Challenges of Terminology Extraction from Legal Spanish Corpora

Patricia Martín-Chozas[0000−0002−8922−7521] and
Pablo Calleja[0000−0001−8423−8240]

Ontology Engineering Group, Universidad Politécnica de Madrid
http://www.oeg-upm.net/
{pmchozas, pcalleja}@fi.upm.es

Abstract Untangling the complexities of legal documentation is an imperative need for non practitioners of the legal profession. The terminology used in the domain is complex and it usually requires expert knowledge to be fully understood, since the legal framework is constantly being updated and the meaning of terms vary accordingly. Non-proprietary Automatic Terminology Extraction (ATE) tools are required in this particular domain in which documents contain private and sensitive data. This paper describes methods for obtaining accurate legal terms from labour law corpora, overcoming the difficulties present in the area, and also analyses the peculiarities of the legal jargon, specifically, in Spanish language. The performed experiments, executed with JATE, a well-known open source library in the ATE literature, are still preliminary, but promising.

Keywords: Legal terminology · Automatic Term Extraction · Natural Language Processing · Semantic Web Technologies

1 Introduction

As evidenced by legal summaries published in Eurlex[1], legal documentation such as laws, directives, decrees and even regular notices are often complex to understand by general public: legal terminology can be a real headache for people that are not used to this type of language. At the same time, many citizens and businesses across the European Union have to deal with significant compliance problems: breach of contracts, overdue debts, excessive working hours, etc.

With the aim of softening such complications, this paper proposes to retrieve *legal terms* from Spanish corpora through Automatic Term Extraction (ATE) techniques. Such legal terms are understood as words and multi-word expressions with a specific meaning within a legal text; collections of such terms are considered *terminologies*. These terminologies could be afterwords interlinked with other language resources to share information, which contributes to obtain definitions, translations and context, easing the comprehension of legal documentation.

[1] https://eur-lex.europa.eu

ATE is a well-known technique in Natural Language Processing (NLP) that supports important tasks such as machine translation, speech recognition or information retrieval, to mention but a few. Several tools are already offering this kind of technology; however, many of them still present unresolved limitations, such as noise generation, disambiguation issues or performance, delays with big corpora, which are some of the most frequent. Here, two additional limitations have been identified as crucial when dealing with legal information: domain specificity and data privacy. Current ATE tools find difficulties when extracting highly-specific legal expressions and, on the other hand, they might include personal data such as proper names and identity numbers in the resulting term lists.

For this reason, this contribution proposes to configure the tool JATE [13] to extract terminology from the legal domain by analysing the peculiarities of the legal language and adapting the extraction patterns to this jargon.

The first use case has been developed within the Lynx project, which has provided the Spanish legal corpus[2] used in the experiments. Lynx project[3] is an H2020 Innovation Action towards the creation of a Knowledge Graph of legal and regulatory data from different jurisdictions and languages. This Legal Knowledge Graph interlinks multilingual legal information and provides explanations and context for legal expressions appearing in each document.

The terms extracted with the model developed here contribute to the creation of this platform that help European citizens understand legal documentation without having to invest time and money in specific legal consulting.

JATE was originally developed for English terminology extraction; since this use case deals with Spanish corpora, it was also required an extension of the tool to cover Spanish language.

This paper is organised as follows: Section 3 presents the related work on term extraction technologies, Section 2 exposes the motivation behind this contribution and the analysis of the problem, Section 4 describes the experiments performed during this work and finally Section 5 includes conclusions and future work to be performed in next stages.

2 Motivation

Some hints of the motivation behind this project have already been presented in the introduction: legal documentation is an unsolvable puzzle for laymen and non-practitioners of the legal domain.

In addition, many of the available legal language resources are not in machine readable formats yet. The major part is published as PDF, which hinders their look-up, and some of them are still distributed in physical format.

The work proposed here tackles this situation by researching on a most accurate terminology extraction methodology from those machine-readable legal

[2] http://data.lynx-project.eu/dataset/llcorpuses
[3] http://lynx-project.eu/

documents with the aim of creating new language resources that can be processed by the newest technologies.

A higher level of accuracy in the extracted terms means less time and money invested in human post processing of the resulting term list. However, the most important advantage lays on the efficacy of Semantic Web technologies: accurate legal terms can easily be linked with other legal language resources, offering users more information such as translations, synonyms, context and related terms. This additional information can help them improve their comprehension of the legal documentation.

Available tools present several limitations, and only a few of them are open source. This means a major drawback since one of the ideas of this contribution is to avoid using proprietary software to address privacy issues.

