is provided by the portal and each pattern usually states the name of one or more domains it refers to. The selection of the CPs of interest, out of the 157 patterns listed in the portal, was performed analysing the competency questions associated to each pattern and evaluating its suitability for the data protection domain. As this domain is a multidisciplinary field that involves also the management of workflows, the scheduling of tasks and the handling of some events, the selected CPs do not only belong to the law field, but also to other different related domains (e.g. Management, Scheduling, Organization and Event Processing). Moreover, several patterns belonging to the General domain (i.e. patterns not specialised or limited to a range of subjects) were included. Table 1 shows the list of patterns that were selected after this analysis.

Among the selected patterns, only two of them are strictly related to the legal domain, i.e. the Complaint Design Pattern[7] [16] and the Privacy Policy Personal Data pattern[8] [17] . While the former allows the modelling of the different constituents found commonly in a complaint, the latter allows the representation of the information contained into a privacy policy describing how the personal data are processed.

Different groups of CPs can be identified considering the similarities holding among the use cases they model. For instance, some of the CPs focus on the modelling of a situation in which an agent (intended as a human being) is involved. By contrast, other CPs try to represent actions and events that require the modelling of temporal parameters. Table 2 shows a possible organisation of the CPs of interest according to different criteria.

[6] http://ontologydesignpatterns.org/wiki/Submissions:ContentOPs
[7] http://ontologydesignpatterns.org/wiki/Submissions:Complaint_Design_Pattern
[8] http://ontologydesignpatterns.org/wiki/Submissions:PrivacyPolicyPersonalData

Table 2. A list of CPs representing agents involved in some situation (left), a list of CPs representing actions and events involving the modelling of temporal aspects (centre) and a list of CPs related to the law field (right). Some of the CPs could appear in more than one column.

Agents	**Actions and events**	**Law field**
Acting For	Activity Specification	Complaint Design Pattern
Agent Role	Action	Privacy Policy Personal Data
Complaint Design Pattern	Communication Event	
Part Of	Participation	
Participation	Time Indexed Participation	
Privacy Policy Personal Data	Time Indexed Situation	
Time Indexed Participation	Task Execution	
Time Indexed Person Role	Time Indexed Person Role	

4. Finding use cases inside privacy policies

A preliminary study on the retrieval of evidences of the selected CPs inside a corpus of domain-related legal texts was performed. The study focused on a single CP, i.e. the aforementioned Privacy Policy Personal Data pattern[8]. Some evidences of it were looked for inside a small corpus of twelve privacy policies addressed to EU citizens and released after the entry into force of the GDPR. The assumption underlying the experiment is that, if an ODP should represent a recurrent ontology design problem, then evidences of this recurrence could be retrieved in the texts belonging to the domain of interest modelled by the pattern.

To verify this assumption, the text of each privacy policy was manually segmented identifying in it the paragraphs whose content was related to the semantic areas represented in the pattern. As not all the semantic areas that are relevant in a privacy policy are represented by the CP (e.g., it does not model the data subject's rights), only the paragraphs relevant for the pattern were selected. In particular, the semantic areas that were identified in it are: *(i)* types of personal data collected by the company and provided by the data subject, *(ii)* types of personal data collected by the company and provided by third parties, *(iii)* type of processing performed on personal data, *(iv)* third parties the personal data are shared with, *(v)* personal data retention period, *(vi)* lawful basis for processing. The paragraphs of the twelve privacy policies were then grouped according to the semantic area they refer to.

To automatically discover evidences of the selected CP, the ClausIE tool was applied on the paragraphs collected for each semantic area. The extracted triples were then filtered, considering those labelled by ClausIE with the label SVO, i.e. triples containing a subject (S), a verb (V) and an object (O). Finally, those triples were ordered according to the frequency they appear in the paragraphs belonging to the same semantic area. Table 3 shows the top-5 most frequent triples for each identified semantic area.

The obtained triples showed promising results for all the semantic areas. Triples that could be considered as *markers* of the presence of a relevant information to be mapped on some class of the pattern were extracted with high frequency. For instance, considering the table referring to the semantic area *(i)* (i.e., types of personal data collected from the data subject) the high frequency of the triple <we, collect, information> in the corresponding privacy policies paragraphs could be considered as an evidence of the presence in a sentence of a list of types of personal data that the company collects. Indeed, the

Table 3. Most frequent triples extracted by ClausIE and related to the six semantic area listed in Section 4. Triples in bold are the most relevant for the corresponding semantic area.

triples for semantic area *(i)*	**freq.**	**triples for semantic area *(ii)***	**freq.**
<**we, collect, information**>	**87**	<**we, receive, information**>	**42**
<your, “has”, information>	42	<we, collect, information>	30
<**we, collect, data**>	**31**	<our, “has”, games>	24
<your, “has”, device>	29	<your, “has”, information>	23
<our, “has”, website>	28	<we, collect, data>	23
triples for semantic area *(iii)*	**freq.**	**triples for semantic area *(iv)***	**freq.**
<your, “has”, information>	83	<your, “has”, information>	78
<**we, use, information**>	**58**	<**we, share, information**>	**76**
<your, “has”, data>	36	<your, “has”, data>	44
<our, “has”, information>	30	<your, “has”, name>	31
<your, “has”, consent>	29	<**we, share, data**>	**30**
triples for semantic area *(v)*	**freq.**	**triples for semantic area *(vi)***	**freq.**
<your, “has”, information>	41	<your, “has”, information>	25
<**we, retain, information**>	**27**	<**your, “has”, consent**>	**19**
<our, “has”, information>	19	<**we, process, information**>	**8**
<your, “has”, account>	16	<your, “has”, data>	7
<we, share, information>	14	<our, “has”, right>	6

privacy policies usually contain sentences like *we collect information that identifies your mobile device*. For this sentence, ClausIE extracts the following triples: <we, collect, information> and <your, “has”, device>, where the second triple is automatically inferred when the verb *to have* is preceded by a personal adjective. Thus, by analysing the frequency of each triple as well as its co-occurrence with other related triples, it could be possible to evaluate which are the concepts and the properties that a CP models and that can be retrieved inside a legal text belonging to the domain of interest. Considering the aforementioned example, each element of the triples could be mapped in some parts of the corresponding CP: the verb *collect* its an evidence for the *DataCollectionStep* class, the *your* adjective (intended as the “you” pronoun) corresponds to the *Agent* class and the *mobile device* noun could be mapped in the *PersonalData* class. Similar mappings could be identified also for the other semantic areas.

5. Conclusion and future work

This paper presents a first insight for the extraction of existing ODPs (specifically, CPs) for the data protection domain. The proposed approach uses OIE techniques to extract evidence of a CP from legal texts, aiming to achieve a fine granularity in the extraction of information. A first experiment tested the retrieval of evidences of a CP inside a small corpus of privacy policies. The next challenges to be addressed will concern the exploitation of the N-ary relations extracted by ClausIE in order to improve the retrieval of evidence of the CPs inside the text. Moreover, the evaluation of the types of legal documents where the evidence of a pattern could be looked for will be crucial for the success of the experiments.

References

[1] Harshvardhan J Pandit and Dave Lewis. "Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies". In: *PrivOn@ ISWC*. 2017.

[2] Harshvardhan J Pandit et al. "GDPRtEXT-GDPR as a linked data resource". In: *European Semantic Web Conference*. Springer. 2018, pp. 481–495.

[3] Monica Palmirani and Guido Governatori. "Modelling Legal Knowledge for GDPR Compliance Checking". In: *JURIX*. 2018, pp. 101–110.

[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far". In: *Semantic services, interoperability and web applications: emerging concepts*. IGI Global. 2011, pp. 205–227.

[5] Aldo Gangemi and Valentina Presutti. "Ontology design patterns". In: *Handbook on ontologies*. Springer. 2009, pp. 221–243.

[6] Cesare Bartolini, Robert Muthuri, and Cristiana Santos. "Using ontologies to model data protection requirements in workflows" . In: *JSAI International Symposium on Artificial Intelligence*. Springer. 2015, pp. 233–248.

[7] Monica Palmirani et al. "PrOnto: Privacy Ontology for Legal Reasoning". In: *International Conference on Electronic Government and the Information Systems Perspective*. Springer. 2018, pp. 139–152.

[8] Valentina Presutti et al. "D2. 5.1: A Library of Ontology Design Patterns: reusable solutions for collaborative design of networked ontologies. NeOn Project Deliverable". 2018.

[9] Michele Banko et al. "Open information extraction from the web". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Morgan Kaufmann Publishers Inc. 2007, pp. 2670–2676.

[10] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Association for Computational Linguistics. 2011, pp. 1535–1545.

[11] Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. "Large-scale information extraction from textual definitions through deep syntactic and semantic analysis". In: *Transactions of the Association for Computational Linguistics*. Vol. 3. 2015, pp. 529–543.

[12] Alan Akbik and Alexander Löser. "Kraken: N-ary facts in open information extraction". In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics. 2012, pp.52–56.

[13] Luciano Del Corro and Rainer Gemulla. "Clausie: clause-based open information extraction". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 355–366.

[14] Aldo Gangemi. "Ontology design patterns for semantic web content". In: *International semantic web conference*. Springer. 2005, pp. 262–276.

[15] Valentina Presutti and Aldo Gangemi. "Content ontology design patterns as practical building blocks for web ontologies". In: *International Conference on Conceptual Modeling*. Springer. 2008, pp. 128–141.

[16] Cristiana Santos et al. "Complaint Ontology Pattern-COP". In: *Advances in Ontology Design and Patterns*. Vol. 32. IOS Press. 2017, pp.69–83.

[17] Harshvardhan J Pandit, Declan O'Sullivan, and Dave Lewis. "An Ontology Design Pattern for Describing Personal Data in Privacy Policies". In: *WOP@ ISWC*. 2018, pp. 29–39.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190325

On the Formal Structure of Rules in Conflict of Laws

Réka MARKOVICH
Computer Science and Communications RU, University of Luxembourg
Department of Logic, Eötvös Loránd University

Abstract. Law has different methods and principles to resolve conflicts between norms, most of these come from Roman Law, they are well-known and much discussed. There is a whole branch of law, though, which is much less discussed while having been created exactly in order to resolve special conflicts: conflict of laws. This system within Private International Law is dedicated to providing metarules in legal situations where more than one national legal systems' rules could be applied: CoL rules indirectly settle the situation by declaring which one's should. The formal representation of how these rules work contributes not only to the modelling of this branch of law but also provides methodologies for concerns arising from other conflicting normative systems, such as ethically sensitive situations where there are multiple stakeholders with different moral backgrounds.

Keywords. Conflict of Laws, Input/Output Logic, normative systems, deontic logic

1. Introduction

According to a paradigmatic—true—story, Sándor Farkas, a young Hungarian, living in his house in the countryside of Hungary, one day decided to move to France in the hope of a better life, leaving his house behind. Living in France met his expectations: he found a girlfriend with whom they moved together and lived so for decades, he adapted himself to the French environment, so after a while he chose the French citizenship over the Hungarian one. One day, though, as all the good things, Sándor's life came to its end, leaving behind a grieving old French lady. After his death, his siblings in Hungary initiated the probate—in the hope of getting the house. The public notary sitting in her countryside office should usually answer the obvious question: How inherits the house? In this case, though, before answering that, another one arose: According to which law should she decide about the inheritance: the Hungarian or the French? The branch of law containing the answer to this latter and the similar questions is called Conflict of Laws (CoL), which is the main part of Private International Law (PIL).[1] CoL concerns international legal disputes involving persons or companies. Instead of directly settling these disputes, the task of CoL is the indirect settlement declaring which law (in terms of legal system) of the possible candidates should be applied. In the case above these were the Hungarian law and the French, but, of course, these candidates depend on the involved parties or

[1]These labels can often be seen used as synonyms, but some other questions, like e.g. the EU rules regarding private law and business law issues are usually also considered as parts of PIL, see for instance Nagy [1].

other factors playing a relevant role in the dispute (e.g. the place of entering into a contract). Each country has its own CoL rules, whenever a process is initiated regarding a legal dispute having this international nature, once the jurisdiction has been clarified, that is, the given authority (public notary, court, etc. depending on the type of the case) has the power to carry out the proceeding, the acting authority has to apply the concerning rules of their law's (lex fori) CoL and *then*—according to what these rules say—apply their own law's *or* a given foreign law's rules in order to *substantially* settle the case.

Using different formalisms to law have been a fruitful research area of applied logics and AI&Law (for overviews see, for instance [2,3,4]). While (or maybe exactly since) PIL is quite peculiar, so far it got little attention from the AI&Law community. Malerba et al. [5] discuss interpretation in this context, Dung and Sartor [6,7] provide a model in modular argumentation putting equal emphasis on the two steps in PIL: (1) the rules and process of distributing the cases between authorities, that is, clarifying jurisdiction, and (2) the rules and process of establishing the applicable law. Dung and Satror don't discuss, though, some special consequences of the specific nature of these rules (e.g. the possibility and rules of the so-called renvoi, see later). The following analysis focuses only on (2) and on this specific nature, discussing only the rules establishing the applicable law and the features which determine the adequate logic for its representation: section 2 will discuss the approach and formalism this analysis relies on, section 3 will present the preliminary formal analysis of the rules of CoL , in section 4 the discussion and further research questions will be presented introducing the relevance of understanding and modelling CoL for handling ethical questions related to AI.

2. Formal Framework: Approach, Language and Semantics

For we are talking about the specific structure of a type of legal norms, we need the explicit representation of norms enabling us to handle normative systems instead of talking about obligations alone. This tradition can be originated from Alchuorrón and Bulygin, just like what the approach's propagator, Makinson [8] calls a liberating effect of leaving the explicit (modal) operator behind when "taking the set of promulgations of a normative code to be made up of purely boolean formulae". The specific framework we are going to rely on is the intput/output (I/O) logic developed by Makinson and van der Torre [9] (and further developed by many), in agreement with Parent and van der Torre [10] that "norm always takes the form of a conditional statement." The feature (and the main objective) of I/O logic as a framework that the core mechanism of its semantics is detachment also makes it adequate for the current purposes: the analysis will discuss what mechanism the CoL rules declare to the judicature regarding applying the law. There are several other advantages of the I/O framework we won't discuss here in detail, so for which the reader is referred to [11,10], this current paper we will rely on only the very basic notions of I/O framework's semantics.

Regarding the language, we will need the following (diverging slightly from what is usually used in I/O logic): norms are ordered pairs of classical propositional logic formulae (a,x) with the intuitive meaning of a conditional norm (or rule), that is, "given a, it ought to be the case that x". In I/O logic, the "antecedent", a, representing some condition or situation (statement of facts) is called the body, while the "consequent", x, representing what the norm tells us to be obligatory, is called the head. Let N be a set of norms, that is, a set of such ordered pairs. We will need to handle sets of sets of norms,

$\mathbb{N}_{NAT} = \{N_1, N_2, ..., N_n\}$, too, representing a given national legal system containing all the codes regulating the different domains. We will talk about domains which are finite sets of statement of facts, that is, formulae: $A = \{a_1, a_2, ..., a_n\}$. Let's see an example: the rule "If a media service provider provides both linear and on-demand services, it shall notify each of its media services separately" is a norm (a,x) from the Act CLXXXV of 2010 on Media Services, N_A, regulating the domain of media A, and this act is an element of the Hungarian law $\mathbb{N}_{HUN}$.

As mentioned above, in the I/O framework the core mechanism of the semantics is detachment and it happens with the *out* operation. We use the notation $out(N,A)$ to denote the output of an input set A under the code N. As Parent and van der Torre [12] put it: "Intuitively, the output of A under N is the set of obligations that apply to a current situation." In this paper, we don't discuss the different types of this operation resulting in different logics (as we won't use them), it will be enough for those readers who are not familiar with the I/O framework if they think of its meaning in terms of the semantical consequence relation in dyadic deontic logic: $x \in out(N,a)$ *iff* $N \models O(x/a)$

3. The Specific Structure of Rules in CoL and Their Formal Reconstruction

The classical conflict resolving methods developed in Roman Law are often discussed, mostly within the question of legal defeasibility but also within legal argumentation, see e.g. [13,14,2]. The use of principles *Lex superior derogat legi inferiori, Lex posterior derogat legi priori*, and *Lex specialis derogat legi generali* can have different legal affect ranging from concerning simply applicability (spec-gen) to rendering the lower one invalid (sup-inf), but they share two important features: (i) aiming at resolving possible conflicts *within a given legal system* and (ii) providing priority standards based on *some features of the sets of norms which is pre-given* and that we can rely on in the case of need. Both the situation in, and the rules of, Conflict of Laws are quite different.

PIL concerns relations across different legal jurisdictions between natural persons, companies, corporations and other legal entities. In these international legal disputes the conflict of laws—as its name shows—arises *between legal systems*, that is, between sets of sets of norms, so we step up a level. The conflict itself is given by the situation that more than one national laws (legal systems) could be applied, and resolving it means deciding which one should. The crucial point is the mutual acknowledgment by each country of all legal systems being equal, that is, there is *no previously given ordering* on the set of all legal systems which we could rely on. Another solution is needed.

The currently widespread spirit of the applied methodology comes from the 19th century legal scholar, Carl von Savigny, whose idea was to find the local seat of a given legal case, that is choosing the applicable law according to which legal system the case has the strongest relationship to [15]. That is, the CoL rules do not settle directly a case by declaring what to do, their task is the indirect settlement through assigning rules: rules providing which law should be applied to the direct settlement. The difference is visible in the structure of these norms and the "normal" substantial legal norms. The substantial legal rules have the following structure: a) the hypothesis sets forth the conditions under which a person should be guided by the given norm and b) the operative part is the disposition indicating the obligation (or permission). While the CoL rules are assigning rules with the structure: a) the hypothesis designates a domain (of cases) b) the operative part is a command to apply the appointed legal system.

What does it mean regarding the formal structure of these norms? While the substantial norms are ordered pairs of formulae (a,x), the CoL rules will be *ordered pairs of sets*: the body is a domain A, that is, a set of statement of facts: $A = a_1, a_2, ..., a_n$; while the head is also a set, a set of sets of norms, that is, a (national) legal system $\mathbb{N}_{nat_1}$. That is, the form of norm in CoL is: $(A, \mathbb{N}_{nat_1})$. These special norms create the statute on CoL, that is the set N_{CoL}: $N_{CoL} = \{(A, \mathbb{N}_{nat_1}), (B, \mathbb{N}_{nat_2}), ...(Z, \mathbb{N}_{nat_n})\}$

And this set of these special norms, that is, the statute on CoL is part of the Hungarian legal system, next to the Act on Media, the Criminal Code, the Act on Tax, etc.:
$N_{CoL} \in \mathbb{N}_{HUN}$ $\mathbb{N}_{HUN} = \{N_A, N_B, ..., N_{CoL}\}$
And this is true not only to the Hungarian law, but to each national legal system, so we are better to indicate that in the Hungarian law you can only find the Hungarian act on media: the French, the Italian, the Chinese and so on have their own—just like they have their own statute on CoL:
$\mathbb{N}_{HUN} = \{N_A^{HUN}, N_B^{HUN}, ..., N_{CoL}^{HUN}\}$ $\mathbb{N}_{FRA} = \{N_A^{FRA}, N_B^{FRA}, ..., N_{CoL}^{FRA}\}$
We still need to show the mechanism of the rules in CoL: How does the appointment of the applicable legal system happen? It happens in the spirit of Savigny's thought: the legislator chooses the legal system which might be the closest to the case providing the most fitting solution to the legal case. What did it mean for the Sándor Farkas case? The Hungarian public notary found in the Hungarian CoL rules the following section: *"The legal relationship of inheritance shall be adjudged on the basis of the law which was the personal law of the testator at the time of his death."*This section (and the sections in general in the statute on CoL) unequivocally assign the legal system that is to be applied to a given group of statements of facts (that is, a domain). In case of Sándor this was the French law, as he had taken (got) the French citizenship while living in France and had it during his life afterwards, therefore that was the personal law of him at the time of his death. The methodology providing the priority standard visible in this rule is based on an ordering too. This ordering is on the set of the *factors* of a type of legal cases (that is, coming from a given domain).

