Challenges of Definition Extraction in Spanish Legal Texts

Karen Leticia Vázquez-Flores[1 [0000−1111−2222−3333]]
Patricia Martín-Chozas[1 [0000−0001−5416−6370]]
Elena Montiel-Ponsoda[1 [0000−0003−3263−3403]]

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
kvazquez@delicias.dia.fi.upm.es
{pmchozas, emontiel}@fi.upm.es
https://oeg.fi.upm.es/

Abstract. In this paper, we describe a preliminary experiment on definition extraction over Spanish legal texts based on linguistic patterns and helped by an automatic terminology extraction tool. Through this experiment, we analyse current issues on the general definition extraction task and, specifically, open challenges on legal Spanish definition extraction. Consequently, we propose various ideas on possible solution to those issues, next steps and future work.

Keywords: Definition Extraction · Legal Texts · Terminology.

1 Introduction

Information Extraction (IE) and Information Retrieval (IR) are two areas of research in the Natural Language Processing (NLP) field whose aim is to obtain information from unstructured texts. Depending on the type of information to be identified, (be it names of companies or people mentioned in the text, concepts dealt with in the document, or opinions given in Tweets), we will be referring to Named Entity Recognition, Term Extraction, or Opinion Mining, to mention but a few of the more specific tasks encompassed by these areas. When dealing with Term Extraction, the aim is to identify those terms or keywords that better define the content of a document. Terms may provide a network of superficial clues to understand whether a document fits the expected search criteria [5]. Automatic Term Acquisition may also be used for indexing new documents, or automatically annotating documents with the resulting terminology. In this sense, the richer the information describing the term, the more accurate the annotations will be.

For the purpose of adding definitions to terms extracted from a legal corpus, we have implemented a set of patterns to automatically discover definitions of Spanish legal terms implicitly defined in texts. It is well recognised

that Legal language is a professional jargon, and, as such, it has some particularities not so frequently found in plain language. Spanish legal language is characterised by the frequent use of technical terms, some archaisms, and complex syntactic constructions (abundant periphrasis, long subordinate clauses, gerunds, impersonal and passive verbs, long enumerations, use of future subjunctive and future tense for obligation, preference for the 3rd person singular, aphoristic style, many citations and references, etc.[1].

Definitions are also abundant in legal documents. Many terms will be directly defined in the text the first time they are mentioned. Some legal documents include a glossary of terms and definitions at the end of the document or as an annex (for example, European Directives and Regulations usually devote one article to define the main terms in the document[2]). However, this is an uncommon practice in Spanish national law, and discovering definitions in text is a necessary and challenging task. The way in which legal terms are defined in legal discourse follows certain lexico-syntactic patterns that are not common in plain language. In order to palliate the lack of approaches to automatically retrieve definitions of legal terms directly defined in Spanish legal documents, we present a set of lexico-syntactic patterns and describe the approach followed to retrieve definitions, as well as some preliminary results.

The rest of the document will be structured as follows: firstly, we briefly refer to some works on the extraction of definitions in law texts. Then, we refer to the main features of definitions in law, and describe the patterns we have created and their implementation. Finally, we discuss the results obtained with and without the help of a terminology extraction tool, and refer to future lines of work.

2 Related work

The most commonly applied approaches to extract definitions are pattern based approaches. Their main drawback lays on the linguistic knowledge required to apply them; however, correctly built patterns usually retrieve accurate results, as tested in various previous works. For instance, in 2004, Saggion applied a method that built the basis for a definition-based question-answering system [9]. The reproducibility of this work was, however, limited to a restricted number of definitions. In the case of [4], they worked with verbal patterns in Romanian, being the most productive ones the ones in which definitions were introduced by *denote*, *state*, *represent*, *define*, *specify*, *consist*, *name* and *permit*. A similar work, with similar results, is performed in [8] to retrieve definitions for Slavic languages. There have also been pattern-based approaches specialised in the legal domain, such as [10], in which the

[1] https://www.um.es/tonosdigital/znum9/corpora/juridicos.htm
[2] https://eur-lex.europa.eu/eli/reg/2016/679/oj

authors identify 52 rules based on lemma/POS and dependency parsing to extract definitions from a large corpus of German court decisions. To palliate the low recall they obtained, they used bootstraping techniques based on single seed words at the expense of precision. The authors suggest that a balance between precision and recall is to be further achieved.

Leaving pattern- and rule-based approaches aside, there have also been several experiments relying on Machine Learning. The scalability of those works is always much higher than with pattern-based approaches, but the manual work required for the annotation of a reference corpora is highly expensive. This is the case of [1], where they explored the use of machine learning to extract definitions from non-technical English texts, or [2], where results suggest that naive Bayes and random forest have the most suited learning bias for the task of definition extraction.

