Proceedings of the First workshop on Language Technologies for Government and Public Admnistration (LT4Gov), pages 32–36
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Legal-ES: A Set of Large Scale Resources for Spanish Legal Text Processing

Doaa Samy°*, Jerónimo Arenas-García, David Pérez-Fernández
Instituto de Ingeniería del Conocimiento & Cairo University*,
Universidad Carlos III de Madrid,
Ministerio de Economía (Secretaría de Estado para la Digitalización e Inteligencia Artificial)
°IIC, C/Tomás Francisco y Valiente, Campus Universitario, Cantoblanco, Madrid, Spain
*Cairo University, Main Campus, Giza 12613, Egypt
Universidad Carlos III de Madrid, Leganés, Madrid, Spain
Secretaría de Estado para la Digitalización e Inteligencia Artificial, C/Poeta Joan Maragall 41, Madrid, Spain
doaa.samy@iic.uam.es & doaasamy@cu.edu.eg, jarenas@ing.uc3m.es, dperezf@mineco.es

Abstract

This paper presents work on progress aiming at the development of Legal-ES. Legal-ES is a set of resources for Spanish legal text processing including a large scale corpus with calculated models for word embeddings and topics. The large scale Spanish legal corpus consists of over 2000 million words from open public legislative, jurisprudential and administrative texts representing a variety of sources from international, national and regional entities. The corpus is pre-processed and tokenized. A word embedding is calculated over raw text and over lemmatised texts in addition to some experiments with topic modelling on the legislative subset of the corpus representing the text from the Spanish Official Bulletin of State (Boletin Oficial del Estado-BOE). Within the framework of the Workshop on Language Technologies for Government and Public Administration (LT4Gov), the present paper showcases how Public Data is a valuable input for developing Language Resources. It fits within the second dimension of the workshop, i.e. PublicData4LRs. Legal-ES is the result of an initiative by the team of the Spanish Plan for the Advancement of Language Technologies (Plan TL) aiming at developing resources for the HLT community to promote intelligent solutions by industry and academia destined to Public Administration and the Legal Domain.

Keywords: Language Resources, Legal Corpus, Embeddings, Topic Modelling, Legislative text, Spanish Resources

1. Introduction

In the legal domain, Human Language Technologies (HLT) and Natural Language Processing (NLP) have been gaining more and more attention over the last years. HLT and NLP are marking a significant difference in handling the large sets of documents and data usually managed by stakeholders in the legal domain, especially when applied in Information Retrieval and Information Extraction.

Within the Spanish National Plan for the Advancement of Language Technologies (Plan TL), priority domains are selected to develop pilot projects. The Legal domain has been one of these priority areas in the last two years given its relevance and its impact on society at the different levels: Governmental bodies' level, industry, academia, services for citizens, structural measures, etc. Conversations were held at different levels with Public and Regional Administrations as well as academic and industrial groups to gain first-hand insights on the current situation of NLP applied in the Legal domain within the context of the Spanish language and co-official languages (Catalan, Basque and Galician).

From the perspective of the National Spanish Plan, the Public Administration would play a relevant role in promoting the NLP industry in the legal domain by adopting NLP-based solutions in real case scenarios and by providing more legal public datasets to allow for developing innovative and intelligent components. These components would contribute to the Digital Transformation of Public Administration and would introduce innovative workflows turning traditional procedures into more effective and less-time-consuming tasks towards better services to the citizens. At the European level, initiatives such as e-Codex or e-Justice aiming at improving legal services for EU citizens by facilitating the exchange of legal information, are examples of the opportunities and the needs within this domain.

On the other hand, Spanish language is one of the top widely spoken languages, but most of the language resources developed for the legal domain are mainly in English. So, there is a justified need to develop resources in Spanish.

Development of Corpora of legal texts started some years ago. Vogel et al. (2017) lists some of the available corpora. BLaRC (The British Law Report Corpus) is an example of these efforts. The British English corpus is made up of judicial decisions and issued by British courts and tribunals consisting of 8.5 million words published between 2008 and 2010. The American Law Corpus (ALC) consists of 5.5 million words, while the Corpus of European Law includes a billion word in English and German.