Also, a common shortcoming of web based applications is personal data management. Legal documents contain private data and it may not be safe to upload them on a web-based tool. In the Lynx project, both public and sensitive documents are being handled, thus, data privacy is an important factor to keep in mind.

Since JATE is an open source framework, it allows the use of a given algorithm selected by the user and even the creation of new ones. This work includes the extension of the tool to cover terminology extraction from Spanish corpora, the analysis of the labour law corpora provided by Lynx partners to discover specific patterns of legal language and the configuration of such patterns in the tool with the aim of extracting more accurate terms.

3 Related Work

Prior to the availability of term extraction tools, this activity was carried out by domain experts and terminology professionals. Despite being the most accurate manner for terminology extraction, it is also the most expensive and time-consuming. Taking into account the amount of information generated nowadays, human terminology extraction is unpractical.

For this reason, automatic term extraction technologies have been extensively studied in the literature [4] and several tools based on statistical and linguistic methods have already been developed.

In previous work, a comparative evaluation of available ATE tools has been performed [11]. Each of them presents different features depending on the format of the tool, typology of the targeted corpus, supported languages, types of extracted terms, etc.

Many of the tools analysed are web based applications: Translated.net[4], TermoStat[5] and FiveFilters[6], for instance, are free tools that can be accessed online.

[4] https://labs.translated.net/terminology-extraction/
[5] http://termostat.ling.umontreal.ca/
[6] https://fivefilters.org/term-extraction/

However, some of the issues found in the evaluation above mentioned include the extraction of stop words, visualisation of terms only in the website (not downloadable) and difficulties to find terms from specific domains.

Other tools offer payable services, such as SketchEngine[7], which is a sophisticated tool with a good performance and additional services [10]. Still, this application presents some difficulties in dealing with big corpora: at least in the trial version, files need to be attached one by one.

On the other hand there are other downloadable tools available, such as TBXTools[8] and TermSuite[9]. The first one has been developed to offer domain and language independent services, thus, implemented patterns might be too general; and the latter does not extract compound terms, the main feature of legal jargon.

This contribution uses the open source library JATE[10], since it integrates the most important terminology extraction algorithms and can be integrated in a local Solr indexer. Furthermore, it can also process large corpora and be extended for other languages or for different purposes, such as the Spanish legal domain.

4 Experiments

In the first place, subsection 4.1 describes the extension of JATE for Spanish language and the initial set of patterns applied. Secondly, subsection 4.2 contains an analysis of the features of legal terminology, specifically within Spanish labour law corpora. Afterwards, subsection 4.3 proposes the Spanish language patterns to be configured in JATE based on the analysis above mentioned. Lastly, subsection 4.4 contains a description of the extraction tests performed.

4.1 JATE extension for Spanish

JATE has implemented ten algorithms: TTF, ATTF, TTF-IDF, RIDF, CValue, X2, RAKE, Weirdness, GlossEx and TermEx.

In these experiments, only Cvalue [2] and TTF-IDF [9] have been used, since the former is intended for multi-word term extraction, one of the main features of legal documents, while the latter measures the significance each term of based on its frequency on the corpus.

JATE relies on the OpenNLP library[11] for tasks such as tokenisation, sentence segmentation, part-of-speech (POS) tagging and chunking, being the last two the most significant to process Spanish documents. For the POS tagging in Spanish, a model trained for version 1.3 has been adapted to the latest version. The different POS tags are coded with Cast3LB format [3], a freely available

[7] https://www.sketchengine.eu/
[8] https://sourceforge.net/projects/tbxtools/
[9] http://termsuite.github.io/
[10] https://github.com/ziqizhang/jate
[11] http://opennlp.apache.org/

treebank for Spanish (see morphological tagset in Figure 1). For the chunking in Spanish, there are no trained models available. However, JATE allows the creation of patterns to identify chunks in natural language based on POS tags. This chunks will reflect potential candidate terms. The Spanish patterns were built from a general corpus composed by newspapers and general articles as per the morphological instructions in Figure 1:

In order to create the Spanish patterns, previous work with English tagsets has been taken as a reference [7]. Such patterns have been translated to the Spanish POS tags and modified according to the grammatical structures of the Spanish language [5].