Mixed techniques have likewise been explored to identify definitions with higher linguistic variability than the classic encyclopedic *genus-et-differentia* definition. In 2009, Westerhout [11] proposed a definition extraction method for the automatic creation of glossaries within the area of eLearning. They used a sequential combination of a rule-based approach, based on POS-tags and a Machine Learning classification algorithm based on Balanced Random Forest. Another mixed work is [3], in which a weakly supervised bootstrapping approach was used to classify sentences in Wikidata as being definitions or not. In general, this kind of hybrid approaches show competitive results in comparison with others only based on patterns, but there is no evidence of works in specialised languages or languages other than English.

All in all, most of the reviewed works suggest that identifying definitions is a challenging task that is dependant on the language and the area of expertise, and that probably mixed approaches are expected to provide effective solutions. Little work has been done with regard to legal definition extraction in Spanish. Our aim is to take the first steps in this topic through a preliminary experiment based on patterns and helped by an automatic terminology extraction tool. We analyse the main issues and challenges derived from this experiment, propose ideas to solve them and expose future steps.

3 Definitions in law

A *definition* is a piece of text that helps clarifying the specific sense of an expression that appears several times throughout a document [7]. Definitions are divided into two parts: *definiendum* and *definiens*. The *definiendum* is the word or element that is to be defined, and the *definiens* is the sentence or clause that gives the meaning to the *definiendum*. For example, in the definition "semantic is the science of meaning", the *definiendum* is "semantic" while the *definiens* is "the science of meaning". These parts are usually

joined by a *copula*, which, in this case, would be "is". The most common structure of definitions is, hence, *definiendum*, *copula* and *definiens*.

However, these components are not always present in a fixed order. Specially in legal texts, it is common that the *definiens* component appears before the *definiendum*, and that the *copula* is not always expressed with the verb to be, which can sometimes lead to misunderstandings.

Specifically in Spanish legal texts, it is natural to find the structure *copula* + *definiendum* + *definiens*, for instance, "son documentos públicos los autorizados por un notario [...]" (*are public documents those authorised by a public notary**), where the *copula* is "son", the *definiendum* is "documentos públicos" and the *definiens* is "los autorizados por un notario". The *copula* can also be expressed with other verbs, commonly pronominal structures, as for instance "se considera" (*it is considered as*), "se entiende por" (*it is understood as*).

Notwithstanding, and depending on the legislator, other forms can also prevail in a text. For instance: "there is/exists" + *definiendum* + a condition expressed by "when", "if", "if and only if", or similar. In article 392 of the Spanish Civil Code, we can find the following definition: "hay comunidad cuando la propiedad de una cosa pertenece a varias personas" (*there exists a community when the property of something belongs to several people*). And it is also not uncommon to find definitions introduced by some punctuation marks, that is, *Definiendum: Definiens*, or in apposition to the *Definiendum* between commas.

4 Experiment

As theoretical foundation, we rely on the study of definitions in Spanish legal documents by [7]. Taking into account the different types of definitions that can be found in a legal text, we have identified the most standard patterns for legal definitions in Spanish (see Table 1).

Table 1. Standard patterns with examples.

	Spanish Patterns (and English translation)	Examples in English
1	Hay/Existe + definiendum + cuando/siempre/if + definiens (There is/exists + definiendum + when/if + definiens)	*There is community when there are more than five people.*
2	Se considera + definiendum + definiens (It is considered + definiendum + definiens)	*It is considered collective agreement an agreement that...*
3	Definiendum + es/son + definiens (Definiendum + is/are + definiens)	*Workers are people that work.*
4	Es/son + definiendum + definiens (It/They is/are + definiendum + definiens)	*It is a worker the person who works.*
5	Se define + definiendum + como + definiens (It is defined + definiendum + as + definiens)	*It is defined worker as a person who works.**
6	Se define + como + definiendum + definiens (It is defined + as + definiendum + definiens)	*It is defined as worker a person who works.*
7	Se entiende que + verb + definiendum + definiens (It is understood that + verb + definiendum + definiens)	*It is understood that there is a collective contract when...*

Our object of study is one of the most representative texts of Spanish labour law, the Spanish Workers' Statute[3]. As usual, when applying linguistic patterns, we have previously performed a part-of-speech tagging of the text with CoreNLP[4] and spaCy[5].