Recent work concerning resources has focused on compiling large datasets and on applying deep learning techniques to train word2vec models. Chalkidis & Kampas (2019) shared word embeddings trained over a large dataset of legislations from UK, EU, Canada, Australia, USA and Japan among others. Nay (2016) published "Gov2vec" in which policies are compared across institutions by embedding representations of the legal corpus of each institution and the vocabulary shared across all corpora into a continuous vector space. The corpus used included 59 years of all U.S. Supreme Court opinions, 227 years of all U.S. Presidential Actions and 42 years of official summaries of all bills introduced in the U.S. Congress. Sugathadasa et al. (2017) used word2vec and lexicons for semantic similarity in the legal domain. Embeddings were calculated over a corpus of 35000 legal case documents, pertaining to various areas of practices in law from US Supreme Court.

Other examples of related work in the legal domain regarding specific applications or aspects of legal text processing include, among other, predictive models for decision support in administrative adjudication (Branting et al. 2017), contract element extraction (Chalkidis, 2017), legal question answering (Do, 2017) (Kim, 2015),

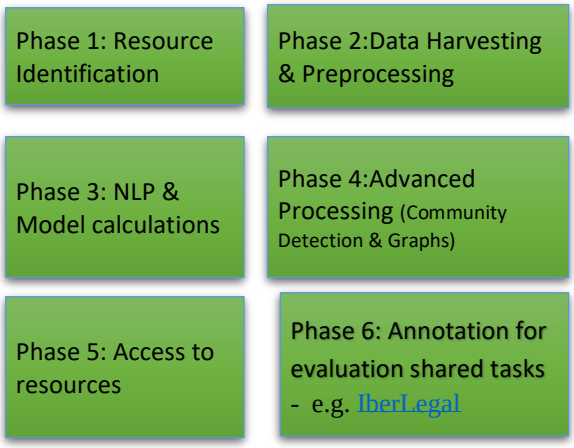
extracting requisite and effectuation parts in legal texts (Nguyen et al., 2018) and classification of sentential modalities (O'Neill et al, 2017).
State of art reveals the increasing interest, the variety of applications and the vast opportunities for HLT and NLP in the legal domain. Nevertheless, the dominance of the English language in most of the resources and the work done is obvious and there is a clear need for resources in other languages including Spanish, especially given that the industrial uptake would have an international wide impact on Spanish speaking countries.

2. Legal-ES: The Corpus

Legal-ES is a large scale Spanish corpus of over 2000 million words representing different types of legislative, administrative and jurisprudential texts. All datasets are gathered from open public portals, mainly from Spain, Europe and International organizations in addition to resources from Mexican and Argentinian portals. For the harvesting, we opted for the availability and the openness of the resource rather than balanced representations in certain time frames.

Figure 1. Resources for Legal-ES

Legal-ES is designed in six consecutive phases to allow for a wider coverage and access. Phases are applied on the different subsets according to the characteristics of each dataset with different timeframes:



At Phase 1, a number of sources were identified. Also, the legal aspects of some resources are subject of study and analysis by legal experts within the team. Nevertheless, a regular update is needed to add newly identified resources and to update the already available ones given the dynamic nature as legislations, sentences, public procurement etc. are in a continuous increase. Resources are identified in **four sets** according to the type. Table 1 summarizes ***Set 1 the preliminary list*** including: Legislations from the Official Bulletin (BOE), Opinions from State Council, Consultations from State Tax Agency, State Advocacy, Fiscal Doctrine, the Spanish subset of the European EurLex and JRC-Acquis dataset.

Name	Number of words
Legislación BOE	547.615.892
Doctrina Fiscalía	2.684.855
Dictámenes Consejo de Estado	135.348.664
Abogacía del Estado	6.123.007
Códigos electrónicos	24.261.786
Consultas tributarias	401.586.826
JRC-Acquis	59.155.891
EurLex	58.005.420

Table 1. Set 1: Identified Datasets and Size

Set 2: Additional legislative resources (238 million words) including:

- Legislations from Mexico Data Portal
- Legislations from Argentina Data Portal

Set 3: Additional jurisprudential resources including:

- Open public sentences from Spanish Supreme Court
- Open public sentences involving Spanish regional authorities (Madrid & Barcelona among others)
- Open public sentences from regional courts in Mexico
- Resolutions from the International Court of Justice (Spanish versions)

Set 4: Further administrative resources including:
- Public Procurement from the Spanish Platform
- Spanish versions of Public Procurement posted on EU Tender Daily Platform.
- Public Procurement from the Mexican Public Procurement Platform.