Categories	Cat-Gloss	Subcategory	Features
N	Noun	Proper, common	Gender; number
V	Verb	Main, auxiliary, semiauxiliary	Tense, mood, person, number, gender
A	Adjective	Qualifying, ordinal	Gender; number
R	Adverb	General, negative	–
P	Pronoun	Personal, demonstrative, numeral, indefinite, relative interrogative, possessive	Gender; number person
D	Determiner	Article,demonstrative numeral, indefinite, relative	Gender; number person
S	Preposition	Simple-complex	Gender; number
C	Conjunction	Subordinating, coordinating	–
I	Interjection	–	–
W	Date	–	–
F	Punctuation	–	–
Z	Number	–	–
Y	Abbreviation	–	–

Figure 1. Morphological tagset of Cast3LB

The initial tag patterns used for the first extraction tests against a general corpus were structured as in Figure 2.

Based on this figure, examples of the first patterns presented are:

- `(\bNC\b)`: **N**oun **C**ommon (e.g. *regulation*)
- `(\bNC\b) (\bNC\b)` : **N**oun **C**ommon + **N**oun **C**ommon (e.g. *family background*).
- `(\bAQ\b) (\bNC\b)` : **A**djective **Q**ualifying + **N**oun **C**ommon (e.g. *national jurisdiction*).

Some peculiarities of the patterns showed in Figure 2 are, for instance, that only common nouns have been considered to be extracted as simple terms. Verbs and adjectives of general knowledge tend to be less relevant when building a domain independent vocabulary. Hence, their tags have not been added to the patterns to avoid noise generation.

```
default (\bNC\b)
default (\bNC\b) (\bNC\b)
default (\bAO\b) (\bNC\b)
default (\bAQ\b) (\bNC\b)
default (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bAO\b) (\bNC\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bAQ\b) (\bAQ\b) (\bAQ\b)
default (\bAO\b) (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bNC\b) (\bSP\b) (\bAQ\b)
default (\bNC\b) (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b)
default (\bNC\b) (\bNC\b) (\bSP\b) (\bDA\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b) (\bAQ\b)
```

Figure 2. First configuration of Spanish general terminology patterns

4.2 Analysis of legal corpora

The patterns exposed in the previous section were intended for general information extraction. However, legal language has its own peculiarities that need to be considered [6] [8]:

- Long and intricate sentences
- Scarce punctuation marks
- Expressions in foreign languages (usually Latin) (e.g. *inter alia*)
- Rare and complex expressions only used in a legal context:
  - Legal terms of art: technical words with exact meaning that cannot be replaced by other terms (e.g. *comodato*, meaning "bailment")
  - Legal jargon: terms and expressions used by lawyers, often archaic and obsolete words (e.g. *lo antedicho*, meaning "the aforesaid")
  - Terms from the general language with a different meaning in the legal domain (e.g. *furnish*, meaning "to provide something or send something")

These terms should be identified since, although in a general context they may have little significance, they are relevant in a legal context.

The labour law corpus used for this experiment is composed by 20 documents containing information on national company agreements from different regions in Spain. Keeping the previous characteristics in mind, this corpus was analysed before the extraction and the following considerations regarding legal terminology were raised:

- The major part of the terms are multiword expressions (e.g. *convenio colectivo*, *Boletín Oficial del Estado*, *grupo profesional*)
- Many of these multiword expressions are built with prepositions and contractions connecting their elements (e.g. *comité de empresa*, *prevención de riesgos laborales*, *estatuto del trabajador*)

– Some types of words that in a general context are not considered terms, such verbs and adjectives, in the legal domain do need to be extracted as terms since they have specific meaning (e.g. *vigente*, *enunciativo*, *devengar*, *retribuir*)
– For the purposes of these experiments, proper names are not considered legal terms and must be avoided in the extraction stage (e.g. *Comunidad de Madrid*, *Principado de Asturias*, *empresa Hermanos Fernández*)
– Similarly, URLs, numbers and dates must also be kept out from the resulting termlists (e.g."https://www.boe.es/")
– Finally, terms including ordinal adjectives that in other jurisdictions may have a unique meaning, such as the "Third Amendment", do not exist in Spanish legislation, so these types of words should not be extracted

4.3 Proposal of Spanish legal language patterns

Thus, the first version of the Spanish general patterns was modified accordingly. Some of the patterns were removed and other new tag configurations were added:

```
default (\bAQ\b)
default (\bVM\b)
default (\bNC\b)
default (\bAQ\b) (\bNC\b)
default (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bAQ\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bSP\b) (\bAQ\b) (\bNC\b)
default (\bAQ\b) (\bNC\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bAQ\b) (\bSP\b) (\bDA\b) (\bNC\b)
```

