Bilingual Dataset for Information Retrieval and Question Answering over the Spanish Workers Statute

Pablo Calleja
Ontology Engineering Group
Universidad Politécnica de Madrid
pcalleja@fi.upm.es

Patricia Martín Chozas
Ontology Engineering Group
Universidad Politécnica de Madrid
pmchozas@fi.upm.es

Elena Montiel-Ponsoda
Ontology Engineering Group
Universidad Politécnica de Madrid
emontiel@fi.upm.es

Víctor Rodríguez-Doncel
Ontology Engineering Group
Universidad Politécnica de Madrid
vrodriguez@fi.upm.es

Elsa Gómez
Cuatrecasas

Pascual Boil
Cuatrecasas

***Abstract*—A bilingual dataset of questions and answers over a key document in Spanish labor law legislation, the Workers Statute, is presented. The document contains 150 questions and their respective answers in the form of one part number from the 130 parts in which the Workers Statute is divided (articles and other provisions), and with the most relevant excerpt of information for the answer. A simple system to retrieve the answers using standard technologies is also described, providing baseline numbers for accuracy in this task. This dataset may be of interest for researchers in the area of Q&A over legal documents. Both the question and answers are available in English and Spanish languages.**

***Index Terms*—Dataset, Q&A, Information Retrieval, Labor Law, Knowledge Graph**

I. Introduction

Information Retrieval and Question Answering have become core tasks in the so called knowledge-based society we live in. Search engines are our best allies when searching for information, regardless of topic or level of expertise. As described by Manning [5], Information Retrieval (IR) used to be an activity that only a few people engaged in such as reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval when they use a web search engine.

In the legal domain, Information Retrieval and Question Answering take a substantial part in the daily work of lawyers when performing domain-specific searches due to the large amounts of text-based information generated in this area. Such a specific area relies heavily on the use of the appropriate terms and the relations between them as established in the search queries.

However, research improvements could not be possible if there are not evaluations to measure the performance of the new techniques, algorithms and systems. Particularly in the legal domain, there are a lot of scenarios in which particular use cases have to be achieved. Moreover, legal information is subject to constant changes and updates, and, in the European context, dependant of each jurisdiction.

In this paper, a new dataset for Information Retrieval and Question Answering has been created in collaboration with domain experts from the Spanish law-firm Cuatrecasas. The dataset is comprised of 150 labor law questions related to the Spanish Workers Statute. The information in the document is divided into Titles, Chapters, Sections, and Articles. Each question in the dataset is matched to the corresponding article in the Workers Statute in which the answer is contained, and also to the specific excerpt of the article in which the question is answered. The questions, articles and paragraphs are available in Spanish and English languages. In addition, the segmented Workers Statute is distributed along with the dataset.

Moreover, this paper presents an initial experiment in the context of the European Project Lynx in which the dataset is used for evaluation purposes.

This paper is structured as follows: In Section II we briefly refer to similar initiatives. In Section III the

This work has been supported by the Lynx project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602.

dataset is described, and Section IV is devoted to the experiment. Finally, Section V presents the results of preliminary evaluations performed on the datasets, and conclusions and future steps are reflected in Section VI.

II. Related Work

Most of the datasets for Information Retrieval and Question Answering in the legal domain are oriented to exploit court cases in English language [8] which are also distributed in public research Tasks [3], [4]. For Spanish language, the recent publication of Legal-ES [7] poses a large scale Spanish corpus comprised of documents from different sources. However, the Spanish Workers Statute is not reflected in the corpus.

The value of this paper is to create a dataset focused on Spanish language for labor law questions related to the Workers Statute document.

III. Dataset

The dataset has been created during the developing of the European Project Lynx. Cuatrecasas, as part of the Lynx consortium, is interested in Information Retrieval systems over documents related to the Spanish legislation. The concrete use case is focused on the Spanish Workers Statute.

The dataset has been created by two experts of the law-firm that has collaborated jointly with other partners of the project. The gold standard is comprised of a set of 150 labor law questions in Spanish language related to the Workers Statute. Each question has been related with the title of the article in which the question is answered as shown in Table I.

The dataset with the rest of the documents used in this paper is published online[1] having followed the FAIR principles (findable, accessible, interoperable, reusable).

IV. Experiment

The experiment has been performed over a particular use case of Information Retrieval in Spanish language in the context of the European Project Lynx. The use case has been reported by a small business model of Cuatrecasas which is oriented to use the Spanish Workers Statute to solve client queries. The experiment is comprised of three main components. The first one is the document representation, how the corpora is represented and its format. Secondly, the database and its configuration properties. Finally, the description of the terminology extraction process.

[1]https://zenodo.org/record/4256718#.X6bIbGhKiUk

A. Document Representation

In the context of the European Project, the data model used for documents is the Legal Knowledge Graph (LKG) Ontology[2]. Lynx Documents are compliant with NIF (NLP Interchange Format)[3] specification and heavily reuses ELI[4] metadata elements. Lynx Documents may be grouped in Collections, and may be enriched with Annotations. Lynx Documents are the basic information units (pieces of text) and are described with Metadata and they are structured in Lynx Document Parts. Moreover, Collections groups documents with any logical relation but it can be achieved with metadata fields (`partOf` to reflect that a Lynx Document is part of another Lynx Document). An example Lynx Document is presented below.