A selection of resources from the different sets have passed the legal check to ensure compliance to the open data licenses and thus proceeded to Phase 2, i.e. they are already harvested and pre-processed. This selection includes: All legislative sources in Set 1, the jurisprudential texts from the Supreme Court (Set 3) and the administrative texts from Spanish Public Procurement (Set 4). We will refer to this subset as Legal-ES/IberLegal.

3. Word Embeddings and Topic Modelling

For Phase 3: NLP & Model Calculations, two experiments for Word Embeddings were conducted. The first over the raw text in Set 1 and the second over the lemmatised subset of BOE legislations in ***Set 1***. Embeddings were trained over 300 dimensions and were collapsed to 2 dimensions for representation purposes as in Figure 3 at the end of this setcion. Both experiments showed interesting results. In figure 3, the square on the left shows a cluster with varieties of wine, while the square on the upper right shows words related to posts with near embeddings. The right square down shows words related to taxes.

Moreover, we tested introducing some words in different semantic fields to check the words with the nearest embeddings. For example, by introducing the words "impuesto [tax]", we found the nearest embeddings "renta [income]", "tributo [tribute]", etc. Also, by checking the word "ley [law]", the nearest embeddings were "orden_ministerial [ministerial_order]", "decreto [decree]", "decreto_real [royal_decree]", etc. In the agricultural domain, when the word "wine [vino]" was introduced, the nearest embeddings were types of wine such as "chardonnay", "merlot", "pinot", etc.

Legislations

-'orden_ministerial'*[Ministerial-order]*, (0.8387089967727661),
-'reales_decretos'[*Royal_Decree*], (0.8019422888755798),
-'decreto'[Decree],(0.7804737091064453),
-'real_decreto-ley' [Royal_Decree_Law], (0.7155911326408386),
-'orden' [Order], (0.6945813894271851),
- 'ley'[Law], (0.6891162991523743)

Taxes

[-('irpf.', 0.7073756456375122),
-('i.r.p.f.', 0.5817404389381409),
-('impuesto'[*Tax*], 0.5020797848701477),
-('renta'[*Income*], 0.46861714124679565),
-('tributo' [tribute], 0.46004000306129456),
-('impuestos[tax]', 0.45298290252685547),
('impuesto_de_sociedades'[*Societies´tax*], 0.444973886013031),

Regarding the Topic Modelling, different models (25, 40, 50 and 150) were trained over the subset of BOE legislations. Trained models of 25 and 40 showed good results with clear topics identified as shown in the examples in Table 2 and Figure 4. Table 2 represents an extract from the topic model-25 with the most representative words for the selected topics. In Figure 4, an example of topic representation in a document where there is a clear dominance of the topic 13 related to "Fishing legislations". This would also contribute in identifying semantic similarities among documents based on topic representation.

Topic	Word	Word	Word
Education & Universities	universitario	educación [education]	enseñanza [teaching]
Taxes	ayuda [aid]	gasto [expenditure]	pago [payment]
Workers	trabajo [work]	empresa [company]	convenio [agreeme
Agreements & Cooperation	partes [parts]	protocolo [protocol]	país [country]

Table 2. Example of Topics (Model-25)

In Phase 4- Advanced Processing, further experimentations were carried out to detect communities of similar documents, however results are still at very preliminary stage that needs further analysis. An example of the graph representations of the communities is the following. Nodes are documents and edges link documents with high semantic similarities:

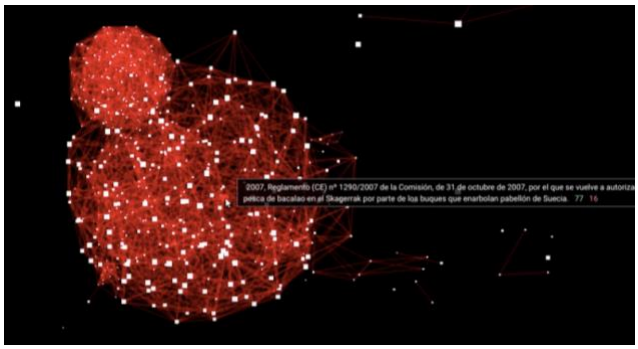


Figure 5. Community Detection in BOE

Finally, at Phase 5, access to the code that facilitates the downloading of the open public resources of BOE was made available on Github.