Figure 3. First configuration of Spanish legal terminology patterns

Following the legal language analysis described in Section 4.3 and the morphological tagset showed in Section 4.1, some examples of the patterns added are:

– `(\bAQ\b)`: **A**djective **Q**ualifying (e.g. *vigente*, meaning "in force")
– `(\bVM\b)`: **V**erb **M**ain (e.g. *devengar*, meaning "accrure")
– `(\bNC\b) (\bAQ\b) (\bSP\b) (\bDA\b) (\bNC\b)`: **N**oun **C**ommon + **A**djective **Q**ualifying + **S**imple **P**reposition + **D**eterminer **A**rticle + **N**oun **C**ommon (e.g. *Boletín Oficial del Estado*, meaning "Official Bulletin of the State")

On the other hand, other patterns have been removed to avoid noise generation, since they were not considered relevant for legal terminology:

- `(\bAO\b) (\bNC\b)`: **A**djective **O**rdinal + **N**oun **C**ommon (e.g. *Third Amendment*)
- `(\bNC\b) (\bNC\b)`: **N**oun **C**ommon + **N**oun **C**ommon. This pattern was deleted since it was observed that the POS tagger sometimes tags proper nouns as NC (common nouns), and the tool extracts structures that are not real terms, such as *empresa Apple*.
- Patterns that link several common nouns or several adjectives have also been removed, since they are very extended in English but not in Spanish grammar which, normally, uses prepositions to separate each component of the term.

4.4 Extraction tests

The testing corpus provided by Lynx partners contains 20 files comprised of 21,475 tokens in TXT format that have been automatically converted apart from PDF files from the labour law domain in Spanish. Two tests have been performed over the Lynx corpus, one per pattern set:

- Extraction tests with Cvalue and TTF-IDF algorithms applying the General Spanish patterns.

  Samples of the most relevant extracted terms are *convenio colectivo* (collective agreement), *dirección de la empresa* (company management), *empresa* (company) and *trabajador* (worker).
- Extraction tests with Cvalue and TTF-IDF algorithms applying the Legal Spanish patterns.

  Most of the main terms recognised with the general Spanish patterns still remain in the same positions with similar scores. However, legal patterns have introduced new relevant terms in the legal domain such as *vacaciones por antigüedad* (seniority holidays), *asambleas convocadas por el comité* (committee-organised assemblies), *documentos relativos a la liquidación* (liquidation documents), *jubilado* (retired) or *disciplinarios* (discplinary).

From the resulting lists of terms, sorted by relevance, the first 200 terms[12] have been considered as the most meaningful, since they have significance in their scores. In this context, "relevant terms" are those legal expressions that can be used to annotate and classify documents by topic or typology, this is, terms that represent the legal domain. Table 1 collects some of the new extracted by JATE, applying the Spanish Legal Patterns[13].

[12] http://doi.org/10.5281/zenodo.2385437
[13] https://github.com/oeg-upm/terminology-extractor

New terms Cvalue	New terms TTF-IDF
• *trabajador tendrá* (worker will have) • *conocimientos adquiridos en el desempeño* (acquire knowledge during the performance) • *ley de prevención de riesgos* (risk prevention law) • *representación legal de los trabajadores* (legal representation of workers) • *firma del presente* (signing the present) • *texto refundido de la ley* (combined text of law) • *bocmboletín oficial de la comunidad* (bocmofficial bulletin of the community) • *entrada en vigor del presente* (implementation of the present contract) • *miembros del comité de empresa* (members of the company committee) • *comisión mixta de interpretación* (mixed interpretation commission) • *boletín oficial de la comunidad* (official bulletin of the community) • *boletín oficial de la junta* (official bulletin of the council)	• *anónima* [AQ] (anonymous) • *flexible* [AQ] (flexible) • *dura* [AQ] (severe) • *sanitario* [AQ] (sanitary) • *discontinuo* [AQ] (discontinuous) • *grave* [AQ] (serious) • *mixta* [AQ] (mixed)

Table 1. New terms extracted by the legal patterns

From this table, the following considerations can be educed:

- New terms retrieved by CValue algorithm are comprised of multiword terms. The algorithm has been enriched with more complex nominal chunks that are relevant in the corpus. However, for the TTF-IDF algorithm, the main results retrieved are based on the individual terms, in this case adjectives.
- The tool presents some tagging mistakes: the term *trabajador tendrá* has not correctly been extracted, since the last component is a verb and there is not such pattern in the set. From the configuration of the patterns, the tool has tagged this verb as a common noun or a qualifying adjective (see Figure 3).
- Also, there might be some tokenisation mistakes in the source corpus, since the term *bocmboletín oficial de la comunidad* has not been correctly extracted either. Another possibility is that the PDF to TXT conversion inserted mistakes such this one in the source files. The quality of the source file must be reviewed and improved.