Each statement of facts is a conjunction of sentences: $a \leftrightarrow \varphi \wedge \psi \wedge \chi \wedge ...$ What makes the situations covered by CoL special is that the factors come from different jurisdictions. In the case of Sándor, his original nationality was Hungarian, his citizenship at the time of his death was French, the house was in Hungary, etc. and forming these like sentences in the conjunction will "bear" these different nationalities: $a \leftrightarrow \varphi_{nat1} \wedge \psi_{nat2} \wedge \chi_{nat3} \wedge ...$ In CoL, the legislator creates a partial order with a maximal element on the (finite) set of the conjuncts: $S_a = \{\varphi, \psi, \chi ...\}$ and assigns the relevant national legal system: $(A, \mathbb{N}_{nat\varphi})$

So far so good. The Hungarian public notary learned from the Hungarian CoL rules that she needs to apply the French law. As we have seen above, though, the set called the French law has many elements: statutes, that is sets of norms, and one is among them is the French statute on CoL. As entering the French law, the legal dispute over Sándor Farkas' inheritance is still an international one, therefore, first the CoL rules need to be checked to learn which law should be applied. The French CoL rules said that the legal relationship of inheritance has to be adjudged on the basis of the law which was the personal law of the testator at the time of his death (so far the same as the Hungarian law), although, if there is a real estate in the inheritance, then it should be adjudged on the basis of the law of its location. Which is the Hungarian law! As the reader likely suspects by now, this could lead to an infinite regress. To prevent that, the Hungarian rules on CoL (which back then were codified in a so-called Law-Decree) specified that

once the applicable law is given, the substantial rules of the given legal system should be applied (that is, not its CoL): *"If, in accordance with this Law-Decree, foreign law is applicable, the rules of the applicable foreign law directly settling the issue in question shall govern."* Formally this norm is: $(A, \mathbb{N}_{nat\varphi} \setminus N_{CoL}^{nat\varphi})$
Although, it is always easier to any authority to apply its own law, so we can find a supplement: *"If, however, the foreign law refers back to the Hungarian law in the issue concerned, with regard to this rule, the Hungarian law shall be applicable."* This solution leads to the called renvoi and formally looks like:
If $((A, \mathbb{N}_{HUN}) \in out(N_{CoL}^{nat\varphi}, a))$ then $(A, \mathbb{N}_{HUN} \setminus N_{CoL}^{HUN})$
We need to define what the output is in the case of CoL rules, for which we need to define the output of a national law, that is, set of sets of norms:
$x \in out(\mathbb{N}_{nat1}, a)$ iff $a \in A \wedge N_A^{nat1} \in \mathbb{N}_{nat1} \wedge x \in out(N_A^{nat1}, A)$
$x \in out(N_{CoL}^{nat2}, a)$ iff $a \in A \wedge \mathbb{N}_{nat1} \in out(N_{CoL}^{nat2}, A) \wedge x \in out(\mathbb{N}_{nat1}, a)$

4. Relevance in the Ethics of AI and Further Research

The rules of CoL are quite special, their formal representation requires some modification of what we have used so far to represent norms and normative systems, but the formalism of the input/output framework can be easily adapted to it. The main specificity of these rules is assigning a set of sets of norms to a set of statements, i.e., the applicable legal system to a given domain. What makes the whole methodology of CoL special, compared to other (norm-)conflict resolving methods, is that there is no previously given ordering on the set of the legal systems that we could rely on (as the nations acknowledge each-other's legal systems as equal) and, therefore, the legislator needs to provide a context-dependent solution by appointing the maximal element of the set of the factors, and, by virtue of that, appointing the relevant legal system to apply.

The contribution of this paper is providing an approach and explanatory formal analysis of the specific nature of rules in CoL. There are several tasks and questions to be answered to make it complete: what logical properties do we need? For instance, halting infinite regress might mean that the *out* operation cannot be transitive. Indeed, the renvoi (when the applicable law's CoL "sends back" the case to the forum's law) is not the only case to be handled: it might happen that the applicable law's CoL rules command to apply a third law and so on (called transmission)—in principle it also might mean the possibility of an infinite regress which has to be stopped and the different countries have different solutions to that (not allowing transmission at all, allowing only limited (small) number of steps, etc.). Also, there are various approaches applied in the different national rules on CoL whose interaction provides a fertile ground for formal research; just like the question of characterization of the cases in different legal systems, that is, how the extension of the domains as sets influence the output.

Conflict of Laws, or PIL in general might seem too peculiar—as Dung and Sartor put: exotic—to be dealt with, but it's importance will only continue to raise up until we have different legal systems and substantial differences in their rules. The development of CoL in the previous centuries was motivated by the increasing volume of international traffic of goods and people, a trend that won't go away anytime soon. A proper formalization might help see clearly and solve problems like the so-called forum shopping when the output, that is, the result of the case depends on which jurisdiction one files for action in (what would have happened if grieving girlfriend had initiated the probate in France?).

But a comprehensive formalization will also be interesting for other areas, too: namely, the ethics of AI, one of the most salient topics today. The results of the robust survey of MIT, the Moral Machine [16] has shown: there are no globally accepted, generally valid values or set of rules to rely on when we talk about the ethics of our (soon-to-be-developed) AI tools. However, it is a major concern requiring some solution soon. We might say that an AI tool doesn't have to enhance everyone's ethical considerations, only that of those who are affected by it. But the issues and debates regarding autonomous vehicles bring clear emphasis on that there are multiple stakeholders. Realizing this, Liao et al. [17] developed an architecture, called Jiminy, which is supposed to advise AI tools in ethically sensitive situations, when there are multiple stakeholders with different normative systems expecting to comply with different moral rules. Handling different peer normative systems is exactly what CoL does, therefore, its techniques seeking for context dependent resolution can definitely provide insights this very 21st century problem.

Acknowledgement

The author has received funding from the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agr. No. 690974 (MIREL).

References

[1] C.I. Nagy, *Private International Law. Hungary*, Wolters Kluwer, 2012.
[2] D. Grossi and A. Rotolo, Logic in the law: A concise overview, in: *Logic and Philosophy Today*, 2011.
[3] G. Sartor, *Legal Reasoning. A Treatise of Legal Philosophy and General Jurisprudence*, Springer, 2005.
[4] H. Prakken and T. Bench-Capon, Introducing the logic and law corner, *Journal of Logic and Computation* **18** (2008), 1–12.
[5] A. Malerba, A. Rotolo and G. Governatori, Interpretation Across Legal Systems, in: *Proceedings of JURIX 2016*, pp. 83–92.
[6] P.M. Dung and G. Sartor, A Logical Model of Private International Law, in: *Proceedings of DEON 2010*, Springer, Berlin, 2010, pp. 229–246.
[7] P.M. Dung and G. Sartor, A modular logic of private international law, *Artificial Intelligence and Law* **19**(2–3) (2011), 233–261.
[8] D. Makinson, On a fundamental problem of deontic logic, in: *Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science*, IOS press, 1999, pp. 29–53.
[9] D. Makinson and L. van der Torre, Input/Output Logics, *Journal of Philosophical Logic* **29**(4) (2000), 383–408.
[10] X. Parent and L. van der Torre, Input/output Logic, in: *Handbook of Deontic Logic and Normative Systems*, D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, eds, College Publications, 2013, pp. 353–406.
[11] D. Makinson and L. van der Torre, What is Input/Output Logic?, in: *Foundations of the Formal Sciences II. Trends in Logic*, Vol. 17, B. Löwe, W. Malzkom and T. Räsch, eds, Springer, 2003.
[12] X. Parent and L. van der Torre, *Introduction to Deontic Logic and Normative Systems*, College Publications, 2018.
[13] H. Prakken and G. Sartor, On the Relation Between Legal Language and Legal Argument: Assumptions, Applicability and Dynamic Priorities, *Proceedings of ICAIL '95* (1995).
[14] G. Governatori, F. Olivieri, S. Scannapieco and M. Cristani, Superiority Based Revision of Defeasible Theories, in: *Semantic Web Rules*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 104–118.
[15] F.C. von Savigny, *Private International Law. A Treatise on the Conflict of Laws: And the Limits of Their Operation in Respect of Place and Time*, T. and T. Clark, 1869.
[16] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, F. Bonnefon and I. Rahwan, The Moral Machine experiment, *Nature* **536** (2018), 59.
[17] B. Liao, M. Slavkovik and L. van der Torre, Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders, ACM, 2019.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190326

PrOnto Ontology Refinement Through Open Knowledge Extraction

Monica PALMIRANI[a1], Giorgia BINCOLETTO[a], Valentina LEONE[b], Salvatore SAPIENZA[a] and Francesco SOVRANO[c]
[a] *CIRSFID, University of Bologna*
[b] *Computer Science Department*, *University of Turin*
[c] *DISI, University of Bologna*

Abstract. This paper presents a refinement of PrOnto ontology using a validation test based on legal experts' annotation of privacy policies combined with an Open Knowledge Extraction algorithm. Three iterations were performed, and a final test using new privacy policies. The results are 75% of detection of concepts and relationships in the policy texts and an increase of 29% in the accuracy using the new refined version of PrOnto enriched with SKOSXL lexicon terms and definitions.

Keywords. legal ontology, GDPR, OKE, refinement.

1. Introduction

We have already published several papers about PrOnto ontology [17][18][19][22] that aims to model the concepts and their relationships presented in the GDPR (General Data Protection Regulation EU 2016/679). This article intends to present a validation process of PrOnto ontology using a bottom-up approach, starting from the language adopted in real examples of Privacy Policies. The research investigates: i) if the existing PrOnto classes are sufficiently exhaustive to support NLP tools in detecting GDPR concepts directly from Privacy Policies; ii) if some classes are missing with respect to the pragmatic language forms; iii) if some frequent terminology could be added to the conceptualisation modelling using e.g., SKOSXL; iv) whether it is possible to create a ML tool that is capable of detecting GDPR concepts in the Privacy Policies. The paper first presents the used methodology; secondly, it presents the legal analysis of the Privacy Policies chosen for the validation and the related mapping of the linguistic terminology in the PrOnto classes; then, the work introduces the ML technique applied to detect the PrOnto concepts from the other Privacy Policies and its results; finally, the conclusion shows the refinements made to the PrOnto ontology thanks to the validation with the Privacy Policies.

2. Methodology

PrOnto was developed through an interdisciplinary approach called MeLOn (Methodology for building Legal Ontology) and it is explicitly designed in order to minimise the difficulties encountered by the legal operators during the definition of a

[1] E-mail: {monica.palmirani, salvatore.sapienza2, giorgia.bincoletto2, francesco.sovrano}@unibo.it, leone@di.unito.it

legal ontology. MeLOn applies a top-down methodology on legal sources. It is based on reusing ontology patterns [12] and the results are evaluated using foundational ontology (e.g., DOLCE [8]) and OntoClean [11] method. The validation is made by an interdisciplinary group (engineers, lawyers, linguists, logicians and ontologists) that integrates the contributions of different disciplines. The methodology is based on the following pillars [1][3]: (i) two legal experts selected ten privacy policies from US-based companies providing products and services to European citizens; (ii) the privacy policies were analyzed using the comparative legal method to discover the frequent concepts mentioned in the texts; (iii) selected portions of text were mapped into the PrOnto ontology with also different linguistic variations; (iv) computer science team developed Open Knowledge Extraction technique starting from the GDPR lexicon, PrOnto ontology and the literal form variants (point 3); (v) results were validated by the legal team that returned them to the technical team; (vi) the steps from (ii) to (v) were iterated three times to refine the ontology and the software model; (vii) finally, new privacy policies were selected by the legal experts[2] in order to evaluate the effectiveness of the refined algorithm and ontology.

3. Legal Analysis of the Privacy Policies

We have selected ten Privacy Policies[3] from an equal number of companies in the sector of sale of goods, supply of services and sharing economy. We chose these companies due to their international dimension, their relevance in their market sectors and the diversity of data processing techniques, with European target. We distinguished between the legal strict terminologies (e.g., data subject) to the communicative language (e.g., customer or user). The legal experts have manually reviewed the Privacy Policies to discover the concepts of legal relevance for data protection domain (provisions, legal doctrine, WP29 and case law) that are remarkably recurrent in the text. The interpretation has also kept into account the existing version of PrOnto ontology, in particular to identify the different terms that express the same concept recognised through a legal analysis at an equal level of abstraction. These terms have been analysed, compared and eventually included in the PrOnto ontology, using techniques like SKOSXL for adding the different linguistic forms (e.g., `skosxl:leteralForm`). This extension of PrOnto definitely improves the capacity of the OKE tools to detect the correct fragment of text and to isolate the legal concept as well as populating the PrOnto ontology. We also noted that the Privacy Policies tend to use the ordinary, everyday language for reasons of transparency and comprehensibility of the texts. Despite the advantage for the costumer/user, the analysis underlined that certain terminologies are not accurate from a legal perspective. For instance, the expression "*giving permission*" is a communicative substitute of "*giving consent*" and "*obtain consent*". Some terminologies are misused because the ordinary language in the policy does not reflect the legal sense e.g., "*anonymous data*" (Recital 26 GDPR) is not in the scope of the Regulation and it is misled with "*anonymized data*". We found terminology coming from computer science like "*to hash*", "*log files*", "*use encryption*" convey a technical meaning that is not classified in the GDPR, which is drafted in a technically neutral way.

[2] Rover, Parkclick, Springer, Zalando, Louis Vuitton, Burger King, Microsoft-Skype, Lufthansa, Booking, Zurich Insurance.

[3] Amazon, Dell, McDonald, Nike, American Airlines, TripAdvisor, Hertz, Allianz U.S. AirBnB, Uber.

4. PrOnto Manual Enhancing

Following this analysis, we have mapped the synthesis of the different lexicon expressions with the PrOnto classes. This step allowed to detect some missing modules that are described below. Under the GDPR, personal data processing (Art. 4.1(2)) is lawful only if motivated by a purpose that must be legitimated by a legal basis (Art. 6 GDPR). Therefore, a lawfulness status was thus added as a Boolean data property of the `PersonalDataProcessing` class. However, from the validation using Privacy Policies, it is extremely important to elicit the Legal Basis because several other implications (rights, obligations, actions) depends to the kind of legal basis (e.g., Art. 22). For this reason, we have modelled new module (Fig. 1 new classes are in orange).

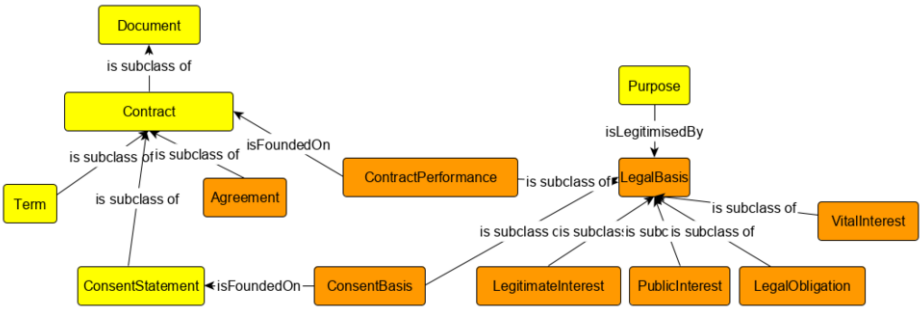


Figure 1 – Legal Basis Module

Archiving and Services are encountered frequently in the Privacy Policies and they are added to the Purpose Module, with also a specific kind of service (`Information-SocietyService`) relevant for the child privacy (Art. 8 GDPR). The Privacy Policies underlined some obligations, and related rights, like the `ObligationToProvide-HumanIntervention` connected with `RightToHaveHumanIntervention` and related with `AutomaticDecisionMaking` that is an action added to the Action module.

5. Open Information Extraction for PrOnto

We built a software for detecting GDPR concepts from Privacy Policies taking inspiration from the PrOnto ontology and using a tool conceptually based on ClausIE [6]. ClausIE is a clause-based approach to Open Information Extraction, which extracts relations and their arguments from natural language text. Open Information Extraction (Open IE) builds information graphs representing natural language text in the form of SVO (Subject, Verb, Object) triples (slightly different from RDF). This method was used in other relevant works in the past and several problems arise: (i) linguistic variants of the same legal concept inside the agreement/contract text are numerous and they include some overlappings of meaning; (ii) while legislative text uses rhetoric sentences, policy text is usually simpler and uses common language to be more understandable; (iii) occasionally, legal provisions are written in passive form in order to emphasize prescriptiveness when addressing the command; (iv) legal text has normative references that affect the knowledge extraction; (v) legal concepts change over time; (vi) frequency is not a good indicator of relevance. The main difference between many classical Open IE techniques and ClausIE is that the latter makes use of the grammatical dependencies extracted through an automatic dependency parser, to identify the SVO triples. ClausIE

is able to identify SVO triples, but we need also to correctly associate them to ontology terms and their literal variants provided by the legal expert team. Let the GDPR and the Privacy Policies be our corpus C. In order to perform the automatic text annotation of our corpus with PrOnto concepts, we follow these steps: 1. we identify a list of all the terms (subjects, objects-classes; verbs-properties) in C, by using a simple variant of ClauseIE; 2. we use PrOnto labels of classes and properties, with additional mapping of linguistic and lexicon variants; 3. we map every possible class/property in C to its closest class/property in PrOnto, using a previous project[4]. This algorithm exploits pre-trained linguistic deep models in order to easily compute a similarity score between two terms.

6. PrOnto Refinement Using OKE

From the Privacy Policies linguistic analysis with OKE, it emerges that some inputs produced important enhancements in PrOnto ontology. **New Child Class**: in the Privacy Policies is frequently mentioned "*child*" that is a particular "data subject" missing in the PrOnto ontology. Initially, we intended to use rules to define child concept because the definition changes for each jurisdiction according to the local implementation of the EU Regulation. However, in light of the important rights and obligations defined in the GDPR for the minors, we decided to include a new class in the `Role` module as subclass of `DataSubject`. `Child` class is related with `ParentalResponsabilityHolder`. **New AnonymisedData Class**: from the Privacy Policies linguistic analysis emerges that "*Anonymised Data*" and "*Anonymous data*" (Recital 26 GDPR)[5] are often misled. The pragmatic language attempts to simplify the legal terminology creating mistake in the conceptualization of those two classes. To stress this distinction, we modelled the relationship `PersonalData isTransformedIn AnonymisedData`.