In the experiment, the Spanish Workers Statute has been transformed into 130 Lynx Documents, one per article represented in the Statute

```
{
    "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.
        json",
    "id": "BOE-A-2015-11430_Part_69",
    "type": [ "lkg:Legislation",  "nif:Context", "lkg:
        LynxDocument" ],
    "text": "Articulo 67. Promocion de elecciones y mandato
        electoral.......",
    "metadata": {
      "skos:prefLabel": "Articulo 67 en Real Decreto Legislativo
          2/2015, de 23....",
      "lkg:partId": "BOE-A-2015-11430#offset_225601_231131",
      "jurisdiction": "ES",
      "language": "es",
      "title": {
        "es": "Articulo 67. Promocion de elecciones y mandato
            electoral."
      },
      "lkg:partOf": "BOE-A-2015-11430",
    },
    "parts": [],
    "annotations": [],
    "offset_ini": 0,
    "offset_end": 5530,
    "translations": {
      "en": "Article 67. Election promotion ..."
    }
  },
```

B. Database

The experiment is oriented to work with the document-oriented database Elasticsearch[5]. Elasticsearch organized documents in Indexes and these documents and their fields, which are represented in JSON format, can be configured to specify the how the values of the fields are stored and indexed in the database. For instance, for indexing text fields, usually the process involves the use of a tokenizer over the text. Moreover, a stemming process over resulting tokens is performed

[2]https://lynx-project.eu/doc/lkg/
[3]https://github.com/NLP2RDF/ontologies
[4]https://op.europa.eu/en/web/eu-vocabularies/eli
[5]https://www.elastic.co/

	Question	Document	Paragraph
Spanish	¿Qué situaciones interrumpen el periodo de prueba?	Artículo 14. Período de prueba	Las situaciones de incapacidad temporal, riesgo durante el embarazo, maternidad, adopción, guarda con fines de adopción, acogimiento, riesgo durante la lactancia y paternidad, que afecten al trabajador durante el periodo de prueba, interrumpen el cómputo del mismo siempre que se produzca acuerdo entre ambas partes.
English	What situations disrupt the trial period?	Article 14. Probationary Period.	Situations of temporary incapacity, maternity and the adoption or fostering of children affecting the worker during the probationary period interrupt the computation of the term, provided that agreement is reached between both parties.

TABLE I
Sample of the information captured in the Spanish Worker Statute Dataset: A question, the document where it is answered and the paragraph which contains it. In Spanish and English languages.

to ease the matching over variable tokens (e.g., cats, catlike, and catty have a common stem which is cat). Elasticsearch has stemmers for different languages such as English and Spanish.

Lynx Documents in their JSON format can be directly consumed by Elasticsearch. Also, for improving the results of the indexing process, a specific indexing configuration has been created for each language used in the context of the project. Tokenizers and stemmers for English and Spanish languages have been prepared for the documents of the Workers Statute.

Elasticsearch will use the fields of 'metadata.title' and 'text' for matching and scoring the queries.

C. Terminology Generation

As part of the Lynx project tasks, a terminology of the most relevant terms in the Spanish Workers Statute has been created. A key process within Information Retrieval is terminology extraction, that allows the identification of the most relevant terms within a document. For extraction process, we tested two different systems: the proprietary service provided by Tilde[6] as Lynx partner, that combines linguistic and statistical extraction; and an statistical open source application, TBXTools [6], which can be easily implemented and modified, when necessary. As a result, we obtained a combination of statistically and linguistically extracted terms that, with little post-processing, served as the basis for the generation of a terminology of labour law.

The list of retrieved terms was afterwards enhanced with translations in English, German and Dutch, synonyms, definitions and relations retrieved from the following existing language resources, being some of them close and proprietary, and others published in Semantic Web formats as part of the Linguistic Linked Open Data cloud[7]:

- Unesco Thesaurus: a controlled thesaurus containing mostly terms on social matters, structured in RDF and available to browse and download[8].
- EuroVoc: this was an originally hand-crafted thesaurus that was also transformed into RDF [1] and it is now available through an SPARQL endpoint[9] maintained by the Publications Office of the European Union.
- InterActive Terminology for Europe (IATE): this is a public resource containing terminology gathered by professionals in the European Union. The most recent version is available through a JSON API[10], although a previous version was converted into RDF following the lemon model [2].
- Lexicala: this API[11] is provided by KDictionaries, also part of Lynx consortium, and it contains general multilingual dictionaries.
- Wikidata: it is an open and collaborative knowledge base that is available through a user interface and also through a SPARQL query service[12].