4. Future Steps: Annotation and Evaluation Shared Tasks

Currently, the process of annotation of Named Entities has started in samples from Legal-ES/IberLegal. The annotation is carried out via bootstrapping, i.e. initial manual annotation is fed into automatic annotation through an iterative method with a final manual validation and an inter-annotator agreement to obtain a gold standard set. The annotation considers five types of Named Entities:

- Persons
- Institutions
- Time expressions
- Locations
- References to laws and legislations.

Sample fragments from legislations, sentences and public procurement are annotated over Brat Platform.

The annotation is still at an exploratory stage, but it is quite challenging specially that the distribution of the Named

Entities and the complexity of the annotation varies among the different types of texts. For example, legislations from BOE follow a normalised style making the annotation easier, while administrative texts of public procurement are much more cumbersome given the broad diversity of the procurement texts and the lack of normalised forms.

The annotated set will be made available within an evaluation shared tasks named after the sample corpus "IberLegal". The task is organised within the Spanish Evaluation Campaign "Iberlef". For the annotation, experiments were carried out for Named Entities annotation using open libraries such as Spacy, FreeLing or IXA Pipes, however, the generic tools performed poorly on this type of texts, specially in detecting references of laws. This revealed a clear need for Named Entity Recognition and Classification components adapted to the legal domain in Spanish. Based on these findings, the team at the Spanish Language Technologies Plan decided to organise IberLegal as the first shared task focusing on Named Entities Recognition in Spanish legal and administrative text.

5. Conclusion

In this paper, we presented a showcase of Public Data as Language Resources where we introduced the ongoing work to develop Legal-ES as a large scale language resource for Spanish legal text processing. The paper outlined the development phases and the progress achieved to date. Through this showcase, we share a possible roadmap of how to start from an open public data resource until reaching to a mature language resource and how to engage the community in the development of measurable components and advances. The actions taken are just steps on the way, but more work and effort is needed to achieve a solid infrastructure that allows for further developments.