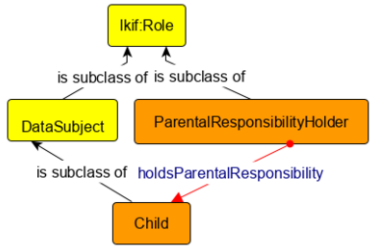


Figure 2: Child class.

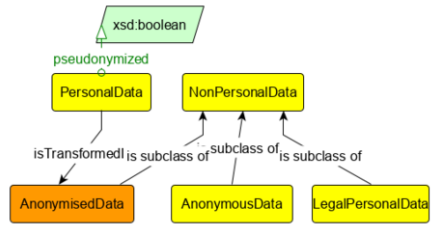


Figure 3: AnonymisedData class.

The best manner to detect an *action* is through verbs. However, within OWL ontology, verbs play the role of predicates that connect domain and range (relationships not classes). For this reason, the legal team modifies the action's classes with the "ing" form according also other scholars [10]. **New Actions** are detected like `Collecting` and `Profiling`. The legal analysis collocates the `Profiling` class as subclass of `AutomatedDecisionMaking` following Art. 22 and the Recital 71. In this case, the OKE provides a very good input to the legal experts that provided an improvement of the legal ontology by relying on their legal analysis. **Lexicon Forms**: it is important to connect the legal concepts to lexicon form variants. We use SKOS and SKOSXL that is

[4] https://gitlab.com/CIRSFID/un-challange-2019.

[5] COM (2019) 250 final anonymised "data which were initially personal data, but were later made anonymous.". Recital 26 GDPR "6. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

a canonical method for connecting OWL and linguistic variants, using `skosxl:literalform`. In this manner, we link PrOnto Core Ontology with other existing lexicon-controlled vocabulary[6].

7. Related Work

Several ontologies model privacy domain. Some of them are oriented to the linguistic tools e.g., UsablePrivacy and PrivOnto [15] to define glossary starting from the bottom-up annotation of the privacy policies (crowdsourcing annotation). GDPRtEXT [20] lists concepts present in the GDPR text without claiming to model norms and legal axioms. GDPRov describes the provenance of the consent and data lifecycle in Linked Open Data [21]. GConsent models the consent action, statement and actors. The SPECIAL Project develops tools for checking compliance in privacy domain. ODRL provides predicates and classes for managing obligations, permission, prohibitions, but not deontic logic operators (e.g., penalty). LegalRuleML [16] ontology was included inside of PrOnto. EUROVOC and IATE are some examples of linguistic ontologies released by the European Union to semantically structure the terminology of documents of the EU institutions [23]. Those resources do not clarify the distinction between legal concepts and their instances and additional knowledge is necessary on legal theory, legal doctrine and legal sociology [7]. Several models propose interfaces between high-level ontological concepts and their low-level, context-dependent lexicalisations [14]. SKOSXL[5] and OntoLex [4] are included in this version of PrOnto for combining ontology and linguistic literal forms, in support to NLP and search engine. Open IE is capable to extract information graphs from natural language. Examples of Open IE tools are ClausIE [6], OpenCeres [13] and Inter-Clause Open IE [1]. Open Knowledge Extraction (Open KE) builds over Open IE to align the identified subject, predicates and objects (SVOs) to pre-defined ontologies. FRED [9] uses different NLP techniques for processing text and for extracting a raw ontology based on VerbNet situations. The challenge of Open KE is that the SVOs alignment requires to understand the meaning of ambiguous and context-dependent terms. Our algorithm tackles the Open KE problem by exploiting pre-trained linguistic deep models to map information to knowledge.

8. Conclusions and Future Work

We have validated the PrOnto ontology with a sample of Privacy Policies and with a legal analysis following the MeLOn methodology, in order to manually check the completeness of the classes and relationships for representing the main content of the policies texts. This exercise detected some new classes in the PrOnto ontology (e.g., Legal Basis). The legal team detected some inconsistency in the terminologies between the legislative text and the pragmatic language. This produced a map of lexicon variants, then modelled using SKOSXL. PrOnto and these extensions fill up an OKE algorithm to detect concepts in the Privacy Policies. The method was iterated three times and at the end we obtained an increase of 29% in the detection of the concepts respect the first interaction that record an increase of 19%. We are capable to detect the 75% of the concept in the new privacy policies using the new version of PrOnto enriched with SKOSXL terms. This method is also relevant to annotate legal texts with PrOnto and so to create RDF triples for supporting applications (e.g., search engine, legal reasoning)[7].

[6] https://www.w3.org/ns/dpv#data-controller.

[7] https://gitlab.com/palmirani/pronto.

Acknowledgements. This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the MSCA grant agreement No 690974 “MIREL: MIning and REasoning with Legal texts”.

References

[1] G. Angeli, M.J.J. Premkumar, C.D. Manning. Leveraging linguistic structure for open domain information extraction. In ACL-IJCNLP **1** (2017), 344–354.
[2] K.D. Ashley, *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge Univ Press, Cambridge New York, 2017.
[3] J. Bandeira, I.I. Bittencourt, P Espinheira, S. Isotani. FOCA: A Methodology for Ontology Evaluation. ArXiv preprint arXiv:1612.00353 (2016).
[4] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda. Towards a module for lexicography in OntoLex. In Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at LDK 2017, Galway. CEUR-WS **1899** (2017), 74–84.
[5] T. Declerck, K. Egorova, E. Schnur. An Integrated Formal Representation for Terminological and Lexical Data included in Classification Schemes. In Proc. of the LREC-2018 (2018).
[6] L. Del Corro, R. Gemulla. Clausie: clause-based open information extraction. In Proc. of the 22nd intern. conference on World Wide Web. ACM (2013), 355-366.
[7] M. Fernández-Barrera, G. Sartor. The legal theory perspective: doctrinal conceptual systems vs. computational ontologies. In Approaches to Legal Ontologies. Springer, Dordrecht (2011), 15–47.
[8] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider. Sweetening Ontologies with DOLCE. In Inter. Conf. on Knowledge Engineering and Knowledge Management Springer, Springer, Berlin, Heidelberg, (2002) 166–181.
[9] A. Gangemi, A. Presutti, V. Reforgiato, D. Recupero, A.G. Nuzzolese, F. Draicchio, M. Mongiovì. Semantic web machine reading with FRED. *Semantic Web*, **8** (2017), 873–893.
[10] A. Gangemi, S. Peroni, D. Shotton, F. Vitali. The Publishing Workflow Ontology (PWO). *Semantic Web* **8** (2017), 703–718.
[11] N. Guarino, C.A. Welty. An Overview of OntoClean. In Handbook on ontologies (2004), 151–171.
[12] P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi (Eds.). *Ontology engineering with ontology design patterns: foundations and applications, Studies on the semantic web*. IOS Press, Amsterdam. 2016
[13] C. Lockard, P. Shiralkar, X. L. Dong. OpenCeres: When Open Information Extraction Meets the Semi-Structured Web. In *NAACL-HLT 2019* **1** (2019), 3047–3056.
[14] J. McCrae, D. Spohr, P. Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In Proc. ESWC 2011. *LNCS* **6643** (2011), 245–259.
[15] A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T.B. Norton, N.C. Russell, P. Story, J. Reidenberg, N. Sadeh. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web* **9** (2018), 185–203.
[16] M. Palmirani, G. Governatori. Modelling Legal Knowledge for GDPR Compliance Checking. In Proc. Jurix 2018. *Frontiers in Artificial Intelligence and Applications* **313** (2018), 101–110.
[17] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo, 2018. PrOnto: Privacy Ontology for Legal Reasoning. In Proc. EGOVIS2018, September 3-5. *LNCS* **11032** (2018), 139–152.
[18] M. Palmirani, M. Martoni, A. Rossi. C. Bartolini, L. Robaldo. Legal Ontology for Modelling GDPR Concepts and Norms. In Proc. JURIX 2018. *Frontiers in Artificial Intelligence and Applications* **313** (2018), 91–100.
[19] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo. PrOnto: Privacy Ontology for Legal Compliance. In Proc. ECDG 2018, ACPI Reading UK (2018), 142–151.
[20] H.J. Pandit, K. Fatema, D. O’Sullivan, D. Lewis. GDPRtEXT - GDPR as a Linked Data Resource. In Proc. ESWC 2018. *LNCS*, **10843** (2018), 481–495.
[21] H.J. Pandit, D. Lewis. Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies. In Proc. of the 5th Workshop PrivOn2017 co-located with ISWC 2017 (2017)
[22] A. Rossi, M. Palmirani, 2019. DaPIS: an Ontology-Based Data Protection Icon Set. *Frontiers in Artificial Intelligence and Applications* **317** (2018), 181–195.
[23] C. Roussey, F. Pinet, M.A. Kang, O. Corcho. An introduction to ontologies and ontology engineering. *Ontologies in Urban development projects* **1** (2011) 9–38.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190327

Towards a Computational Theory of Action, Causation and Power for Normative Reasoning

Giovanni SILENO [a,1], Alexander BOER [b] and Tom VAN ENGERS [b,a]
[a] *Informatics Institute, University of Amsterdam, Netherlands*
[b] *Leibniz Institute, University of Amsterdam/TNO, Netherlands*

Abstract. In order to effectively implement guidance structures in a computational social system, directives which are specified in general terms of duties and rights need to be transformed in terms of powers and liabilities attributed to social parties. The present paper is a work in progress report on an axiomatization of power structures in a logic programming setting, covering the intentional level in specifying actions, the connection between productive characterization of actions and causation, the default nature of action specifications, failures and omissions, the relations of causation and power, and the concept of interfering actions.

Keywords. Models of action, Causation, Power, Ability, Susceptibility, Event Calculus, Reactive rules, Normative Reasoning, Logic Programming, ASP

1. Introduction

For enabling automated normative reasoning, norms need to be represented in a computationally processable way, just as the world on which such norms are deemed to apply. Ideally, two types of normative reasoning can be distinguished: (a) *reasoning with norms*, i.e. applying given norms to qualify behaviour and situations (possibly to take decisions upon); (b) *reasoning about norms*, that can be further inflected in internal view (i.e. check whether a certain norm is *valid* and *applicable* with respect to a given set of norms) and external views to legal systems (i.e. whether the norm is *effective* in guiding behaviour and/or it is *efficient* in terms of costs required for its maintenance). Although the internal/external distinction is mostly evident in (b), a more attentive analysis shows that elements of the second re-enter in (a). More specifically, in order to be applied on a social system, i.e. to effectively implement guidance structures (for instance in a computational social system) *directives which are specified in general terms of duties and rights needs to be reinterpreted in terms of powers and liabilities attributed to social parties.*

There is a long-standing debate between proponents of a purely deontic approach to norms (in legal philosophy see e.g. MacCormick and Raz [1]: "Though powers are essential to the explanation of rights, they are not in themselves rights"), and "paritary" approaches to deontic and potestative categories as the one advanced by Hohfeld [2]. This debate is mirrored in logic and related computational disciplines, although most solutions starts implicitly or explicitly from some flavour of *deontic logic* (see e.g. [3]). Inspired by Hohfeld, a much smaller number of studies focuses on directed obligations and rights

[1]Corresponding Author: g.sileno@uva.nl. This research was partly funded by NWO (VWData project).

including powers (e.g. [4,5,6]). Relevant to this group, although different in spirit, is the contribution by [7], highlighting the *teleological* aspects of normative relations.

The research direction motivating this paper takes a somehow even more radical stance: even acknowledging the primary role of deontic categories to specify optimality (and sub-optimality) in the world, we recognize the crucial role of potestative categories to deal with the implementation of normative mechanisms in the social system, in the attempt to guide it towards this optimality. Indeed, at the level of social system, *everything boils down to power structures, enabling or disabling action* (institutional and physical). Accepting this formulation, it is crucial to be able to model and reason with power, as well as with related concepts, as action and causation. The present paper can be seen as a work in progress report on an axiomatization in a *logic programming* (LP) setting.[2]

2. Representing action

Procedural, productive and intentional characterizations It is generally acknowledged that three general characterizations of actions exist in human language, mapping to three levels of abstraction [8]: **task** (e.g. "Brutus stabbed Ceasar"), **outcome** ("Brutus killed Ceasar"), and **intent** ("Brutus murdered Ceasar").[3] More in detail:[4]

- The *behavioural* or *procedural* characterization relates to the task abstraction level, describing the type of behaviour that the agent has just followed (or is following, if the action is not atomic). We will denote it as `performs(X, A)`, as e.g. in `performs(brutus, stabbing)`.
- The *productive* characterization relates to the outcome level, describing the result that the agent's behaviour has produced. We will denote it as `brings(X, R)` as e.g. in `brings(brutus, dead)`.
- The *purposive* or *intentional* characterization is associated with the intent level, describing the intent which drives the agent's behaviour, whose content can be either procedural `aims(X, A)` as e.g. `aims(brutus, stabbing)` or productive `aims(X, R)`, as e.g. in `aims(brutus, dead)`.

Definition of actions Introducing a general predicate `does` for actions, we can rewrite the variations of the initial example in terms of characterizations:

```
does(brutus, stabbing) <-> performs(brutus, stabbing).
does(brutus, killing) <*> brings(brutus, dead).
does(brutus, murdering) <->  aims(brutus, killing), does(brutus, killing)
```

where `<->` indicates logical equivalence, and `<*>` stands for a *default* inference mechanism that will be investigated further in the text. At second sight, we note that the task-form (and similarly the `does` predicate) denotes the performance of an attempt, but in no case it implies that the associated result has been achieved. In general, result is defined as *completion* of the action (in the sense of successful execution), as e.g. `stabbed` or `killed`, and, in case of actions specified with a productive characterization, as an *effect*

[2] A prototype implementation in *answer set programming* (ASP) of most ideas presented here is publicly available at `http://leibnizcenter.org/resources/JURIX2019/actions.lp`.

[3] On similar lines, Sartor [7] considers the procedural and productive characterisations for the types of actions to be used for norm modelling. Amongst others, Clark and Clark [9] include also a *stative* characterisation.

[4] As a convention, we use the *-ing* verbal form for identifiers denoting the action as a *process* or performance, and the *-ed* form for the action denoted as an *object*, or act.

in the world, as e.g. `dead`. Note that the outcome-form specifies the final result, but does not necessarily refer to intent (as in case of accidents). The intent-form makes instead clear that the outcome of the action is performed with intent.

General properties Procedural characterizations can be associated to immediate intents (cf. Searle's *intention-in-action*, or by seeing intentions as selected plans, as in Bratman's account, basis for most BDI agent architectures):

```
performs(brutus, stabbing) -> aims(brutus, stabbing).
```

Intentional procedural content can be brought to productive:

```
aims(brutus, stabbing) <-> aims(brutus, stabbed).
```

By definition, all actions comes along with an implicit productive content, e.g.:

```
performs(brutus, stabbing) <*> brings(brutus, stabbed).
```

The symbol `<*>` was used above to highlight that the logical equivalence between performance and outcome does not always hold, as performances cover also failed attempts of action. We propose here a possible logical model, in the simplifying assumption of dealing only with atomic actions (i.e. their duration is irrelevant w.r.t. the model granularity). Clearly, if an act has been completed, then performance has occurred:

```
brings(brutus, stabbed) -> performs(brutus, stabbing).
```

In contrast, we can assume that performance is completed *by default*, unless it is known otherwise. We introduce then a *strong negation* predicate `neg`, but we also rely on the unary operator *default negation* `not` provided by the logic programming semantics:

```
performs(brutus, stabbing), not neg(brings(brutus, stabbed))
  -> brings(brutus, stabbed).
```

Note that, because actions of any characterization can be described in the task form, this property is inherited by the `does` predicate. In sum, by generalizing these examples we can identify a few axioms mapping *observations of performances* (`performs/2`), *attribution of causal responsibilities* (`brings/2`) and of *intentions* (`aims/2`), to and from *action descriptions* (`does/2`).

Perfect, imperfect actions, etc. Let us consider actions identified by a task description `A` and an outcome description `R`, related by the predicate `actionResult/2`. Let us consider that performance has a certain duration, but that the production of the outcome is (qualitatively) immediate. The following qualifications of an action `A` can be defined as conjunctions of `does(X, A)` and `actionResult(A, R)` with these other conditions:

- *perfect action*: `brings(X, R)`
- *imperfect action*: `neg(brings(X, R))`
- *ongoing action*: `not(brings(X, R))`
- *successful intention*: `aims(X, R), brings(X, R)`
- *failed intention*: `aims(X, R), neg(brings(X, R))`
- *ongoing attempt*: `aims(X, A), not(brings(X, R))`

where the `not/1` predicate is true if no conclusion about the term is possible, i.e. `not(P)` is true when `not P` and `not neg(P)` are true. So, by relying on the idea of imperfection, action can be defined *negatively*:

```
does(X, neg(A)) <-> imperfect(does(X, A)).
```

meaning that the action has been performed, but has not reached the expected result (*failure*). Note in contrast how `neg(does(X, A))` means that performance has not been initiated (*omission*).

3. Representing causation

In a computational system, causal mechanisms triggered by an action `A` performed by an agent `X` in condition `C` and resulting in producing or consuming an object `r`, can be implemented as *reactive rules*, similarly to *event-condition-action* (ECA) production systems:

```
performs(X, A) : holds(C) => +r.  % initiation
performs(X, A) : holds(C) => -r.  % termination
```

In our case, consequences (neglecting temporal aspects) consist in the initiation (`+`) or the termination (`-`) of one or more objects.