Finally, the enhanced terminology was subsequently represented in SKOS[13], manually curated by domain experts from Cuatrecasas and published with open license in Zenodo. The resulting terminology is publicly available at[14]

V. Evaluation

The evaluation has been addressed by three experiments in relation of how the query is sent to the Elasticsearch database. The evaluation has been focused only the Spanish language because the experimentation is aligned with Cuatrecasas contribution and results in the project. The first experiment consisted in sending the question as it is written in the gold standard. In

[6]https://term.tilde.com/
[7]https://linguistic-lod.org/llod-cloud
[8]http://vocabularies.unesco.org/browser/thesaurus/en/
[9]http://publications.europa.eu/webapi/rdf/sparql
[10]https://iate.europa.eu/developers
[11]https://api.lexicala.com/
[12]https://query.wikidata.org/
[13]https://www.w3.org/TR/swbp-skos-core-spec/
[14]https://zenodo.org/record/3843561

	Perfect	Correct	Others	Not Found
Experiment 1. Q	66% (99 queries)	18% (27 queries)	11% (17 queries)	4% (6 queries)
Experiment 2. Q+ET	66% (99 queries)	18% (28 queries)	10% (15 queries)	4% (7 queries)
Experiment 3. Q+ET+ATE	68% (103 queries)	16% (24 queries)	10% (15 queries)	4% (7 queries)

TABLE II
Evaluation results for each experiment. Results are presented as a percentage from 0 to 100 and the number in the parenthesis represents the number of queries. Experiment 1 only takes into account the original query. In Experiment 2 the query is enriched with the Extracted Terminology and Experiment 3 includes the Automatic Term Extraction with Rake.

the second one, the question is enriched repeating those terms that are in the query and in the extracted terminology (relevant terms of the domain). Finally, in the last experiment the query is enriched as in the previous experiment and also repeating the query terms that have been identified by the library Rake[15]. The target of the library is to identify the important terms (nominal chunks) that appear in the query to reflect their importance in the scoring process.

The purpose of repeating terms in the query is to trunk the scoring values of the Elasticsearch search engine (coordination factor), increasing the value of these terms when the natural language query is sent.

The results for each query is a list of up to ten documents which are classified into four groups: perfect, correct, others and not found. Perfect means that the target document appears in the first result of the list (the most important one), correct means that it appears in the first four results, others means that the target document appears in the rest of the list and if it does not appear the list is classified as not found. The experiment results are presented in Table II.

The evaluation shows that even with the simple query in Experiment 1 the results reach a high performance in perfect and correct classifications (66% and 18%). The Elasticsearch stemming process for Spanish language has been critical to identify the correct documents in the database because the words of the query usually do not match perfectly with the words reflected in the documents due to the derived forms that are present in Spanish language (e.g., *trabajador* masculine worker, *trabajadora* feminine worker, *trabajadores* masculine workers, *trabajadoras* feminine workers).

Surprisingly, the use of the extracted terminology in Experiment 2 has not affected the overall results; only one question has passed from the category others to correct. Future experiments will analyze the use of the synonyms reflected in the terminology to enrich the query. However, the combination of the extracted terminology and the automatic term extraction over the query has reached the highest performance score with 103 documents (68%) classified as perfect.

[15]https://pypi.org/project/rake-nltk/

VI. Conclusions and Future Work

This work presents a new bilingual dataset oriented for the natural language processing tasks of Information Retrieval and Question Answering. The dataset is comprised of 150 labor law queries of the Spanish Workers Statute and it has been created in collaboration with domain experts of the law firm Cuatrecasas. Moreover, a simple Information Retrieval experiment has been developed in the context of a use case of the Lynx project in which the dataset is used for the evaluation. The dataset and the documents used in the experiment are distributed openly to the community.

As presented before, further experiments will exploit the use of synonyms extracted from the sources of information that are used to enrich the terminology. The main goal is to validate if these sources of information contain the correct synonyms that are used by end-users. Regarding the dataset, the main target is to increase the number of questions and answers to reach a trainable dataset that can be used by new state-of-the-art techniques such as BERT which requires a high amount of data.

VII. Acknowledgments

This paper has been funded by the project European project Lynx (Nº780602) and the European project Prêt-à-LLOD (Nº825182).

References

[1] M. L. Alvite Díez, B. Pérez León, M. M. Martínez González, D. J. Vicente Blanco, et al. Propuesta de representación del tesauro eurovoc en skos para su integración en sistemas de información jurídica. 2010.

[2] P. Cimiano, J. P. McCrae, V. Rodríguez-Doncel, T. Gornostay, A. Gómez-Pérez, B. Siemoneit, and A. Lagzdins. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517, 2015.

[3] Y. Kano, M.-Y. Kim, R. Goebel, and K. Satoh. Overview of coliee 2017. In *COLIEE@ ICAIL*, pages 1–8, 2017.

[4] A. Mandal, K. Ghosh, A. Bhattacharya, A. Pal, and S. Ghosh. Overview of the fire 2017 irled track: Information retrieval from legal documents. In *FIRE*, 2017.

[5] C. D. Manning, H. Schütze, and P. Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[6] A. Oliver and M. Vàzquez. Tbxtools: a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 473–479, 2015.

[7] D. Samy, J. Arenas-García, and D. Pérez-Fernández. Legal-ES: A set of large scale resources for Spanish legal text processing. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 32–36, Marseille, France, May 2020. European Language Resources Association.

[8] M. Saravanan, B. Ravindran, and S. Raman. Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, 17(2):101–124, 2009.