To identify the *definiendum*, we have followed two different strategies: 1) self generated by POS tokens and 2) helped by an Automatic Term Extraction tool. In the first case, we have generated sequences of terminological patterns based on each token's part of speech (see Table 2). In the second case, we have made use SketchEngine[6] [6], a widely used corpora managing tool that also implements cutting-edge automatic terminology extraction (ATE) algorithms. Our hypothesis is that by relying on the identification of the most relevant terms in the text, we may reduce the number of false positive results obtained, which identify as *definiendum* structures that do not correspond to any legal concept.

After executing the two different approaches, we then compare the results obtained to check if previous terminology work can help in definition extraction, or if, on the contrary, identifying terms with part-of-speech tags works better in this specific case.

Table 2. Term patterns with approximate translations in English

	Term patterns	Term examples (with approximate English translation)
1	haber + noun	ley (law)
2	denominar + noun + conj + noun	empleador o empresario(employer or businessman)
3	considerarán + noun	familiares (familiar)
4	considerarán + noun + adj	relaciones laborales (labour relations)
5	considerará + noun + adp + noun	centro de trabajo (workplace)
6	considerará + adp + noun + adj	de carácter colectivo (collective character)
7	considerará + verb + det + noun	terminado el contrato (finished contract)
8	entender + adj + adv	prorroga automática (automatic extension)
9	entender + noun + adj	grupo profesional (professional group)
10	entender + noun + adp + noun	reducción de jornada (reduction of working hours)
11	entender + verb + adp + det + noun + adj	excluida de el ámbito laboral (excluded from the workplace)
12	entender + adj + adp + noun + adj	celebrado por tiempo indefinido (for indefinite period)
13	entender + adj + conj + adp + noun	nulos y sin efectos (null and void)
14	entender + verb + noun + adj	causas técnicas (technical reasons)
15	entender + verb + noun + adp + det + noun	insolvencia del empresario (insolvency of the employer)
16	entender + det + noun + aux + adj	la disminución es persistente (persistent decrease)
17	denominar + propn + adp + propn	código de trabajo (work code)
18	denominarán + noun	empresario (entrepreneur)
19	denominarán + noun + adj + adp + det + adj	personalidad jurídica de el contratante (legal entity of the contractor)
20	denominarán + noun + adp + det + noun	trabajadores de el contratista (workers of the contractor)
21	denominarán + noun + adp + noun	semanas de disfrute (weeks of leave)
22	denominarán + punce + det + noun + conj	el contratista o subcontratista (contractor or subcontractor)

[3] https://www.boe.es/eli/es/rdlg/2015/10/23/2/con
[4] https://stanfordnlp.github.io/CoreNLP/
[5] https://spacy.io/
[6] https://www.sketchengine.eu/

5 Results and Discussion

The main issue common to both approaches is, as expected, the main limitation of linguistic patterns: the great number of false positives retrieved. However, our hypothesis was confirmed: when relying on the previously extracted terms, the number of false positives decreased. Still, we have encountered specific issues in each approach.

As for the definitions extracted with terminological patterns, the major issue concerns the identification of non-terminological structures, such as "en su totalidad" (on the whole) and "misma titulación" (same titulation). Such issues are mainly caused by the POS-tagging tools applied, that may fail when assigning tags. Regarding those definitions retrieved after a previous terminology extraction task, we also need to rely on the quality of the performance of the automatic term extraction tool, which may also wrongly identify tokes as terms. For instance, we retrieved the following definition: *A los anteriores efectos, se considerará salario la cantidad reconocida como tal en acto de conciliación o en resolución judicial por todos los conceptos a que se refiere el artículo 26*. The term defined in this case is "salario" (salary). However, since the ATE tool retrieved "salario la cantidad" (salary the amount) as a more specific term, this is the one wrongly associated to the definition.

Therefore, the automatically extracted terms need to be post-processed, either by humans or with an automatic post-processing approach to avoid this kind of mistakes. Similarly, we also retrieved: *Si en la empresa no hubiera ningún trabajador comparable a tiempo completo, se considerará la jornada a tiempo completo prevista en el convenio colectivo de aplicación o, en su defecto, la jornada máxima legal*. The *definiendum* in this case is "jornada a tiempo completo"; however, this term does not appear in our term list, and our patterns wrongly matched this definition with the term "jornada", that is a broader concept.

A different issue encountered occurs when we find part of the term before the copula and part of the term after the copula. For instance in this defintion of the term fair dismissal, we have the term "despido" (dismissal) before the copula, and the modifying adjective "procedente" (fair) after the copula. See: *El despido se considerará procedente cuando quede acreditado el incumplimiento alegado por el empresario en su escrito de comunicación*. A human can reason that the term defined is "despido procedente", but our algorithm cannot. Therefore, in this sentence, we have different term choices: *How then can a machine reason which of the terms is being defined?* This is doubtlessly a very interesting challenge to work on.