Causation in logical reasoning At further inspection, events as e.g. `performs(X, A)` have an implicit temporal annotation, because the agent might perform several times the same type of action. Thus, assuming actions to be atomic (immediate) and interleaved (an actor cannot performs the same action twice at the same moment), `performs(X, A, T)` would denote a well-specified action instance. Further, in the moment in which we are dealing with time, dynamic facts have to be transformed into fluents: any (predicate) object `O` requires to be situated in time, as in `holds(O, T)`. Neglecting the enabling condition `C`, causal mechanisms could be then rewritten by making explicit the *change* of state for the fluent caused by the action:

```
performs(X, A, T), initiates(A, R), neg(holds(R, T-1)) -> holds(R, T).
performs(X, A, T), terminates(A, R), holds(R, T-1) -> neg(holds(R, T)).
```

(note that that, written in this form, `A` is an action type, while `R` is an object instance.) Unfortunately, these axioms are not sufficient for a logically sound reasoning. As shown in *situation calculus* [10], *event calculus* [11] and functionally similar solutions, additional axioms are required to capture *inertia*, *circumscription* and related epistemic properties. Let us consider for instance the simplest version of event calculus:

```
%% event calculus axioms (F fluent, A action type, T, T1, T2 times)
holds(F, T) :- initially(F), not clipped(0, F, T).
holds(F, T2) :- occurs(A, T1), initiates(A, F, T1), T1 < T2,
                not clipped(T1, F, T2).
clipped(T1, F, T2) :- occurs(E, T), T1 <= T,  T < T2, terminates(A, F, T).
```

Here, actions and fluents are reified as terms rather than as predicates. Intuitively, this is because change occurs at a meta-level with respect to the level of objects, and then everything has to be brought at meta-level to reason with it. In contrast, the notation of *reactive rules* enables in principle to abstract temporal attributes, as it introduces constraints only at the level of events. The following reactive rule implements a causal mechanism:

```
performs(X, A): initiates(A, R) => +R.
```

but it corresponds to a logical dependence at event level (`+` act as a unary predicate instead of an operator). Then, some other computational mechanism is responsible for executing the initiation and termination of fluents. For their compactness, it is tempting to maintain the description of causal mechanisms as reactive rules separated from that of necessary constraints holding between the objects, even knowing that they are not independent: certain causal mechanisms can create implicit constraints, as well as given constraints can inhibit certain causal ramifications. However, it is important to remind here that it is possible to semantically unify them, e.g. by using event calculus.

4. Representing power

Power—of an agent `X` towards an object `Y` to obtain a consequence `R` (concerning `Y`) by performing an action `A`—can be seen as the reification of a causal mechanism:

```
power(X, Y, A, R) <-> [performs(X, A) => +R(Y)].
```

The biconditional can be nested in the reactive rule:

```
performs(X, A) : power(X, Y, A, R) => +R(Y).
```

unveiling that the `initiates` predicate seen above is nothing else than a coarser description of `power`. With respects to conditions, power, even more when acting on symbolic objects (as for institutional power), is grounded on three qualification processes: (1) parties `X` and `Y` qualify to certain roles; (2) action `A` qualifies to a certain type/form; (3) context (here implicit, typically concerning where and when and the absence of overruling by another normative source). Each of these components brings conditions on the application of the causal mechanism:

```
power(X, Y, A, R) :- role(X, x), role(Y, y),
  action(A, a), actionResult(A, R), context(C, c).
```

Ability and susceptibility In the general causal interpretation, power primarily addresses the agent party (the one performing the action), so it can be renamed as *ability*:

```
ability(X, Y, A, R) <-> power(X, Y, A, R).
```

In duality, we can define the notion of *susceptibility* by primarily addressing the recipient party. A recipient is susceptible to an action (and then to the agent performing it) if it suffers a change because of its occurrence:

```
susceptibility(Y, X, A, R) <-> power(X, Y, A, R).
```

Negative powers By analogy to physics, in which forces can be attractive and repulsive, given a certain power, we can define its opposite by changing the sign of the outcome (cf. negative power/liability in [12]):

```
neg-power(X, Y, A, R) <-> power(X, Y, A, neg(R))
```

On the other hand, we can define the absence of power as the irrelevance of the action with respect to a certain outcome:

```
no-power(X, Y, A, R) <-> not power(X, Y, A, R), not neg-power(X, Y, A, R).
```

Negative susceptibilities and no-susceptibilities can be defined accordingly.

Preparatory/interfering actions, enabling/disabling powers An action `IA` interferes with an action `A` if, when the first is performed, it inhibits the outcome usually expected for performing the second. This notion is crucial for defining e.g. protection measures against interference as for *freedom of speech* (see e.g. [7]). Interestingly, it can be expressed in terms of powers; as a matter of facts, the interfering action modifies the power associated to the action target of the interference. The modification can be *structural* (it holds after `IA`'s completion) or *contingent* (it holds as long as the performance of `IA` is occurring), constraints that can be captured respectively at event level and at object level:

```
% structural (at event level)
power(Z, power(X, Y, A, R), IA, neg)
  <-> [ performs(Z, IA) => +neg(power(X, Y, A, R)). ]
```

```
% contingent (at object level, neglecting the time variable T)
power(Z, power(X, Y, A, R), IA, neg)
  <-> [ not performs(Z, IA) -> power(X, Y, A, R).
        performs(Z, IA) -> neg(power(X, Y, A, R)). ]
```

Enabling powers, associated for instance to *preparatory* or *support actions*, can be described in a dual way.

5. Conclusions and future developments

Implicitly or explicitly, most systems referring to regulations, policies and similar constructs in the computational domain refer to some form of *deontic logic*. Plausibly because of the strict control structure inherent to computational systems, the potestative category is usually neglected. However, because computational systems are becoming more and more social systems with *de facto* decentralized control structures, it becomes crucial to form a theory of power, so that institutional design in computational settings can intervene directly at the *social coordination* level of the guidance problem. In principle, this representational standpoint should help to study the entrenchments holding between physical and institutional actions.

Directed by this higher-order goal, the present paper presents our starting point for an operational axiomatization of power structures in a logic programming setting, motivated by recent results in LP research and applications. It explicitly introduces the intentional level in specifying actions, it elaborates on the connection between productive characterization of actions and causation, it defines a way to compute failures and omissions, and establishes a connection between causation, ability/susceptibility and power, enabling a definition of interfering actions. Future extensions of this work will focus on a wider number of institutional patterns (ex-ante vs ex-post enforcement, punishment-based vs reward-based enforcement, delegation, etc.) and concepts (recklessness, negligence, etc.).

References

[1] N. MacCormick and J. Raz, Voluntary Obligations and Normative Powers, *Proceedings of the Aristotelian Society* **46**(1972) (1972), 59–102.
[2] W.N. Hohfeld, Fundamental Legal Conceptions as Applied in Judicial Reasoning, *The Yale Law Journal* **26**(8) (1917), 710–770.
[3] D.M. Gabbay, J. Horty and X. Parent (eds), *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013.
[4] L. Lindahl, *Position and Change: A Study in Law and Logic*, Synthese Library, Springer, 1977.
[5] D. Makinson, On the formal representation of rights relations, *Journal of Philosophical Logic* **15** (1986).
[6] A.J.I. Jones and M. Sergot, A Formal Characterisation of Institutionalised Power, *Journal of IGPL* (1996).
[7] G. Sartor, Fundamental Legal Concepts: A Formal and Teleological Characterisation, *Artificial Intelligence and Law* **14**(1) (2006), 101–142. doi:10.1007/s10506-006-9009-x.
[8] J.F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, MIT Press, 2000.
[9] H.H. Clark and E.V. Clark, *Psychology and Language: An Introduction to Psycholinguistics*, Harcourt Brace Jovanovich, 1977. ISBN ISBN 9780155728158.
[10] R. Reiter, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, MIT Press, 2001.
[11] M. Shanahan, The Event Calculus Explained, *Artificial Intelligence Today* (1999), 409–430.
[12] G. Sileno, A. Boer and T. van Engers, On the Interactional Meaning of Fundamental Legal Concepts, in: *Proceedings of JURIX 2014*, Vol. FAIA 271, 2014, pp. 39–48.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190328

Application of Character-Level Language Models in the Domain of Polish Statutory Law

Aleksander SMYWIŃSKI-POHL [a], Krzysztof WRÓBEL [b], Karol LASOCKI [a] and Michał JUNGIEWICZ [a]

[a] *AGH University of Science and Technology, Krakow, Poland*
[b] *Jagiellonian University, Krakow, Poland*

Abstract. Polish statutory law so far is distributed as PDF, HTML and text files, where the structure of the rules and the references to internal and external regulations is provided only implicitly. As a result, automatic processing of the regulations in legal information systems is complicated since the semi-structured text needs to be converted to a structured form. In this research, we show how character-level language models help in this task. We apply them to the problems of detecting the cross-references to structural units (e.g. articles, points, etc.) and detecting the cross-references to statutory laws (titles of laws and ordinances). We obtain 98.7% macro-average F1 in the first problem and 95.8% F1 in the second problem.

Keywords. character-level language models, cross-reference recognition, language modelling, legal text processing, Polish law

1. Introduction

The Polish statutory law is available for everyone in the system called the Internet System of Legislative Acts (Internetowy System Aktów Prawnych – ISAP[1]). The system contains all the bills, ordinances, rulings of the Constitutional Tribunal and international agreements adopted from 1918, the year Poland has regained its independence, until now. All acts in the system are distributed as PDF files. Some of the metadata of the acts are provided directly on the web page dedicated to the individual document, but the actual content of the document is structured only visually.

Our goal is automatic structuring of the body of the Polish legislative acts. Much of this structure may be processed with regular expressions (especially the structural units of acts, since their patterns are very rigid). In this research, we concentrate on two issues that are harder to tackle with simple, rule-based techniques: detection of cross-references to structural units of the acts, such as articles, paragraphs, and points and detection of cross-references to titles of legislative acts. We apply the same algorithm to both problems, namely we use a recurrent neural network and a character-level language model (cLM).

[1] `http://isap.sejm.gov.pl`

2. Related Work

To our knowledge, there are very few works on detecting legal cross-references using machine learning methods. Almost all of the systems described in literature use a rule-based approach.

One approach to the problem based on machine learning methods is described in [1] and was tested on Japanese texts. The novelty of this work is two-fold. It lies in using machine learning for legal references resolution. Secondly, the authors claim their innovation is resolving references not only to document targets but also to sub-document parts. Their system achieves 80.06% in the F1 score for detecting references, 85.61% accuracy for resolving them, and 67.02% in the F1 score for end-to-end setting task on the Japanese National Pension Law corpus.

A more classical approach for detecting cross-references in legal texts is [2]. The authors used certain NLP patterns to build a rule-based system. These patterns were developed on Luxembourg's legislation, written in French. The system was tested on the Personal Health Information Protection Act (PHIPA) by the Government of Ontario, Canada, written in both French and English and on several Luxembourgish legislative texts.

As a rule-based baseline system for detecting references in Polish legal texts we used one developed as a part of the SAOS (Court Judgement Analysis System) project [3]. The general schema of the algorithm is to first tokenize the text, and then extract certain ranges of the tokens as candidates for references. After that, based on some rules and regular expressions, it looks for fragments of text that contain legal references and splits them into classes.

On the other hand, the application of language-model based algorithms for various NLP tasks seems to be a standard approach at least in the last 2 years. In the past, in tasks such as text classification or named entity recognition, *words* were treated as the main units of processing [4]. The vector space model (VSM) was one of the formalisms best suited for providing a coherent representation of words for ML algorithms. They used to be represented as one-hot encoded vectors, where the size of the vector equals the size of the vocabulary.

A relatively recent solution to the problem of limited vocabulary is word embeddings (WE) – dense vector representations of words. These embeddings are obtained in an unsupervised manner, thus they are easily adaptable to new languages and problems. The most successful methods are based on neural networks and factorization of co-occurrence matrices. Popular systems, such as word2vec [5], GloVe [6] and fastText produce [7] so-called static WE, since the representation is independent of the word context. As such it limits the expressiveness of the models since the vectors are unable to capture polysemy. The „traditional" word embeddings also face a problem of the composition of multiple words into one vector – the vectors might be linearly combined (e.g. averaged) or units for multi-word entities have to be defined separately.

Contextual word embeddings are the latest representation able to solve that problem. The most recent systems: ELMo [8], BERT [9] and Flair [10] encode not only the word in question but also its surroundings. Moreover, Flair and ELMo do not employ tokenization, since they use character-based or byte-pair-encoding (BPE) based embeddings. This allows for computing dense representation for unrestricted spans of text.

The most recent studies show [11] that such models can solve a large number of problems: language modeling, named entity recognition, machine translation, text gen-

Figure 1. An example of references to structural units appearing in the Polish statutory law.

Figure 2. An example of a reference to a legislative act appearing in Polish statutory law.

eration, text summarization, natural language inference, and question answering – with very little or even no manually annotated data for the downstream task. Yet we haven't found any paper that uses contextual WE for the problems we tackle.

3. Problem Description

3.1. Cross-References to Structural Units

The problem of cross-reference to structural units of statutory law is depicted in Figure 1. The example comes from an amending act, which are typically packed with all types of references. We call these references *cross-references to structural unites*, since they point to particular, structural fragments of laws, such as chapters, articles, paragraphs, points, letters, indents as well as particular sentences.

The cross-references to structural units in the Polish statutory law can be roughly divided into two groups: those that are used in the amending bills, where the sequence of units almost always starts with an article[2], which is further placed within a particular law, and those that are more common in non-amending bills, when the top-level element may be any valid unit. In the second case, the higher-order units are indicated implicitly as the units the reference appears in.

We define the problem of detecting cross-references to structural units as the detection of the exact span of the reference and as a qualification of the span as one of the following (13) types: *article*, *point*, *paragraph*, *letter*, *indent*, *chapter*, *division*, *branch*, *title*[3], *book*, *part*, *subchapter*, and *sentence*. However, since the rule-based tool devised to detect the cross-references in the Polish law detects only 3 types of references: *article*, *paragraph*, and *point*, to make the comparison fair we only provide the results for these three categories.

3.2. Cross-References to Statutory Laws

The problem of *cross-references to legislative acts* is depicted in Figure 2. Usually, the title of an act starts with *ustawa* (bill) or *rozporządzenie* (ordinance), followed by date of publication and ends with a detailed location of the act, allowing for its unambigu-

[2]Rare cases of amendments include a chapter which is completely removed or added and an amendment in the title of the law.

[3]No to be confused with the title of the law.

ous identification. Although the priming word is always present, date might be omitted and title does not have to be followed by the location details. These two features make detection of act titles a challenging problem. We define the problem of detecting cross-references to legislative acts as the detection of the exact span of the title.

4. Applied Algorithms

4.1. Character-Level Language Model

We use Flair toolkit [10] to train the cLM and compute the contextual embeddings. The cLM allows for obtaining embeddings of any fragment of text. To achieve the best results two language models are trained: forward and backward.

The training of cLM starts with preparation of the corpus, definition of the character dictionary and determination of the training parameters. The loss function is cross-entropy, which translates to perplexity (exponent of cross-entropy).

One of the most important training parameters is the size of the internal state of RNN. The authors of Flair use 1024 or 2048 [10], resulting in 2048 or 4096 components in the final embedding. This is a large number in comparison to popular static WEs that range from 100 to 300 in size. On the other hand, the dictionary is much smaller since it only includes a limited subset of Unicode characters. The default learning rate is set to 20, which decreases with the training process. In our experiments, when training the cLMs the size of the internal state of the RNN was set to 2048.

4.2. Cross-References Detection as NER

Flair also includes a module which performs Named Entity Recognition (NER). The text is split into tokens and the contextual embeddings of the tokens are computed based on the cLM one character after the token (for the forward model) and one character before the token (for the backward model). The vectors are concatenated and they are used as the representation of the token in a biLSTM network. There is a Conditional Random Field (CRF) layer at the top, which performs the final assignment of the tags to the tokens. This model was used directly in both experiments, since the detection of both types of cross-references may be treated as a NER-like problem.

5. Data

The features of the corpus used to train the cLM are given in the second column in Table 1. The number of tokens is not very large, compared to typical corpora used to train language models, yet thanks to its domain specificity, we have achieved good perplexity (92,4) training for 3 days on one node with two K40 GPUs. To prepare the documents for the problems, we have collected approximately 10 acts from each year, starting in 1994 and ending in 2018, resulting in 243 documents.

The annotationwas performed by 5 annotators with good knowledge of law (at least 5 years of studies in law) or linguistics (a master degree was required). We used the Inforex system [12] and followed a scheme where each document was annotated by two annotators and then a super-annotator resolved the conflicts. In fact the number of

Table 1. The statistics of the corpus used to train the language model (Acts) and the annotated sub-corpus.

Measure	Acts	Annotated
Number of tokens	9 776 676	396 963
Number of distinct lemmas	36 716	14 737
Number of sentences	371 082	14 737
Number of documents	1 892	243
Size in MBs	56	3.6
Average sentence length	26.3	26.9

Table 2. The F_1 score for the detection of the cross-references to structural units.

System	art	pkt	ust	micro	macro
rule-based	0.9454	0.9360	0.9364	0.9401	0.9393
cLM-based	**0.9797**	**0.9942**	**0.9874**	**0.9850**	**0.9871**

Table 3. The precision, recall and F_1 score for the detection of the cross-references to titles of legislative acts.

System	Precision	Recall	F_1
rule-based	**1.000**	0.6316	0.7742
cLM-based	0.9579	**0.9579**	**0.9579**

differences in annotations was very small and usually these were omitted or superfluous punctuation marks. The annotation of the data (the first round with two annotators and the second round with the super-annotator) took approximately 120 man-hours.

6. Experiments

We have split the annotated data (on the document level) into sub-corpora used for training of the model (*Train*), for tuning of the hyper-parameters (*Dev*) and for testing the model (*Test*) in ratio 70%/15%/15%. We have compared the performance of our model with SAOS extractors designed to perform the same task but in the domain of court rulings.

Table 2 contains the results for detection of cross-references to structural units. Our system achieves better results for all classes than the rule-based system. For articles, the performance is almost perfect. Table 3 contains the results for detection of cross-references to the titles of the legislative acts. The rule-based system has perfect precision, but its recall reaches only 63%. Our system is not completely precise (though 96% is a very decent result), but its recall is significantly higher (also 96%), thus the F_1 score is much better. Comparing to the first problem, it is apparent that the detection of titles is more challenging, but the system works very well.

7. Conclusions and Future Work

We have presented the results of the two experiments where we applied a cLM to the problems related to the processing of statutory law. The results of the experiments with the detection of cross-references obtained using that model are better than the results of a rule-based system. In all cases, the F_1 scores were above 95% showing that the models may be used practically.

In our future work, we will apply similar models to automatic detection and structuring of the amending acts, as well as to the detection of relations between cross-references to structural units.

Acknowledgments. This work was supported by the Polish National Centre for Research and Development – LIDER Program under Grant LIDER/27/0164/L-8/16/NCBR/2017 titled "Lemkin – intelligent legal information system". This research was also supported in part by PLGrid Infrastructure.

References

[1] O.T. Tran, N.X. Bach, M.L. Nguyen and A. Shimazu, Automated reference resolution in legal texts, *Artificial Intelligence and Law* **22** (2013), 29–60.

[2] N. Sannier, M. Adedjouma, M. Sabetzadeh and L. Briand, An automated framework for detection and resolution of cross references in legal texts, *Requirements Engineering* **22**(2) (2017), 215–237. doi:10.1007/s00766-015-0241-3.

[3] SAOS text mining extractor, GitHub, 2015.

[4] D. Jurafsky and J.H. Martin, *Speech and language processing*, 2009.

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).

[6] J. Pennington, R. Socher and C.D. Manning, GloVe: Global Vectors for Word Representation, in: *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. `http://www.aclweb.org/anthology/D14-1162`.

[7] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics* **5** (2017), 135–146.

[8] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, in: *Proc. of NAACL*, 2018.