5 Conclusions and future work

Results present little variations: 6% of new terms applying legal patterns with Cvalue and 3.5% with TTF-IDF. This situation is mainly given by POS tagging mistakes: the tool tends to tag any unknown word as common noun since it is the most frequent type of word in texts. Thus, some proper names, URLs, verbs

and adjectives are extracted as common nouns, spoiling the patterns. Training the model again with larger corpora from the domain would avoid many of these mistaken tags.

Also, not all the extracted terms are relevant for the legal domain. Some terms (e.g. *sanitario*, *dura*, *mixta*) do not belong to the legal terminology, so it would be required to generate a list with terms from the general usage to escape such extractions in future experiments.

On the other hand, many of the wrong extractions are caused by tokenisation mistakes in the original corpus: clean and well structured source documents would avoid this issue. Moreover, other libraries such as IXA Pipes [1] cover better NLP tasks for Spanish Language. It is proved that its POS tagger retrieves better results for Spanish corpora and it uses EAGLE tags [14] for the word classes, which contain more information about them (e.g, gender, number, form). Even though patterns would require to be coded again, they could represent much more grammatical content.

Another immediate convenient step to improve and generate a sound set of legal patterns is to consult grammatical peculiarities of legal language with actual legal experts: lawyers, prosecutors, judges, law students, etc., since these professionals are the best knowledge source of this domain.

On the whole, these experiments have been performed to highlight the importance of developing terminology extractors without the need of using online platforms for domains that deal with sensible data that cannot be distributed to third parties.

Regarding the use of JATE, the best approach here is to use several algorithms (in this case, Cvalue and TTF-IDF) with different features and performances to get more comprehensive results and to get a better overview of terms of the corpus. For instance, in the experiments, the results have shown the importance of the nominal chunks in the corpus although there are adjectives and nouns that are importance by themselves.

Also, since the tool allows the implementation and creation of new algorithms, another interesting experiment is the testing of other existing algorithms such KEA [12], used by other available tools, and the generation of a customised algorithm for legal language to test its accuracy.

Finally, although in these experiments only public data have been handled, a future line of work would be focused on identifying sensitive data to identify and remove named entities of persons and organizations in a given corpus to treat personal and private data. Furthermore, entity linking processes of the retrieved terms are also considered in order to validate accuracy of the terms, and search for related terms in relevant knowledge bases.

References

1. Agerri, R., Bermudez, J., Rigau, G.: Ixa pipeline: Efficient and ready to use multilingual nlp tools. In: LREC. vol. 2014, pp. 3823–3828 (2014)

[14] http://www.lsi.upc.es/ nlp/tools/parole-sp.html

2. Ananiadou, S.: A methodology for automatic term recognition. In: Proceedings of the 15th conference on Computational linguistics-Volume 2. pp. 1034–1038. Association for Computational Linguistics (1994)
3. Civit, M., Martí, M.A.: Building cast3lb: A spanish treebank. Research on Language and Computation **2**(4), 549–574 (2004)
4. Costa, H., Zaretskaya, A., Pastor, G.C., Seghiri, M.: Nine terminology extraction tools: Are they useful for translators? Multilingual (2016)
5. Española, R.A.: Nueva gramática de la lengua española (2009)
6. Haigh, R.: Legal English. Routledge (2018)
7. Handschuh, S., QasemiZadeh, B.: The acl rd-tec: a dataset for benchmarking terminology extraction and classification in computational linguistics. In: COLING 2014: 4th International Workshop on Computational Terminology (2014)
8. Hidalgo, A.: La ambigüedad en el lenguaje jurídico: su diagnóstico e interpretación a través de la lingüística forense. Anuari de Filologia. Estudis de Lingüística (7), 73–96 (2017)
9. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering **1**(1), 9–27 (1995)
10. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1**(1), 7–36 (2014)
11. Martín Chozas, P.: Towards a Linked Open Data Cloud of language resources in the legal domain. UPM (2018)
12. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: Practical automated keyphrase extraction. In: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129–152. IGI Global (2005)
13. Zhang, Z., Gao, J., Ciravegna, F.: Jate 2.0: Java automatic term extraction with apache solr. In: LREC (2016)