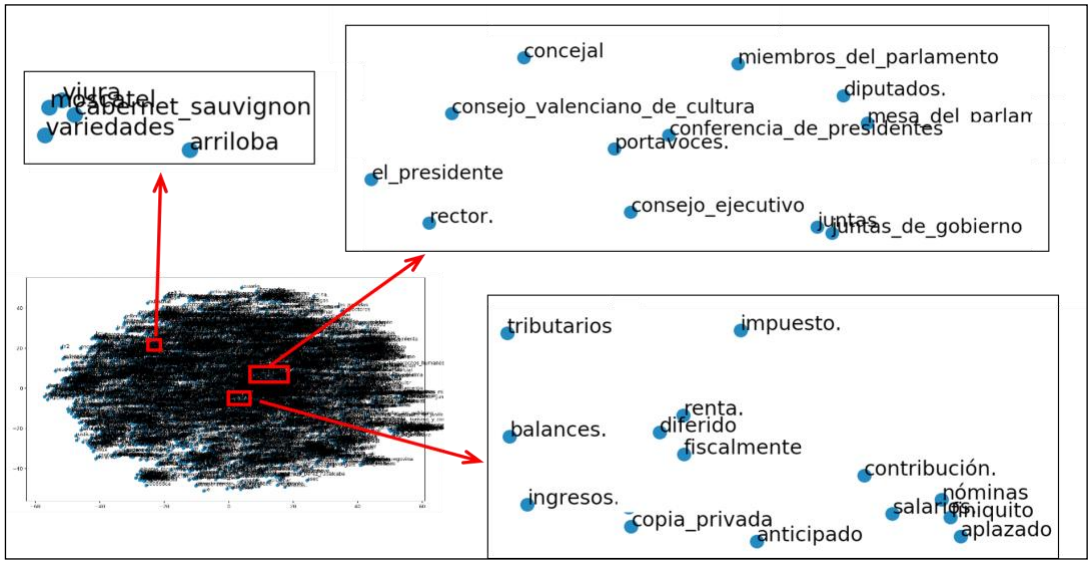


Figure 3. Representations of Word Embeddings-Set 1

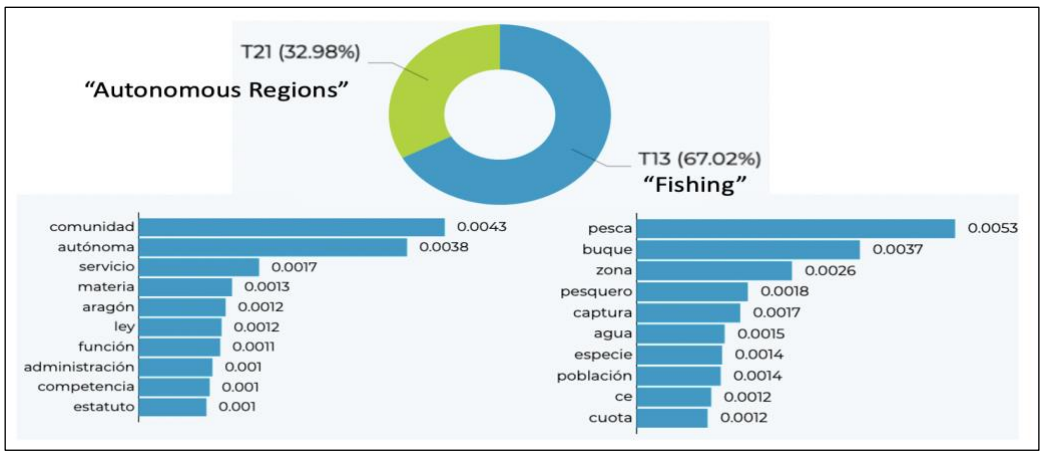


Figure 4. Document Representation through Topics

6. Bibliographical References

Branting LK, Yeh A, Weiss B, Merkhofer E, Brown B (2017). Inducing predictive models for decision support in administrative adjudication. In: Proceedings of the MIREL Workshop on the The 16th International Conference on Artificial Intelligence and Law, London, UK.

Chalkidis I, Androutsopoulos I (2017) A deep learning approach to contract element extraction. In: Proceedings of the 30th International Conference on Legal Knowledge and Information Systems, Luxembourg, pp 155–164.

Chalkidis, I., Kampas, D. (2019). Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 27, 171–198 (2019). https://doi.org/10.1007/s10506-018-9238-9.

Do PK, Nguyen HT, Tran CX, Nguyen MT, Nguyen ML (2017). Legal Question Answering using Ranking SVM and Deep Convolutional Neural Network. CoRR abs/1703.0. arXiv:1703.05320.

O'Neill J, Buitelaar P, Robin C, Brien LO (2017). Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. In: Proceedings of the 16th international conference on artificial intelligence and law, London, UK, pp 159–168.

Nay JJ (2016) Gov2vec: Learning distributed representations of institutions and their legal text. In: Proceedings of the first workshop on NLP and computational social science. Association for Computational Linguistics, pp 49–54, Austin, Texas. DOI:10.18653/v1/W16-5607

Nguyen T, Nguyen L, Tojo S, Satoh K, Shimazu A (2018). Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. Artif Intell Law 26(2):169–199

Kim My, Xu Y, Goebel R (2015) A convolutional neural network in legal question answering. In: Ninth International Workshop on Juris-informatics (JURISIN).

Sugathadasa, K., Ayesha, B., Silva, N.D., Perera, A., Jayawardana, V., Lakmal, D., & Perera, M. (2017). Synergistic union of Word2Vec and lexicon for domain specific semantic similarity. 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), 1-6. DOI:10.1109/ICIINFS.2017.8300343.

Vogel, F., Hammann, H. and Gauer, I (2017). Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies. Law & Social Inquiry, 2017. DOI: https://doi.org/10.1111/lsi.12305.