[9] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[10] A. Akbik, D. Blythe and R. Vollgraf, Contextual string embeddings for sequence labeling, in: *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, Language Models are Unsupervised Multitask Learners (2019).

[12] M. Marcińczuk, M. Oleksy and J. Kocoń, Inforex—A collaborative system for text corpora annotation and analysis, in: *Proceedings of the international conference recent advances in natural language processing, RANLP*, INCOMA Shoumen, 2017, pp. 473–482.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190329

Combining Textual and Visual Information for Typed and Handwritten Text Separation in Legal Documents

Alessandro TORRISI, Robert BEVAN, Katie ATKINSON, Danushka BOLLEGALA and Frans COENEN
Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK

Abstract. A paginated legal bundle is an indexed version of all the evidence documents considered relevant to a court case. The pagination process requires all documents to be analysed by an expert and sorted accordingly. This is a time consuming and expensive task. Automated pagination is complicated by the fact that the constituent documents can contain both typed and handwritten texts. A successful auto-pagination system must recognise the different text types, and treat them accordingly. In this paper we compare methods for determining the type of text data contained within paginated bundle pages. Specifically, we classify pages as containing typed data only, handwritten data only, or a mixture of the two. For this purpose, we compare text classification methods, image classification methods, and ensemble methods using both textual and visual information. We find the text and image based approaches provide complimentary information, and that combining the two produces a powerful document classifier.

Keywords. Pagination of Legal Bundles, Image Classification, Text Classification

1. Introduction

Legal document pagination [1] is an important process that is typically performed prior to a court hearing. The purpose of pagination is to produce an indexed court bundle containing all of the information and evidence related to a case. Processing legal documents in this way improves an advocate's ability to present a case during a hearing. During pagination, a domain expert must filter large volumes of information, meticulously sorting documents according to subject and often chronology. For example, in the medical negligence domain [2], medical records represent an important source of evidence and can easily contain hundreds or thousands of pages; during pagination any sections that are relevant to the medical negligence case need to be extracted from this vast amount of information. Pagination is further complicated in this instance due to the often non-contiguous distribution of evidence contained within medical records. In addition to evidence of any negligent acts, a patient's medical history may be relevant to the case, as well as any negative outcomes they experienced as a result of the negligence, which may occur months or years after the initial negligent act.

Regardless of the complexity of the case under examination, the pagination of medical records is generally a time-consuming and expensive task. The expense involved in pagination calls for the development of automated methods to help speed up the process.

Legal documents typically contain both typed and handwritten texts. Identification of typed and handwritten texts is a necessary precursor to building an effective automatic pagination tool. For example, if a system can determine that a page contains only typed data, the page can be analysed using classical Optical Character Recognition (OCR) approaches, whereas if the page contains handwritten data, handwriting recognition will need to be applied. A further advantage of identifying the type of text contained within a page is that it can help with the task of page categorisation. For example, in the medical negligence domain, consultation notes and records of correspondence between parties often contain typed text only, whereas pages containing handwritten text only, and those containing a mixture of typed and handwritten data, often correspond to doctors' notes, consent forms and/or laboratory examination reports respectively.

In this paper we compare different methods of classifying paginated bundle pages into three categories: (a) typed text only pages (**typed**), (b) handwritten text only pages (**handwritten**), and (c) pages containing both types (**mixed**). We experimented with two different approaches. For the first approach the problem was treated as a text classification task. For the second approach it was treated as an image classification task. We compare different methods of visual feature extraction including construction of visual keypoints using classical feature extraction methods, and feature extraction using pre-trained convolutional neural networks (CNNs) [3]. We also experimented with fine-tuning a CNN pre-trained on a large dataset [4]. Finally we combined the separate methods in an ensemble, and observed that the two different approaches provided complimentary information, with a best accuracy of over 95%.

2. Proposed Approach

Law firms typically receive medical records in hard copy. These are then scanned and converted into an electronic format using OCR software. Modern OCR software performs very well, but typically typed text is handled better than handwritten text. OCR applied to handwritten text can be error prone. This discrepancy in the quality of texts produced when applying OCR to the different text types motivated our text classification approach. We treat the problem as a standard text classification task. Specifically, we built a dictionary of unigrams, bigrams, and trigrams using a distinct set of medical records from those used in our experiments. Documents were converted into a machine readable format using a Bag of Words (BoW) model. A logistic regression classifier was trained to classify the documents, optimized using a grid-search approach.

Humans are easily able to distinguish typed and handwritten text by eye. The purpose of the image classification approach was to exploit visual features. Additional motivation was the fact that OCR software will sometimes fail to identify any text in pages that do in fact contain text, limiting the effectiveness of the text classification approach. Image classification relies on the use of visual words related to small parts of a page (converted to an image) which carry information related to features such as colour, shape or texture. In the Computer Vision community, a number of local feature operators have been presented [5]. After the advent of the well known Scale Invariant Feature Transform (SIFT) [6], different alternatives were proposed which satisfy a more efficient and effective calculation. Two examples are ORB [7] and BRISK [8] which are considered in this paper. Both algorithms detect keypoints inside an image, and assign a feature vector to each keypoint, which is of dimension 32 and 64 for ORB and BRISK, respectively.

To provide a robust estimate of the best feature operator, a balanced dataset of 15M keypoints was extracted from training data using both BRISK and ORB. The extracted keypoints were clustered using a standard k-means with the aim of grouping together all the keypoints related to similar objects. At the end of the clustering, a dictionary was obtained composed of the centroids of the resulting clusters (sets of visual words).

A second image classification was performed using Convolutional Neural Networks (CNN). CNNs produce state-of-the-art image classification performance when trained with very large datasets. In our application, the dataset was very small. Fortunately, it is possible to leverage the power of CNNs without a large training set through transfer learning [3]. There are two main approaches to transfer learning for image classification: fine-tuning and feature extraction combined with a linear classifier. In the first the CNN's output layer, and any fully-connected layers at the top of the network, are replaced with the number of units in the output layer equal to the number of classes in the problem; model training is resumed with the new dataset. Due to the hierarchical structure of the features extracted from the different network layers, typically, only a subset of the layer weights towards the top of the network are updated during training, as features extracted in the lower layers are less specialized and therefore more likely to be useful. In the case of CNNs for extracting features, typically the fully-connected layers at the top of the network are replaced with a single pooling layer and a softmax classifier. The resulting network is trained to classify the new dataset, with only the softmax layer weights updated. It is also possible to truncate the network at a lower level prior to adding the pooling and softmax layers, which can lead to superior classification performance due to the hierarchical feature structure.

Both feature extraction and fine-tuning approaches were considered. First, we trained linear classifiers using features extracted with the following pre-trained networks: Xception, ResNet152V2, InceptionV3, InceptionResNetV2, MobileNet, DenseNet201, and NASNetLarge [9, 10, 11, 12, 13, 14, 15]. In each instance the fully-connected layers at the top of the network were replaced with an average pooling layer and softmax layer containing three units corresponding to the three classes. In addition, we experimented with extracting features using different sub-architectures of the InceptionResNetV2 network (the network was truncated at different levels prior to feature extraction). Next, we fine-tuned the pre-trained MobileNet network [13], replacing the fully-connected layers at the top of the network with three dense layers with ReLu activation function and a softmax output layer with three units. MobileNet was selected for fine-tuning in the belief that it was the least likely to overfit due to its relatively low parameter count. Each network was optimized using Adam [16].

3. Evaluation Data

To evaluate the proposed approach, 50 pre-paginated medical bundles were used of the form that might be used in accident claims litigation. 30 bundles were randomly selected as the training data. 3000 different pages were extracted, 1000 for each category. The last 20 bundles were used as test data. A total of 1800 pages were extracted (600 pages for each category). Creation of ground truth information was conducted using two different domain experts. Selection of candidate samples was undertaken to include as many handwritten writing styles as possible. All the collected documents were in PDF format. Text was at first extracted and then pages were converted to images to perform image

classification. Consent was obtained from clients to use their medical data for this research. Non-anonymous sensitive information was included and this prohibits us from making this data publicly available.

4. Experimental results

For the evaluation of both text and image classifications, the metrics used were Precision, Recall and the F1 measure. Results of the CNN feature extraction experiments are shown in Figure 1. Each of the CNNs produced useful classification features: the worst performing classifier, trained using features extracted with DenseNet201, achieved a class-averaged F1 score of greater than 80%. The best performing classifier, trained using features extracted with InceptionResNetV2, achieved an F1 score of over 89%. We found that extracting features at an earlier stage of the InceptionResNetV2 network improved classification performance by ∼2% (Figure 1). This may be because the features extracted at the top of the network are more specialized to the original training task.

Each network was optimized using Adam ($\eta = 0.001$; $\beta_1 = 0.9$; $\beta_2 = 0.999$). Prior to training, 20% of the training data was randomly selected for use as a validation set. Fine-tuning was performed for 50 epochs with early stopping according to validation loss. In the feature extraction setting, classifiers were trained for 500 epochs without early stopping. In both settings, model checkpoints were saved at epochs where the validation performance exceeded the previous best, and the best performing model was selected for use in the evaluation. Each experiment was repeated 5 times, and the best performing models in each trial were combined in an ensemble for the evaluation in order to minimize the effect of model initialisation.

Table 1 shows the evaluation of seven different classifiers, three of them consider an ensemble of textual and visual information. The best image classification (conducted considering a classical image classification approach) was achieved considering BRISK as feature operator and an Extra Tree Random Forest (ETRF) classification model. Seven different values of k in the range (50, 2000) were tested to find the optimal size of the code-word representing each page of a medical bundle. The achieved results are statistically comparable but a best F1 measure equal to 90.3% was registered considering BoVW vectors composed by 750 features (see second row in Table 1). Class probabilities obtained conducting a text classification improved the results of image classification in all the conducted experiments. F1 measures equal to 93.5% and 95.7% were achieved when text classification is combined with the image classification conducted through traditional approaches and using a fine-tuned CNN such a MobileNet, respectively. A fine-tuned MobileNet improved the class-averaged F1 score by 12% when compared with the classifier trained with features extracted using MobileNet, without any fine-tuning.

Use of visual information was useful in this case to resolve labels for pages not containing any text. 70% (28 out of 40) of empty pages were correctly classified using MobileNet. Another advantage of using a CNN instead of a classical image classification approach is related to time performance. It takes about a second to classify a document image considering classical approaches. It includes the extraction of keypoints, their conversion to a BoVW vector and classification. Using a fine-tuned CNN such as MobileNet drastically reduces the computational time, ensuring a classification of up to five images per second.

There are two main sources of error produced by the proposed approach. The first one is related to the straight lines which might be eventually included in pages of medical

CNN	Prec	Rec	F1
DenseNet201	84.9	82.1	82.3
ResNet152V2	85.5	84.3	84.4
MobileNet	85.3	84.5	84.6
Xception	86.4	85.7	85.8
InceptionV3	88.9	88.2	88.3
NASNetLarge	89.0	88.4	88.4
InceptionResNetV2	**89.7**	**89.3**	**89.4**

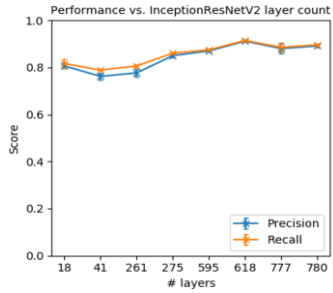


Figure 1. Left: Comparison of models trained using different CNNs as feature extractors. **Right:** Classifier performance for models trained using features extracted from various points of the InceptionResNetV2 network. All metrics were macro-averaged across classes.

Table 1. Evaluation of seven different decision systems.

Classifier	Prec	Rec	F1
Text Classification	86.6	86.5	86.5
ETRF	90.3	90.3	90.3
Text Classification + ETRF	93.6	93.5	93.5
MobileNet	94.7	94.7	94.7
Text Classification + MobileNet	**95.7**	**95.7**	**95.7**
InceptionResNetV2	91.4	91.1	91.2
Text Classification + InceptionResNetV2	94.5	94.4	94.4

records. As most of the pages containing straight lines are represented by medical forms, charts and tables, the current classifier is more prone to associate this information to pages belonging to the class “mixture”. A few other errors were discovered in pages not containing enough contrast between background and foreground pixels, making the keypoints extraction using BRISK more difficult. These errors suggest to us to conduct an image pre-processing step prior to classification in order to remove noise or any other artefact which can alter the classification verdict.

4.1. Discussion

The pagination process requires that a complete set of documents is at first allocated to analysts with expertise in a particular legal area. In the medical domain, pages of medical records are sorted and collated according to the instructions of a referring solicitor. Usually, this process requires the extraction of three main sub-bundles composed by correspondence, clinical information and General Practitioner (GP) records, respectively. The machine learning approach proposed in Torrisi et al. [1] to assist the pagination was originally implemented for such purposes but it considered only text information. This means that it becomes less feasible in the presence of a badly scanned document or when an OCR device does not provide enough accuracy. The approach proposed in this paper is intended to resolve such eventualities and it can be combined with the functionalities already implemented in [1] for providing further data categorisation. A further advantage of using the proposed approach is as a precursor step for developing bespoke OCR solutions for handwriting recognition, which is one of our ongoing steps.

5. Conclusion

In this paper an approach for separating typed from handwritten data was proposed for assisting the pagination process in litigation claims. Classification was conducted considering textual information, visual information and an unweighted combination of both types of data. Two different image classification approaches were tested: one considered feature extraction and classification using standard machine learning models; a second image classification was achieved considering the use of pre-trained neural networks. Best classification performance was conducted considering the combination of text classification with MobileNet, which resulted in a final F1 measure equal to 95.7% on test data composed of 1800 samples. Given the promising performance, our current target is increasing the size of the employed dataset, and also including documents from different sources so that a multi-context machine printed / handwritten text separator can be achieved. We are also investigating further alternatives to sort the documents according to subject. This is made possible through testing of the proposed application by analysts who operate in the medical negligence context.

References

[1] A. Torrisi, R. Bevan, K. Atkinson, D. Bollegala and F. Coenen, Automated Bundle Pagination Using Machine Learning, in: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, 2019, pp. 244–248.

[2] R. Bevan, A. Torrisi, D. Bollegala, F. Coenen and K. Atkinson, Extracting Supporting Evidence from Medical Negligence Claim Texts, in: *Proceedings of the 4th International Workshop on Knowledge Discovery in Healthcare Data*, KDH '19, 2019.

[3] M. Hussain, J. Bird and D. Faria, A Study on CNN Transfer Learning for Image Classification, *CoRR* (2018).

[4] A. Krizhevsky, I. Sutskever and G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, 2012, pp. 1097–1105.

[5] S.A.K. Tareen and Z. Saleem, A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK, *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (2018), 1–10.

[6] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision* **60**(2) (2004), 91–110.

[7] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, ORB: an efficient alternative to SIFT or SURF, 2011, pp. 2564–2571.

[8] S. leutenegger, M. Chli and R. Siegwart, BRISK: Binary Robust invariant scalable keypoints, 2011, pp. 2548–2555.

[9] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, *CoRR* (2016).

[10] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, *CoRR* (2015).

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the Inception Architecture for Computer Vision, *CoRR* (2015).

[12] C. Szegedy, S. Ioffe and V. Vanhoucke, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, *CoRR* (2016).

[13] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, *CoRR* (2017).

[14] G. Huang, Z. Liu and K.Q. Weinberger, Densely Connected Convolutional Networks, *CoRR* (2016).

[15] B. Zoph, V. Vasudevan, J. Shlens and Q.V. Le, Learning Transferable Architectures for Scalable Image Recognition, *CoRR* (2017).

[16] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in: *the 3rd International Conference for Learning Representations, 2015*.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190330

Legal Text Generation from Abstract Meaning Representation

Sinh Trong VU [a] Minh Le NGUYEN [a] and Ken SATOH [b]
[a] *School of Information Science, Japan Advanced Institute of Science and Technology*
[b] *National Institute of Informatics, Japan*

Abstract.
Generating from Abstract Meaning Representation (AMR) is a non-trivial problem, as many syntactic decisions are not constrained by the semantic graph. Current deep learning approaches in AMR generation almost depend on a large amount of "silver data" in general domains. While the text in the legal domain is often structurally complicated, and contain specific terminologies that are rarely seen in training data, making text generated from those deep learning models usually become awkward with lots of "out of vocabulary" tokens. In our paper, we propose some modifications in the training and decoding phase of the state of the art AMR generation model to have a better text realization. Our model is tested using a human-annotated legal dataset, showing an improvement compared to the baseline model.

Keywords. AMR Generation, Deep Learning, Legal

1. Introduction

Abstract Meaning Representation, or AMR in short, is a semantic annotation scheme that encodes a natural language sentence as a rooted, directed graph. Every vertex and edge of the graph is labeled according to the sense of the words in a sentence [1]. We give an example of AMR annotation in Figure 1, where the nodes (e.g. *"enjoy-01", "right-05"*, ...) represent concepts, and the edges (e.g. *":arg0", ":condition"*, ...) represent relations between those concepts. Recently, AMR gains a lot of attention in the NLP research community, as it is widely used as an intermediate meaning representation for NLP tasks, e.g. machine translation [2], summarization [3].

To obtain success in those tasks, the problem of AMR-to-text generation has to be solved effectively. Several deep learning approaches have been proposed to tackle this problem by leveraging a large amount of silver data [4], [5]. Despite acceptable performance on general domain text, those generating models struggle in dealing with the legal domain, where the sentences are complicated structure and contain domain-specific terms. We figure out that lots of out-of-vocabulary words are generated, and almost the negation and conditional sentences are generated incorrectly.

In our paper, we propose a modification in the training phase and decoding phase of the baseline graph to sequence model to improve the generation quality. Specifically, in the training process, we constrain the encoder-decoder model by a controllable variable to avoid the repetitive token generating as well as guiding the model to recognize the

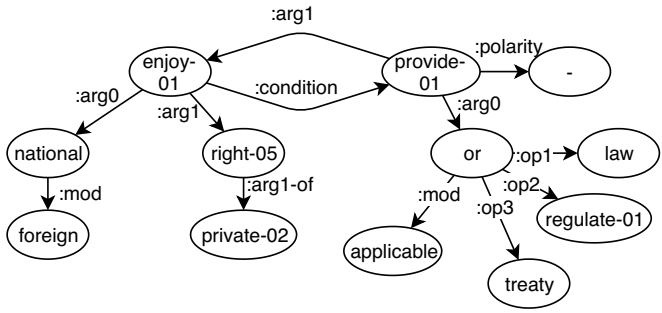


Figure 1. AMR graph for the sentence *"Unless otherwise provided by applicable laws, regulations or treaties, foreign nationals shall enjoy private rights"*.

negation and conditional sentences more appropriately. After training, the model is fine-tune with a silver dataset generated from a civil code in the English version. Moreover, we adopt weighted decoding [6] with a modified beam-search algorithm to avoid out-of-vocabulary words. The model is tested using a human-annotated legal dataset, showing improvement over the baseline model.

2. Background

2.1. Deep learning approaches in AMR-to-text Generation

Given an AMR graph $G = (V;E)$, where V and E denote the sets of nodes and edges, respectively, the goal is to generate a sentence $W = (w_1, w_2, ..., w_n)$ where w_i are words in the vocabulary. Since first introduced as a shared task at SemEval-2017 [7], several approaches have been proposed to tackle this generation problem, with a dominance of deep learning models. Konstas et al. [5] linearized AMR graphs, then adopt an encoder-decoder model to translate these string-like objects into natural language (NeuralAMR). Song et al. [4] modified the encoder side architecture to capture the graph structure data more properly. This resulted in a graph-to-sequence model (Graph2Seq) capable of generating well-written text, obtaining the state of the art BLEU score in this generation problem in 2018. However, these models still struggle when dealing with legal text, i.e. Graph2Seq obtains 9.86 BLEU score on JCivilCode [8], comparing to the score of 32.0 on LDC2017 test set. In our paper, we rely on Graph2Seq to build our baseline model.

2.2. The baseline model

As mentioned before, we adopt the graph-to-sequence model in [4] as our baseline. With a given AMR graph $G = (V;E)$, each node v_i is represented by a hidden state vector h_i, initializing by the word embedding of that node. The graph state g is defined as the set of h_i. Information exchange between a current node v_i and all incoming nodes and outgoing nodes connected to it are captured through a sequence of state transitions $g_0, g_1, ..., g_k$. The encoder side used a long short term memory (LSTM) network to perform this graph state transition. With this state transition mechanism, information of each node is propagated to all its neighboring nodes after each step. After k transition steps, each node

state contains the information of a large context, including its ancestors, descendants, and siblings, where k is the maximum graph diameter in the dataset (we choose $k = 9$ in our experiments). The decoder side is also a LSTM network incorporated with a copy mechanism [9] to deal with decoding objects like name entities, numbers, and date. The detail computation in each step can be found in the original paper.

3. Legal AMR generation

3.1. Conditional training

Conditional Training (CT) [10] is a method to learn an encoder-decoder model $P(y|g,z)$, where z is a discrete control variable and g is the AMR graph. We design z by annotating every (g,y) pair in the training set with the attribute we wish to control, e.g. the length of the linearized graph, or whether g contains negation or not. This attribute value will be determined during training, depend on each training sample. We use an embedding value with size 10 to represent the control variable z. This value will be concatenated to the decoder's input at each step. The objective function of training is given by the cross-entropy loss: $loss_{CT} = -\frac{1}{T}\sum_{t=1}^{T} logP(y|g,z,y_1,...,y_{t-1})$. Parameters of the model are initialized when training with the benchmark general domain dataset, then finetuning with the silver legal dataset to optimize $loss_{CT}$.

3.2. Decoding in legal style

To enhance the probability of generating words with certain features, we adopt Weighted Decoding (WD) that was introduced by Ghazvininejad et al. [11]. On the t_{th} step of decoding, the generated hypothesis $y_{<t} = y_1, \ldots, y_{t-1}$ is expanded by computing the score for each possible next word w in the vocabulary by the formula:

$$score(w, y_{<t}; g) = score(y_{<t}; g) + logP_{LSTM}(w|y_{<t}, g) + \sum_i w_i * f_i(w; y_{<t}, g).$$

In which $\log P_{LSTM}(w|y_{<t}, g)$ is the log probability of the word w calculated by the bi-LSTM network, $\text{score}(y_{<t}; g)$ is the accumulated score of the generated words in the hypothesis $y_{<t}$ and $f_i(w; y_{<t}, g)$ are decoding features with the corresponding weights w_i. There can be multiple features f_i to control multiple attributes, and the weights w_i are hyperparameters. A decoding feature $f_i(w; y_{<t}, g)$ assigns a real value to the word w. The feature can be continuous (e.g. the unigram probability of w) or discrete (e.g. the length of w in characters). A positive weight w_i increases the probability of words w that scores highly with respect to f_i and vice versa.

Another problem of generating text from legal AMR is the out of vocabulary tokens, where lots of words in the legal domain are not included in well-known word embedding, e.g. Word2Vec or Glove. We collect the vocabulary of three datasets: a benchmark dataset in general domains and two datasets obtained from Vietnamese and Japanese civil code. We observe that more than 30% of the words in these vocabulary do not appear in Glove [12]. To deal with this problem, we modified the beam search decoding algorithm. Specifically, after collecting an extra-vocabulary from the legal finetune set, we assign a binary feature to each word w in the test set representing whether w is in the legal vocab or not. This increases the probability of words in the legal vocabulary to be selected to the *top-k* generation, where k is the beam size.

4. Experiments and Results

4.1. Dataset Preparation

In our experiments, we use three datasets: (i) the benchmark dataset LDC2017T10 for training the baseline model, (ii) silver data generated from a Vietnamese Civil Code for fine-tuning the model, and (iii) the JCivilCode dataset [1] [8] for testing the performance. Because of lacking hardware resources, we do not conduct our experiments on silver data sampled from external corpora (like NeuralAMR and Graph2Seq using Gigaword).

Table 1. Statistics of the three dataset used in our experiments

Dataset	LDC2017T10	VN Civil Code	JCivilCode
Number of samples	36.521	3,073	128
Vocabulary size	29,943	3,026	778
Number of words out of vocab	4,453	602	270
:condition edge	1,794	190	69
Negation	10,947	356	57

In dataset (i) we use the linearization and anonymization algorithm provided by Song et al. [4] and Konstas et al. [5]. For dataset (ii), the silver data is obtained by performing two best parsers for legal text: JAMR [13] and CAMR [14] as suggested by Vu et al. [8]. Each sample sentence in the corpus will provide two AMR graphs, this also enlarges the dataset for finetuning our models. The statistics of these datasets can be found in Table 1.

4.2. Results and Analysis

We evaluate our models mainly by BLEU score [15] and METEOR score [16]. We also report the number of OOV words generated from each model. From Table 2, it can be observed that our both proposed modification improve the performance of text generation. While CT increases the BLEU score and METEOR score comparing to the baseline model, Legal Decoding (LD) helps reduce the OOV rate significantly. However, combining both two techniques does not result in the best score overall, where BLEU and METEOR score decrease slightly after LD, since this algorithm sometimes eliminates non-legal words from the *top-k* space.

Our experimental results also confirm the important role of training data. After finetuning with a legal dataset, we obtain 2.81 and 0.96 improvement on BLEU and METEOR score, respectively. When comparing to the state of the art pre-trained models, with a huge amount of data, our proposed modification still got lower results by a small margin.

To have a closer look, we provide some output examples for each model in Table 3. All the models still generate low-quality sentences, with grammatical errors and repetitive words. The baseline model trained without any legal data provides an out-domain word that does not appear in the source AMR graph. After finetuning, the sentences generated become longer but not so meaningful except for the output of CT model, which includes almost correct information. LD, as mentioned earlier, could help reduce the OOV rate overall, but may cause some words or fragments missing and repetitive.

[1] https://github.com/sinhvtr/legal_amr

Table 2. Generation results in BLEU score, METEOR score and number of OOV generated. The baseline Graph2Seq is trained on benchmark dataset only. The next four lines show our proposed modifications, with and without finetuning data. The last two lines are the results of two best pretrained models with extra corpus.

Model	BLEU	METEOR	OOV
Baseline Graph2Seq	5.50	16.78	135
Graph2Seq + CT	6.82	17.42	112
Graph2Seq + Finetune data	8.31	17.74	145
Graph2Seq + Finetune data + Conditional Training	**8.56**	**18.61**	143
Graph2Seq + Finetune data + LD	8.42	17.98	57
Graph2Seq + Finetune data + CT + LD	8.43	18.04	**57**
Graph2Seq Pretrained on 2M Gigaword corpus	9.31	21.38	29
NeuralAMR Pretrained on 2M Gigaword corpus	9.07	20.55	35

Table 3. Output comparison with an example from JCivilCode dataset

Gold data Unless otherwise provided by applicable laws, regulations or treaties, foreign nationals shall enjoy private rights.
Baseline model the foreign national enjoy a private right not if the applicable law or economic treaty
Baseline model + finetune data when it is not provided for by law or the treaties to enjoy the private rights , the foreign national shall have the enjoy private rights .
Baseline model + finetune data + CT the foreign national will enjoy private rights without providing applicable regulate regulate or treaty
Baseline model + finetune data + CT + LD when a foreign national enjoys the private right , if not provided for by law or the provisions of law or the provisions of law .
Graph2Seq Pretrained on 2M Gigaword foreign nationals will enjoy private rights while there are no laws or regulations if the or or without the regulations are provided .

5. Conclusion

In this paper, we figure out the difficulties of AMR generation in the legal domain, where the logical structure is complicated and lots of domain-specific terms are not in the well-known vocabulary. We propose two modifications to the training and decoding phases of the state of the art graph to sequence model to tackle these difficulties. The experimental results prove the effectiveness of our method over the baseline model. Despite the improvement, all models in our experiments still generate low-quality text from legal AMR. The best-reported score is only 9.31 for BLEU and 21.38 for METEOR, leaving a challenge for research in this domain.

Acknowledgments. This work was supported by JST CREST Grant Number JPMJCR1513, Japan.

References

[1] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer and N. Schneider, Abstract Meaning Representation for Sembanking, in: *Proceedings of*

the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 178–186. https://www.aclweb.org/anthology/W13-2322.

[2] B. Jones, J. Andreas, D. Bauer, K.M. Hermann and K. Knight, Semantics-Based Machine Translation with Hyperedge Replacement Grammars, in: *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 1359–1376.

[3] F. Liu, J. Flanigan, S. Thomson, N. Sadeh and N.A. Smith, Toward Abstractive Summarization Using Semantic Representations, in: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1077–1086. doi:10.3115/v1/N15-1114.

[4] L. Song, Y. Zhang, Z. Wang and D. Gildea, A Graph-to-Sequence Model for AMR-to-Text Generation, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1616–1626. doi:10.18653/v1/P18-1150. https://www.aclweb.org/anthology/P18-1150.

[5] I. Konstas, S. Iyer, M. Yatskar, Y. Choi and L. Zettlemoyer, Neural AMR: Sequence-to-Sequence Models for Parsing and Generation, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 146–157. doi:10.18653/v1/P17-1014.

[6] A. See, S. Roller, D. Kiela and J. Weston, What makes a good conversation? How controllable attributes affect human judgments, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1702–1723. doi:10.18653/v1/N19-1170. https://www.aclweb.org/anthology/N19-1170.

[7] J. May and J. Priyadarshi, Semeval-2017 task 9: Abstract meaning representation parsing and generation, in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 536–545.

[8] S.V. Trong and M.N. Le, An empirical evaluation of AMR parsing for legal documents, *ArXiv* **abs/1811.08078** (2018).

[9] J. Gu, Z. Lu, H. Li and V.O.K. Li, Incorporating Copying Mechanism in Sequence-to-Sequence Learning, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2016, pp. 1631–1640. doi:10.18653/v1/P16-1154. http://aclweb.org/anthology/P16-1154.

[10] A. Fan, D. Grangier and M. Auli, Controllable Abstractive Summarization, in: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 45–54. doi:10.18653/v1/W18-2706. https://www.aclweb.org/anthology/W18-2706.

[11] M. Ghazvininejad, X. Shi, J. Priyadarshi and K. Knight, Hafez: an Interactive Poetry Generation System, in: *Proceedings of ACL 2017, System Demonstrations*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 43–48. https://www.aclweb.org/anthology/P17-4008.

[12] J. Pennington, R. Socher and C.D. Manning, Glove: Global vectors for word representation, in: *In EMNLP*, 2014.

[13] J. Flanigan, S. Thomson, J. Carbonell, C. Dyer and N.A. Smith, A Discriminative Graph-Based Parser for the Abstract Meaning Representation, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 1426–1436. doi:10.3115/v1/P14-1134.

[14] C. Wang, S. Pradhan, X. Pan, H. Ji and N. Xue, CAMR at SemEval-2016 Task 8: An Extended Transition-based AMR Parser, in: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, Association for Computational Linguistics, San Diego, California, 2016, pp. 1173–1178. doi:10.18653/v1/S16-1181.

[15] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040.

[16] M. Denkowski and A. Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, in: *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190331

On Constructing a Knowledge Base of Chinese Criminal Cases

Xiaohan WU [a,1], Benjamin L. LIEBMAN [a], Rachel E. STERN [b], Margaret E. ROBERTS [c] and Amarnath GUPTA [c]

[a] *Columbia Law School, USA*
[b] *University of California Berkeley, USA*
[c] *University of California San Diego, USA*

Abstract. We are developing a knowledge base over Chinese judicial decision documents to facilitate landscape analyses of Chinese Criminal Cases. We view judicial decision documents as a mixed-granularity semi-structured text where different levels of the text carry different semantic constructs and entailments. We use a combination of context-sensitive grammar, dependency parsing and discourse analysis to extract a formal and interpretable representation of these documents. Our knowledge base is developed by constructing associations between different elements of these documents. The interpretability is contributed in part by our formal representation of the Chinese criminal laws, also as semi-structured documents. The landscape analyses utilizes these two representations and enables a law researcher to ask legal pattern analysis queries.

Keywords. landscape analysis, Chinese criminal cases, Information Extraction, discourse analysis, context-sensitive grammar, knowledge representation

1. Introduction

Our long-term goal is to develop a knowledge-based information system that would capture the "general knowledge" about a legal universe and the way law is practised in that universe. We use the term "general knowledge" in the sense that it can maintain enough information to enable a user infer "what usually happens" in a given legal scenario and what makes some case exceptional. For example, the one should be able to infer from the system that no defense argument is usually presented for drunk driving cases, and in an exceptional situation where there is one, only a leniency in the punishment is requested. We call these class of questions *legal landscape analyses*.

Prior Work. The primary corpus for our study is the Judicial Decision Documents (JDD) available from the Supreme People's Court (SPC) [1]. As Gupta et al [2] showed, parts of the data, such as the parties to the lawsuit including the plaintiffs and defendants, together with their legal representation, are represented as structurable text, stored in a relational database. However, [2] did not analyze the unstructured part such as the facts found by the court.

[1]Corresponding Authors: Wu E-mail: xw2510@columbia.edu, Gupta E-mail: a1gupta@ucsd.edu

```
CaseParty [role:Defendant, name: [redacted], position:务工, gender:男, birthday:1968-07-17,
ethnic:汉族, ancestralHome:云南省彝良县, address:彝良县, education: 小学,
lawEnforcementActions:[LawEnforcementAction{, reason: 诈骗罪},
LawEnforcementAction{date: 2016年8月25日, agency: 云南省彝良县人民法院, action: 有期徒刑
(FixedTermImprisonment), duration: 一年}, LawEnforcementAction{, action: 判处罚金(Fine),
fine: 100000万元, }, LawEnforcementAction{date: 2016年12月2日, reason: 诈骗罪, agency: 从
昭通监狱, action: 解回(TransferBack), agency2: 大关县看守所, action2: 羁押(Custody)}], ]
```

Figure 1. The semi-structured output of a party involved in a case.

2. Landscape Analysis of Legal Documents - A First Formal Model

We model a collection **C** of JDDs as a triple $(\mathbf{S}, \mathbf{D}, \mathbf{M})$ where **S** is a heterogeneous relation, **M** is a k-dimensional matrix and **D** is a mapping between elements of **S** and the indices of **M**. Here, a *heterogeneous relation* refers to a relation whose attributes can take different forms of semi-structured values. For example, `case-type` is a string valued (e.g., 'criminal' or 'administrative') attribute, while `parties` is a complex value as shown in Fig. 1. Notice how the parser output includes the criminal history of the defendant under the element `LawEnforcementActions` containing a hierarchy of subelements like the duration of the defendant's imprisonment.

The matrix **M** is derived from our analysis of the text-valued `Fact` element. Using parsing methods described in the next section, sentences in the fact can be classified into 8 classes: case background, arguments from plaintiff/prosecutor, evidences provided from plaintiff/prosecutor, requests/opinions from plaintiff/prosecutor, arguments from defendant, evidences from defendant, reviewed facts from court, and evidences accepted by court. In a typical JDD document, multiple consecutive sentences may belong to each class. The sentences in these sections can be further decomposed into an *action schema* given by `[subject, action, object, action_modifier]`. For example, the sentence (translated) "The defendant surrendered himself at police station in Binjiang on Feb.13th, 2017, where he admitted his crime honestly." has the actions: `['name of defendant', 'went to', 'Binjiang police station','voluntarily']`, `['name of defendant', 'stated', 'criminal action','later','honestly']`. In the sentence (translated)(The total value of stolen items is 25,920 yuan.), the system detects the variable `damage`: ['25,920 yuan'] A similar representation of the court decision leads to a structure of the punishment issued by the court. For criminal cases `punishment` is represented by the numeric vector

{`Exemption(免于刑事处罚)`, `Public Surveillance(管制)`,`Detention(拘役)`, `Fixed-Term Imprisonment(有期徒刑)`, `Probation(缓刑)`, `Fine(罚金)`, `Political Rights Deprivation(剥夺政治权利)`, `Confiscation(没收)`, `Life Imprisonment(无期徒刑)`, `Death(死刑)`, `Political Rights Deprivation For Life(剥夺政治权利终身)`} where `Death`, `Exemption`, `LifeInprisonment`, `PoliticalRightsDeprivationForLife` are represented in binary code and other vector elements are represented by a quantified "degree of punishment" either in terms of time or in terms of monetary value.

The representation enables us to represent more than one punishment (e.g., prison time and fine) for a crime. Integrity constraints are applied to ensure that specific combinations of punishments (e.g., `FixedTermImprisonment` and `lifeImprisonment`) do not co-occur. We construct the matrix **M** as a product `action` $\times$ `damage` $\times$ `punishment-bucket` where a `punishment-bucket` is a discretized representation of

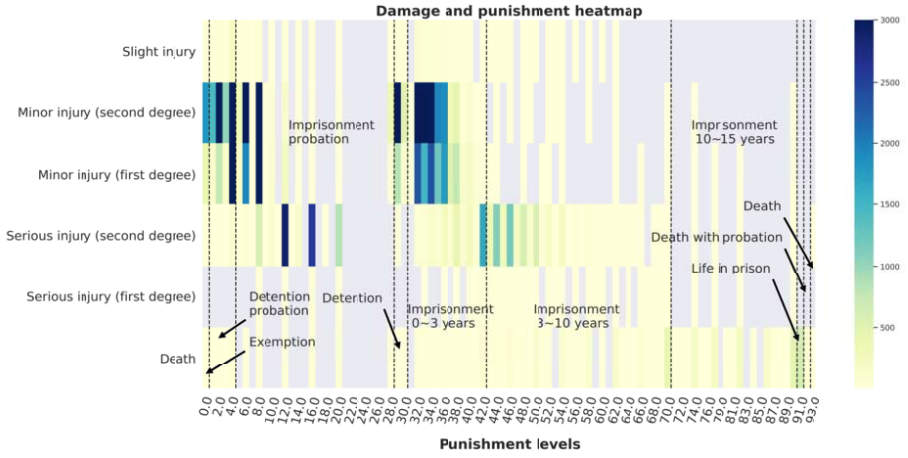


Figure 2. Damage and punishment heat map for assault and battery cases

the punishments. A cell of the matrix represents the number of cases that fall in the action-damage-punishment construct. **M** is partitioned by crime type so that theft is considered separately from murder. While this partitioning introduces some inaccuracy for cases where multiple crimes occur, we tolerate the inaccuracy for landscape analyses where the goal is to understand general properties of the distribution. Figure 2 shows a fragment of this matrix as a heatmap. Note that the color in this map indicates the number of cases for the corresponding combination. Gray means zero case. The unit for punishment levels is 3 months except for *Exemption, life in prison, death with probation and death penalty*, each of which takes one unit. Figure 2 shows how some combination of damages and punishment are more dense while some other combinations are empty, indicating combinations that although theoretically plausible occur rarely in practice. For example, according to Criminal law article 234, "whoever intentionally inflicts injury upon another person,causing severe injury to another person, shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years". However, in practice, many assaulters were sentenced to fixed-term imprisonment of less than three years with probation – indicating judges' discretion in deciding punishments.

The mapping **D** between **S** and **M**, which is used for information retrieval, is a collection of indices. The forward indices serve as a pointer from a schema element like: `JDD.prosecutorArgument.sentence.actions.drunk-driving` to **M**`.traffic-misconduct[3]` where [3] indicates the axis of the matrix where `drunk-driving` is mapped. Similarly, `JDD.prosecutorArgument.sentence.drunk-driving.punishment` may map to **M**`.traffic-misconduct[3][2]` which is the action-punishment slice of the `traffic-misconduct` partition of **M**. In contrast, the reverse index behaves similarly as an inverted index in an information retrieval system where every cell of the matrix is mapped back to a list of case identifiers that populate the cell. Thus, the retrieval function `getCases(`**M**`[3][2][4])` will retrieve the drunk driving cases resulting in property damage up to 1000 yuan where a fine was imposed.

3. Information Extraction

To extract our analytical primitives, we have developed a parsing strategy for linguistic patterns that are characteristically observed in JDDs. The information extraction module assumes that the names of plaintiffs, defendants and their legal counsel are available to the system. In the following, we present a method for extracting the "action" part from the unstructured `Facts` of a JDD. The linguistic patterns observed include:

Long flowing sentences. The flowing sentence is a unique sentence pattern in Chinese. It contains so-called 链式结构(chain structure) – the relationship between 逗断(dòuduàn) was usually indicated by the order of events. Wang [3] defined dòuduàn as the basic unit of Chinese text and dòuduàn can be used as the index to specific communication event. We use dòuduàn as the minimum text processing unit for parsing and discourse analysis to reduces computation and improves parsing accuracy rate [4].

Action-focused defendant-centered description. The majority of sentences in facts, especially arguments from prosecutor and reviewed facts, are descriptions of actions. Even if the description is in passive voice, the subject of an action is usually the defendant. For example, 'The defendant has already obtained the victim's families' forgiveness.' is more common than 'The victim's family has already forgiven the defendant.'

Extracting action triggers. Verbs have been used as triggers in open information extraction [5,6] and news events extraction [7]. These relation patterns, however, is only applicable to English text. Open information extraction research in Chinese is still relatively inadequate[8]. We extract central actions where the subjects are the defendant or the police using the following rules for trigger verb extraction.

1. *Rule 1.* verbs in paths that originated from ROOT in constituency tree and only contains {*'IP','VP','VV','VRD'*}
2. *Rule 2.* verbs that are {*'conjunct','clausal complement'*} dependents of trigger verbs obtained by Rule 1.

For example, in dòuduàn 被告人在15号车厢当面接收张某某发送的手机微信红包(The defendant received Wechat red pockets sent by Zhang in person in car No.15), part-of-speech tagging identified two verbs: 接收(receive) and 发送(send). The central action in this dòuduàn is, [['The defendant'], 'receive', ['wechat red pocket'], ['in person']]. Therefore, the trigger verb is "receive" rather than "send" by *Rule 1*.

Extracting elements of actions. In addition to action trigger verb, we defined *Subject, Object* and *action_modifier* in *action schema*. We extracted these elements based on universal dependencies (a multiliguial generalization of the dependency relationships from the Stanford Dependency parser) of trigger verbs:

- *Subject* extraction has two rules: *Rule 1* extracts nouns that are *'nominal subject'* of the trigger verb. *Rule 2* inherits *Subject* from the latest dòuduàn if *Rule 1* fails.
- *Objects* are usually *direct objects* of trigger verbs. Note that dòuduàn containing '被','将' and '把'are treated as exceptions.
- *action_modifier* are trigger verb's *adverb modifier*. We also excluded *(*遂,并,且,后,但*)* because they turned out to be less important in our landscape analysis.

Extract damages, criminal charges, convicted crime charges and punishments. We extract monetary damages by applying named entity recognition(NER). There are five

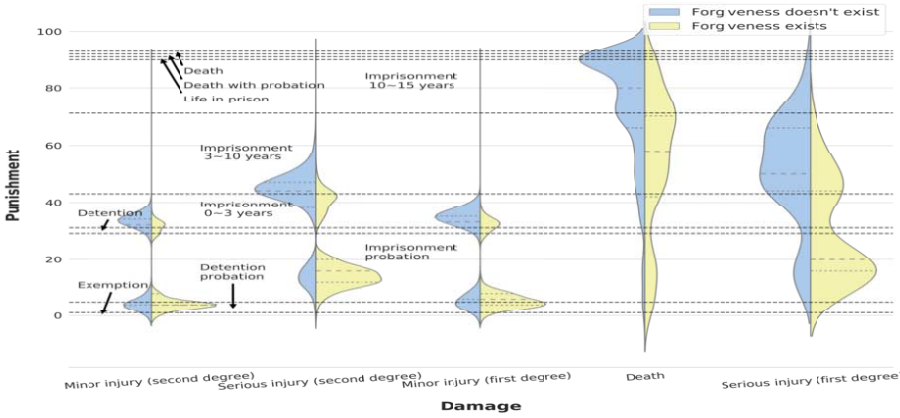


Figure 3. Probability density of punishment levels for battery cases with/without victims' forgiveness

injury levels in Chinese legal system. Since the injury levels are fixed and finite, we extracted human health damages by keyword matching. We use regular expression to extract the name of 469 crimes and convert extracted crime names to standard names to eliminate variations. Since the decision part of criminal cases is more structured, we chose the extraction keywords according to the principal and supplementary punishments in Chinese criminal law Article 33.

4. Answering Analytical Questions

Question 1. What is the distribution of punishments for cases where the defendant received the victims'(or victim families') forgiveness versus where they did not, conditioned by the damage caused by the crime?

We define $C1$ as a subset of cases where the action includes a lemmatized version of the term "forgiveness" with positive *action_modifier* and $C2$ where the cases do not. $C1$ contains 75655 battery cases while $C2$ contains 60627 cases. In Figure 3, the yellow part is probability density of punishments for cases where forgiveness exist while blue part is for cases where forgiveness don't exist. Evidently, judges tend to give lenient punishments to defendants who received forgiveness regardless of the damage severity.

Question 2. What punishments are rare for crime type X. Find the distribution of circumstances for which the punishment is "exemption". Here, we specify a "circumstance" as a combination of crime types, actions and damages. The steps of query evaluation are: (i) P = getMarginals($\mathbf{M}.X$, 'punishment'), (ii) C = getMarginals($\mathbf{M}.X$, 'punishment'='exemption'), (iii) C' = top-k(C, 20)

We set `case type = ''battery''`. In step (i), We found two types of rare punishments – punishments that are extremely lenient or harsh and punishments where the measurement unit is not a quarter of a year. Notice the C is a 2D histogram with axes action and damage. C' returns a fraction of C that only contains k most important actions defined by user – 20 most frequent action-damage pairs by default. We obtained 2,181 battery cases where defendants were exempted from criminal punishments and 2,033 actions associated with these cases. The importance score for each action is action frequency in C divided by action frequency in $\mathbf{M}.battery$. High exclusiveness can also lead to error actions that had very low frequency in both C and $\mathbf{M}.battery$. So we take

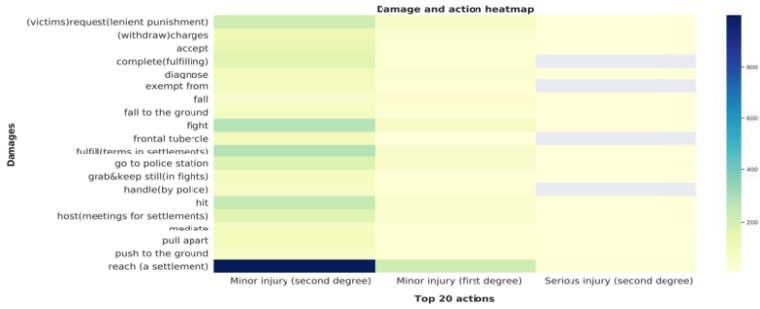


Figure 4. Heat map of damage and top 20 actions in battery cases

5% most frequent actions and select 20 most important actions according to importance score. Figure 4 is the co-occurrence-heat-map of damages and selected actions. This heat map shows that reaching settlements and fulfilling the terms for minor injuries before trial is a key factor for receiving exemption from punishments.

5. Conclusion and Future Work

In this paper, we have sketched our approach to developing a knowledge-base to answer landscape questions revealed by judicial decision documents from Chinese courts. Unlike a facts-and-rules or a graph-based knowledge representation system, we have opted to use heterogeneous relation, a distribution matrix and a mapping between them as our knowledge structure, and showed its usefulness in answering questions. Yet, our representation has taken some simplifying decisions that failed to capture some of the practical nuances of criminal law. In future work, we will refine our representation to accommodate further levels of punishment and action granularity.

Acknowledgment. We acknowledge the NSF RIDIR grant 1738411 for funding.

References

[1] B.L. Liebman, M. Roberts, R.E. Stern and A.Z. Wang, Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law, *S. Science Research Network Collection* (2017).

[2] A. Gupta, A.Z. Wang, K. Lin, H. Hong, H. Sun, B.L. Liebman, R.E. Stern, S. Dasgupta and M.E. Roberts, Toward Building a Legal Knowledge-Base of Chinese Judicial Documents for Large-Scale Analytics, in: *Proc. of Int. Conf. on Legal Knowledge and Information Systems (JURIX)*, 2017, pp. 135–144.

[3] H. Wang and R. Li, On the basic Unit of Chinese Texts and the Causes of the Flowing Sentences, *Essays on Linguistics* (2014), 11–40.

[4] X. Li, C. Zong and R. Hu, A hierarchical parsing approach with punctuation processing for long Chinese sentences, 2005.

[5] A. Fader, S. Soderland and O. Etzioni, Identifying relations for open information extraction, in: *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2011, pp. 1535–1545.

[6] N.G. Silveira, Designing Syntactic Representations for NLP: An Empirical Investigation, PhD thesis, Stanford University, 2016.

[7] D. Rusu, J. Hodson and A. Kimball, Unsupervised techniques for extracting and clustering complex events in news, in: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2014, pp. 26–34.

[8] Q. Bing, L. Anan and L. Ting, Unsupervised Chinese open entity relation extraction, *Journal of Computer Research and Development* **52**(5) (2015), 1029–1035.

Demo Papers

This page intentionally left blank

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190333

The NAI Suite – Drafting and Reasoning over Legal Texts

Tomer LIBAL [a] and Alexander STEEN [b]
[a] *American University of Paris*
[b] *University of Luxembourg*

Abstract. A prototype for automated reasoning over legal texts, called NAI, is presented. As an input, NAI accepts formalized logical representations of such legal texts that can be created and curated using an integrated annotation interface. The prototype supports automated reasoning over the given text representation and multiple quality assurance procedures.

Keywords. Legal Reasoning, Deontic Logic, Automated Reasoning

1. Introduction

Computer systems are playing a substantial role in assisting people in a wide range of tasks, including search in large data and decision-making; and their employment is progressively becoming vital in an increasing number of fields. One of these fields is *legal reasoning*: New court cases and legislations are accumulated every day. In addition, international organizations like the European Union are constantly aiming at combining and integrating separate legal systems [1]. In contrast to this situation, the automation of legal reasoning is still underdeveloped albeit being a growing field of research. Approaches for automatic reasoning over sets of norms have been developed, such as for business [2] and GDPR compliance [3].

One of the reasons for the relatively restricted number of applications of automated reasoning to the legal domain is the lack of editing tools which can be used by non-logicians. Indeed, the applications mentioned above are mainly based on the work of logicians. The most popular approach to the formalization of legal texts is by using a logic programming language. This approach has enjoyed success [4] in the last 40 years and is still popular today. Nevertheless, basic knowledge of logic programming is still required. In order to have a wider use of legal reasoning, other professionals, such as lawyers and jurists, should be able to use the tools.

A second reason is the lack of tools and methodologies for asserting the correctness of the logical representations of the legal texts. Among existing results, one can find a methodology for building legal ontologies [5] and more concretely to our approach, one for validating formal representations of legal texts [6].

Lastly, the scarcity of legal reasoning software prevents the utilization of such formalizations, even if proven correct. One can mention here the engines for defeasible log-

ics [7], Higher-order logics [8] Deontic logics with contrary-to-duty obligations [9], as well as logic programming [4].

In this paper we describe the new normative reasoning framework NAI, which addresses these problems by providing functionality and methodology for lawyers and jurists. NAI is a web application and is readily available at https://nai.uni.lu. NAI is also open-source, its source code is freely available at GitHub[1] under GPL-3.0 license.

NAI features an annotation-based editor which abstracts over the underlining logical language. It also contains an easily accessible functionality for ensuring that the formalization is consistent and that the formalized sentences are independent from each other. NAI also supports a methodology for proving the correctness of formalizations via execution of behavioral tests. Lastly, it provides an interface for the creation of queries and for checking their validity.

The architecture of NAI is modular, which allows using different logics and reasoning engines. It also provides an API, which can be used by other tools in order to reason over the formalized legislation.

In this paper we give a technical description of a new tool for legal formalization and reasoning which utilizes an innovative annotations interface. An example of its usage and a demonstration of a new Agile methodology for formalizing legal texts are presented in the full paper [2] and can be seen on a demo account [3].

2. The NAI Suite

The NAI suite integrates novel theorem proving technology into a usable graphical user interface (GUI) for the computer-assisted formalization of legal texts and applying automated normative reasoning procedures on these artifacts. In particular, NAI includes

1. a legislation editor that graphically supports the formalization of legal texts,
2. means of assessing the quality of entered formalizations, e.g., by automatically conducting consistency checks and assessing logical independence,
3. ready-to-use theorem prover technology for evaluating user-specified queries wrt. a given formalization, and
4. the possibility to share and collaborate, and to experiment with different formalizations and underlying logics.

NAI is realized using a web-based Software-as-a-service architecture. It comprises a GUI that is implemented as a Javascript browser application, and a NodeJS application on the back-end side which connects to theorem provers, data storage services and relevant middleware. Using this architectural layout, no further software is required from the user perspective for using NAI and its reasoning procedures, as all necessary software is made available on the back end and the computationally heavy tasks are executed on the remote servers only. The results of the different reasoning procedures are sent back to the GUI and displayed to the user. The major components of NAI are described in more detail in the following.

[1] See https://github.com/normativeai.
[2] http://arxiv.org/abs/1910.07004
[3] Please login to https://nai.uni.lu using Email address: smoking@nai.lu / Password: nai

2.1. The Reasoning Module

The NAI suite supports formalizing legal texts and applying various logical operations on them. These operations include consistency checks (non-derivability of falsum), logical independence analysis as well as the creation of user queries that can automatically be assessed for (non-)validity. After formalization, the formal representation of the legal text is stored in a general and expressive machine-readable format in NAI. This format aims at generalizing from concrete logical formalisms that are used for evaluating the logical properties of the legal document's formal representation.

There exist many different logical formalisms that have been discussed for capturing normative reasoning and extensions of it. Since the discussion of such formalisms is still ongoing, and the choice of the concrete logic underlying the reasoning process strongly influences the results of all procedures, NAI uses a two-step procedure to employ automated reasoning tools. NAI stores only the general format, as mentioned above, as result of the formalization process. Once a user then chooses a certain logic for conducting the logical analysis, NAI will automatically translate the general format into the specific logic resp. the concrete input format of the employed automated reasoning system. Currently, NAI supports only the Deontic logic described in [9]; however, the architecture of NAI is designed in such a way that further formalisms can easily be supported.

The current choice of Deontic logic is primarily motivated by the fact that it can be effectively automated using a shallow semantical embedding into normal (bi-)modal logic [9]. This enables the use of readily available reasoning systems for such logics; in contrast, there are relatively few dedicated automated normative reasoning systems such as the one described in [7]. In NAI, we use the MleanCoP prover [10] for first-order multi-modal logics as it is currently one of the most effective systems and it returns proof certificates which can be independently assessed for correctness [11]. It is also possible to use various different tools for automated reasoning in parallel (where applicable). This is of increasing importance once multiple different logical formalisms are supported.

2.2. The Annotation Editor

The annotation editor of NAI is one of its central components. Using the editor, users can create formalizations of legal documents that can subsequently used for formal legal reasoning. The general functionality of the editor is described in the following. A more detailed exemplary application on a concrete legal document is presented in the demo.

One of the main ideas of the NAI editor is to hide the underlying logical details and technical reasoning input and outputs from the user. We consider this essential, as the primary target audience of the NAI suite are not necessarily logicians and it could greatly decrease the usability of the tool if a solid knowledge about formal logic was required. This is realized by letting the user annotate legal texts and queries graphically and by allowing the user to access the different reasoning functionalities by simply clicking buttons that are integrated into the GUI. Note that the user can still inspect the logical formulae that result from the annotation process and also input these formulae directly. However, this feature is considered advanced and not the primary approach put forward by NAI.

The formalization proceeds as follows: The user selects some text from the legal document and annotates it, either as a term or as a composite (complex) statement. In the

first case, a name for that term is computed automatically, but it can also be chosen freely. Different terms are displayed as different colors in the text. In the latter case, the user needs to choose among the different possibilities (which roughly correspond to logical connectives) and the containing text can be annotated recursively. Composite statements are displayed as a box around the text.

The editor also features direct access to the consistency check and logical independence check procedures (as buttons). When such a button is clicked, the current state of the formalization will be translated and sent to the back-end provers, which determine whether it is consistent resp. logically independent.

User queries are also created using such an editor. In addition to the steps sketched above, users may declare a text passage as *goal* using a dedicated annotation button, whose contents are again annotated as usual. If the query is executed, the back-end provers will try to prove (or refute) that the goal logically follows from the remaining annotations and the underlying legislation.

2.3. The Abstract Programming Interface (API)

All the reasoning features of NAI can also be accessed by third-party applications. The NAI suite exposes a RESTful (Representational state transfer) API which allows (external) applications to run consistency checks, checks for independence as well as queries and use the result for further processing. The exposure of NAI's REST API is particularly interesting for external legal applications that want to make use of the already formalized legal documents hosted by NAI. A simple example of such an application is a tax counseling web site which advises its visitors using legal reasoning over a formalization of the relevant tax law done in the NAI suite.

References

[1] A.-M. Burley and W. Mattli, Europe before the court: a political theory of legal integration, *International organization* **47**(1) (1993), 41–76.

[2] M. Hashmi and G. Governatori, Norms modeling constructs of business process compliance management frameworks: a conceptual evaluation, *Artif. Intell. Law* **26**(3) (2018), 251–305. doi:10.1007/s10506-017-9215-8.

[3] M. Palmirani and G. Governatori, Modelling Legal Knowledge for GDPR Compliance Checking, in: *Legal Knowledge and Information Systems: JURIX*, Vol. 313, IOS Press, 2018, pp. 101–110.

[4] T.J.B.-C. et al., Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation, in: *Proceedings of ICAIL*, ACM, 1987, pp. 190–198.

[5] M. Mockus and M. Palmirani, Legal ontology for open government data mashups, in: *2017 Conference for E-Democracy and Open Government (CeDEM)*, IEEE, 2017, pp. 113–124.

[6] C. Bartolini, G. Lenzini and C. Santos, An interdisciplinary methodology to validate formal representations of legal text applied to the GDPR, in: *JURISIN*, 2018.

[7] G. Governatori and S. Shek, Regorous: a business process compliance checker, in: *Proc. of the 14th Int. Conf. on Artificial Intelligence and Law*, ACM, 2013, pp. 245–246.

[8] C. Benzmüller, A. Farjami, P. Meder and X. Parent, I/O Logic in HOL, *IfCoLoG Journal of Logics and their Applications* **6**(5) (2019), 715–732.

[9] T. Libal and M. Pascucci, Automated reasoning in normative detachment structures with ideal conditions, in: *Proc. of ICAIL*, 2019, pp. 63–72. doi:10.1145/3322640.3326707.

[10] J. Otten, MleanCoP: A Connection Prover for First-Order Modal Logic, in: *7th International Joint Conference, IJCAR*, 2014, pp. 269–276. doi:10.1007/978-3-319-08587-6_20.

[11] J. Otten, Implementing Connection Calculi for First-order Modal Logics., in: *IWIL@ LPAR*, 2012, pp. 18–32.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190334

Facts2Law – Using Deep Learning to Provide a Legal Qualification to a Set of Facts

Ivan MOKANOV[a1]
[a] *Lexum*

Abstract. Over the course of the last year Lexum has started exploring the potential of deep learning (DL) and machine learning (ML) technologies for legal research. Although these projects are still under the umbrella of Lexum's research and development team (Lexum Lab, https://lexum.com/en/ailab/), concrete applications have recently started to become available. This demo focuses on one of these applications: Facts2Law. The project benefits from a combination of factors. First, the millions of legal documents available in the CanLII database in parsable format along with structured metadata constitute a significant dataset to train AI algorithms. Second, Lexum has direct access to the knowledge and experience of one of the leading teams in AI and deep learning worldwide at the Montreal Institute for Learning Algorithms (MILA) of the University of Montreal. Third, the availability of computer engineers with cutting-edge expertise in the specifics of legal documents facilitates the transition from theory to practical applications. Regarding concrete outcomes, Lexum's Facts2Law can predict the most relevant sources of law for any given piece of text (incorporating legal citations or not).

Keywords. Deep learning, embeddings, citation network, case law

1. CanLII as a Training Data Set

CanLII is the largest repository of Canadian public legal information. The CanLII database includes:

- Data from 14 jurisdictions – federal, provinces and territories; More than 2M decisions from Canadian courts and tribunals. The length of these decisions varies greatly; their average is around ten pages;
- Around 650 statutes and a 2,000 regulations per jurisdiction, the vast majority of which have a table of contents represented as a tree whose leaves are numbered legislative "sections";Item
- A total of over 30,000 cited sections and subsections of statutes and regulations; Over 8.7M citations from one decision to another;
- Over 8M citations from decisions to a statute's section or subsection;
- All citations are hyperlinked and already extracted to a precision close to 90%;
- The citations and table of contents are encoded in a standardized fashion.

[1] Email: ivan@lexum.com

Lexum legal citator, Reflex, uses text analysis and probabilities to establish the associations between parallel citations. In the case of historical material, citation patterns from all major Canadian printed reports are supported. For more recent content, electronic citations patterns are supported (including the neutral citation, QL, Carswell, JE (Jurisprudence Express), Azimut, and REJB - EYB (Yvon Blais)). Reflex makes uses of parallel citations to expand the citation network of CanLII. For example, if Judge A cites the case X v. Y with its DLR1 and CCC2 citations, and Judge B cites the same case with its DLR1and OAC6 citations, Reflex will conclude that DLR1, CCC2, and OAC6 are all parallel citations of the same case. Reflex also use powerful computer heuristics – rules of thumb either manually coded or acquired through statistical analysis – to recognize oft-used citation patterns. Citations that are ambiguous by themselves can then be deciphered by clues given by their context. For example, if a full citation of the Immigration and Refugee Protection Act is detected at the beginning of a paragraph, a later mention of "the Act, at section ##" can be inferred to relate to the same statute.

Thanks to its citator, the CanLII database already includes a "map" of the Canadian Law. All of this data is already highly structured and available to train ML algorithms.

2. Facts2Law - Predicting Legal Citations

Lawyers regularly produce briefs, legal opinions, and other types of legal advice documents. These opinions, provided in writing, examine the various legal aspects of the client's situation. A lawyer who wants to research an aspect of that situation will often perform full-text search queries to buttress his opinion. Unfortunately, full-text search queries are often limited in their scope and might miss some nuances of the situation. To remedy this, some systems provide search results based on the text of the whole opinion.

"More like this" systems are nothing new but they are typically bag of words affairs that are somewhat limited in their understanding of the content. In addition, legal databases are made up of words but also of citations, which are reliable indicators of popularity and authority, measures that are not considered in traditional "more like this" approaches.

Lexum's solution to this challenge is to learn from the existing citations on CanLII to predict which sources of law are relevant to the text of a legal brief, a legal opinion or to the plain language description of a legal issue.

This approach makes it possible to enhance content by providing additional contextual information. It also enables legal researchers to search the law in an entirely new way, by describing their problem in plain language. The results obtained will constitute a good starting point in the sorting of issues and the subsequent exploration of the applicable rules.

While there is no training data for this exact question in the context of the CanLII database, citations in case law can act as stand-ins for the question of relevance. First, court decisions are opinions with content and form very similar to legal opinions and briefs. Second, the citations that appear in these court decisions are typically the result of a research and relevance evaluation by a human being in the course of his research. If a document is cited, it is because it is relevant. If a document is not relevant enough to be discussed, then it will not be cited.

Lexum has developed a preliminary iteration of Facts2Law. It can currently be interrogated via a simple web-based interface. As a baseline control the current tool makes use of the Logistic Regression for ML (https://machinelearningmastery.com/logistic-regression-for-machine-learning/), a technique borrowed from the field of statistics.

The production instance makes use of a Neural Networks and Deep Learning algorithm.

This approach takes as input both a brief and a target document and, as output, its confidence that the brief should cite the target. To do so, Lexum uses whole document embeddings of both documents and a weighted summary of those in neighbouring nodes in the citation graph along with the relevant metadata. Then, the approach consists in feeding both branches to fully connected layers, merge them and modulate the merged inputs to take into account the age of the brief before passing them to another fully connected layer that outputs the prediction. This architecture allows us to use heuristics to select a subset of the corpus that could be relevant to a document and then rank them efficiently to extract documents of interest.

3. Embeddings

One of the key components of the citation predictor project is the ability to reduce a whole document to only a fixed series of numbers (1000 in our case) which in AI language is called an *embedding*. The quality of the embeddings will have a considerable effect on our results, and as such we are constantly on the lookup for improvements and developments in this field.

We started in 2018 with Doc2Vec, which is based on Word2Vec. The idea is to basically learn the embeddings of individual words and average them out to get the embedding of a document. In 2019, new technologies became available: BERT, Transformer XL and XL Net. Although BERT is an improvement over Doc2Vec, the resources (computing power) needed to generate the embeddings as well as the small document size (512 tokens) makes it less optimal. Transformer XL, on the other hand, requires much less processing and can work on much larger text sequences. Transformer XL and Doc2Vec are two completely different strategies. The major difference is that Transformer XL will use the context (from left to right) of a word to predict its embedding. As such, it will be able to give a different semantic meaning for words that have multiple meanings (like fly, it can be an insect or a verb).

XL Net is very similar to Transformer XL, however, it does not take into account the order of the words (while Transformer XL will consider the context in exact order from left to right). This has shown to lead to even better results. We are currently evaluating the performance of these two strategies versus our original Doc2Vec, we expect to see some significant gains, but this has yet to be confirmed.

4. Applications

The demo will focus on a specific way in which this approach can be used to enhance legal research. In Canada, administrative tribunal decisions, for example, are very factual and do not usually contain many references to other cases, as opposed to judicial decisions, which cite other decisions abundantly. Considering this drafting

pattern in administrative decisions, citation parsing algorithms are of little use when we try to identify other cases of interest, because such cases are simply not referred to. Using the approach of Facts2Law, we can nevertheless identify pertinent cases which could have been cited, if the decision-maker had chosen to cite other relevant case law.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190335

ANOPPI: A Pseudonymization Service for Finnish Court Documents

Arttu OKSANEN [a,c], Minna TAMPER [a], Jouni TUOMINEN [a,b]
Aki HIETANEN [d], and Eero HYVÖNEN [a,b]

[a] *Semantic Computing Research Group (SeCo), Aalto University, Finland* [1]
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki*
[c] *Edita Publishing Ltd.*
[d] *Ministry of Justice, Finland* [2]

Abstract. To comply with the EU General Data Protection Regulation (GDPR) publishing court judgments online requires that personal data contained in them must be disguised. However, anonymizing the documents manually is a costly and time-consuming procedure. This paper presents ANOPPI service for automatic and semi-automatic pseudonymization of Finnish court judgments. Utilizing both statistics- and rule-based named entity recognition methods and morphological analysis, ANOPPI is able to automatically pseudonymize documents written in Finnish preserving their readability and layout. The service is currently still in development but pilot tests are going to be carried out in Finnish courts in 2020.

Keywords. automatic pseudonymization, case law, named entity recognition

1. Introduction

Publishing court decisions openly on the web, either as human-readable documents or machine-readable data, enhances the legal protection of citizens by making the administration of justice more transparent. Open electronic access to case law can also be useful to decision-making and research concerning legal practice. In Finland, case law is published publicly online as HTML documents in the Finlex data bank[3] [1] and as linked open data in the Semantic Finlex service[4] [2].

Unfortunately, due to issues of data protection and privacy and the requirement to comply with the EU General Data Protection Regulation (GDPR), currently only a minor part of all the Finnish court judgments is published online. For example, currently none of the judgments of the district courts are available. Publishing court judgments online requires that the documents are pseudonymized so that identifying named entities appearing in the document, such as persons,

[1] Corresponding Author: firstname.lastname@aalto.fi
[2] Corresponding Author: firstname.lastname@om.fi
[3] `http://www.finlex.fi`
[4] `http://data.finlex.fi`

companies and geographical locations, are replaced with referent identifiers. However, currently all of the pseudonymization work is done manually in the courts by experts which is costly and time-consuming. Therefore a tool that automates the process of pseudonymization is highly desired.

This paper presents Anoppi, a web service for semi-automatic pseudonymization of documents written in Finnish. In the on-going project, we are focusing on case law documents as a first use case. However, the purpose of the Anoppi service is to be a general-purpose domain-agnostic pseudonymization tool. The service is currently still under development, but a first demonstrator has already been created, and pilot tests will be carried out in Finnish courts in 2020. The source code will be published with an open license once the service is ready to be brought into real use. We will discuss the underlying ideas of the service as well as the first demonstrator in more detail in the following sections, starting with a description of the pseudonymization method in Section 2, followed by an overview of the application user interface in Section 3, and finally concluding with related work and discussion in Section 4.

2. Pseudonymization Method

The court orders are available in electronic format either as plain text, XML, HTML, or DOCX files. Based on [3], we have developed a tool that is able to find the named entities from these documents and annotate the occurrences of the named entities with special tags. The tool can be used as a RESTful web service that takes as input the document and produces as output the annotated document with a separate list of all the named entities found in the document.

To find the named entities the tool uses multiple different named entity recognizers and combines the results from those. First of all we use ready-made statistics- and rule-based named entity recognition (NER) software such as FiNER[5], a rule-based named entity recognizer for Finnish language, and Stanford NER [4]. Secondly, we have developed our own set of regular expression patterns to recognize things such as vehicle registration plates and property identifiers. In addition, we use an all-inclusive Finnish person name ontology that is based on the open data published by the Population Register Centre[6] to look up person names appearing in the court cases. Finally, we use the Finnish dependency parser [5] to support deciding if a term appearing in the text is a name.

After finding the named entities and their occurrences in the text the occurrences are replaced with pseudonyms. To assign a reasonable pseudonym for a given named entity its category must also be resolved. For example, we must be able to differentiate towns and corporations so that a pseudonym can be correctly determined as either "town A" or "corporation A". Categorical disambiguation is based on a scoring scheme that weighs the results obtained from the different named entity recognizers.

As Finnish is a highly inflected language we must also derive the correct inflected form for the pseudonym so that the pseudonymized text stays readable. To

[5] https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md
[6] https://vrk.fi/en/

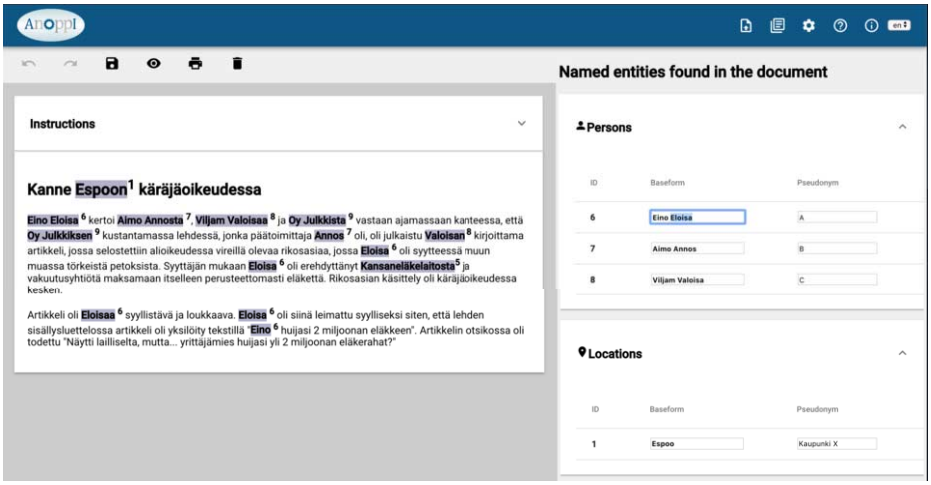


Figure 1. User interface of the ANOPPI application.

achieve this, we use morphological analysis to be able to distinguish, for example, the case and possessive suffix of a noun.

3. User Interface

As we do not expect the result of the automatic pseudonymization to be perfect, a web-based user interface, shown in Figure 1, is provided where the user of the service can make further modifications to the proposed named entities and their pseudonyms. The text with identified entities is shown on the left. On the right hand column, the user interface allows the user to add new named entities and also edit and remove the existing ones. In addition, it is possible to remove entire phrases from the text if de-identification requires it. Once the editing is complete, the user can preview and export the resulting document that aside from the pseudonyms and text removals should be identical to the original one.

Evaluation of the user interface is underway by usability tests, first within the project team and later in 2019 and 2020 in the courts where the service is eventually going to be brought into use.

4. Related Work and Discussion

Automatic or computer-aided pseudonymization is already utilized in judiciaries of various European countries [6]. As an example, in Denmark an anonymization tool for court orders was implemented using solely manually crafted grammar rules to find the named entities in the texts [7]. On-going development projects similar to ours, in which the focus of the automatic pseudonymization is on court orders and where machine learning-based methods are used, are being carried out in France and Austria[7].

[7]Based on oral presentations at `https://eu2019.fi/en/events/2019-09-05/workshop-anonymisation-of-court-judgements-challenges-and-solutions`

For the moment, the Finnish public sector utilizes hardly at all automatic anonymization or pseudonymization tools, and it is difficult to evaluate the sufficiency of de-identification for different types of data and requirements [8]. ANOPPI aims to change the situation by enabling organizations to deploy automatic pseudonymization in their processes cost-effectively using an open source solution. However, the usefulness of the ANOPPI service will eventually be largely dependent on the precision and recall of the NER methods as well as the applicability of the user interface. Edita Publishing Ltd. has previously estimated that on average it takes approximately 38 minutes to pseudonymize a precedent of the Supreme Court manually. In order for ANOPPI to be successful, pseudonymization of the precedents using the service should be more efficient.

Acknowledgments. This work was funded by the the Ministry of Justice in Finland. We thank Saara Packalén, Tiina Husso, and Oili Salminen of the Ministry of Justice, and Risto Talo, Jari Linhala, and Sari Korhonen of Edita Publishing Ltd. for collaboration. CSC – IT Center for Science, Finland, provided us with computational resources.

References

[1] A. Hietanen, Free Access to Legislation in Finland: Principles, Practices and Prospects, in: *Law via the Internet. Free Access – Quality of Information – Effectiveness of Rights*, M. Ragona, ed., European Press Academic Publishing, Florence, 2009.

[2] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen and E. Hyvönen, Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web, in: *Knowledge of the Law in the Big Data Age*, G. Peruginelli and S. Faro, eds, Frontiers in Artificial Intelligence and Applications, Vol. 317, IOS Press, 2019, pp. 212–228.

[3] M. Tamper, E. Hyvönen and P. Leskinen, Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research, in: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019)*, Springer-Verlag, 2019, Forthcoming.

[4] J. Rose Finkel, T. Grenager and C. Manning, Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, in: *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2005, pp. 363–370.

[5] K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski and F. Ginter, Building the essential resources for Finnish: the Turku Dependency Treebank, *Language Resources and Evaluation* **48** (2014), 493–531, Open access. doi:10.1007/s10579-013-9244-1.

[6] M. van Opijnen, G. Peruginelli, E. Kefali and M. Palmirani, On-Line Publication of Court Decisions in the EU: Report of the Policy Group of the Project 'Building on the European Case Law Identifier', 2017, Available at SSRN: https://ssrn.com/abstract=3088495.

[7] C. Povlsen, B. Jongejan, D.H. Hansen and B.K. Simonsen, Anonymization of Court Orders, in: *11th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, Las Palmas, Spain, 2016. doi:10.1109/CISTI.2016.7521611.

[8] A. Bäck and J. Keränen, Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot, 2017, Liikenne- ja viestintäministeriön julkaisuja 7/2017. `http://urn.fi/URN:ISBN:978-952-243-503-3`.

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

Subject Index

Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

Author Index