To evaluate the value of this work, we have taken the definitions we have correctly extracted and looked up their corresponding terms in the most important source of legal lexicographic knowledge in Spain: Pan-Hispanic Dictionary of Legal Spanish (*Diccionario Panhispánico del Español Jurídico,*

DPEJ henceforth)[7]. We consider that this work could be complementary to such lexicographic resources in two ways: first, by providing additional terms and their definitions, and second, by providing definitions that are slightly different than the ones contained in those resources, but that are contextualised in certain legal documents (as the case of the Workers' Statute). The following are examples of those:

- Trabajador nocturno (night worker): *Para la aplicación de lo dispuesto en el párrafo anterior, se considerará trabajador nocturno a aquel que realice normalmente en periodo nocturno una parte no inferior a tres horas de su jornada diaria de trabajo, así como a aquel que se prevea que puede realizar en tal periodo una parte no inferior a un tercio de su jornada de trabajo anual.*
- Familiar (familiar): *Se considerarán familiares, a estos efectos, siempre que convivan con el empresario, el cónyuge, los descendientes, ascendientes y demás parientes por consanguinidad o afinidad, hasta el segundo grado inclusive y, en su caso, por adopción.*
- Código de Trabajo (Work Code): *El Gobierno, a propuesta del Ministerio de Empleo y Seguridad Social, recogerá en un texto único denominado Código de Trabajo, las distintas leyes orgánicas y ordinarias que, junto con la presente, regulan las materias laborales, ordenándolas en títulos separados, uno por ley, con numeración correlativa, respetando íntegramente su texto literal.*

An additional advantage of performing definition extraction over legal texts is that we can identify when the same term has been defined differently in several legal documents, for comparison purposes, and also, when several definitions are available for the same term in the same document, as is the case of "centro de trabajo", work centre, in the document at hand:

1. *A efectos de esta ley se considera centro de trabajo la unidad productiva con organización específica, que sea dada de alta, como tal, ante la autoridad laboral.*
2. *En la actividad de trabajo en el mar se considerará como centro de trabajo el buque, entendiéndose situado en la provincia donde radique su puerto de base.*

This kind of "contextual enrichment" seems highly valuable, since lexicographic resources such as the DPEJ usually contain only one and more generic definition.

6 Conclusions and Future Work

In this experiment, we have tackled the case of standard definitions for Spanish legal texts. Our approach relies on part-of-speech taggers, that have

[7] https://dpej.rae.es/

shown imperfect for Spanish. Research on more accurate taggers and lemmatizers is therefore a requirement. Other issues have been thrown by our own terminological patterns, that we will need to review and adjust; as well as the definitional patterns, that sometimes generated wrong results. With regard to the ATE tool applied, SketchEngine, we observed that several of the terms identified are not completely correct or not much relevant in the text. Thus, another step to take is to test the performance of different ATE tools, which may apply statistical or linguistic approaches, and also to research on a post-processing algorithm to remove noise from that terms. A curated term list would surely improve the extracted definitions.

On the other hand, in this experiment we have left aside the so-called *non-standard legal definitions*. Such definitions can be divided into four different groups: 1) indirect definitions, 2) conditional definitions, 3) incidental definitions and 4) meta-linguistic definitions. The first group includes regular definitions but expressed in a more indirect or verbose manner. The second group refers to definitions that have to comply with more than one condition: if P, then X is A only if X is B. Incidental definitions are those that have been exposed without intention, while exposing another statement. Finally, meta-linguistic definitions are those in which the *definiendum* can be both *definiendum* and *definiens*. Examples on these types of definitions can be found in [7].

It goes without saying that this kind of definitions are much more difficult to extract that the standard ones. In fact, we have also tried to apply linguistic patterns with little success (for the moment). A clear future step is to work on the automatic identification of non-standard definitions in legal texts. A preliminary idea would involve relying on the manual annotation of non-standard definitions of the Spanish Workers' Statue to train a machine/deep learning engine to recognise them. We do not, however, completely discard the use of linguistic patterns as well. A combination of both could also turn into an interesting and effective approach.

References

1. Borg, C., Rosner, M., Pace, G.: Evolutionary algorithms for definition extraction. In: Proceedings of the 1st Workshop on Definition Extraction. pp. 26–32 (2009)
2. Del Gaudio, R., Batista, G., Branco, A.: Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. Natural Language Engineering **20**(3), 327–359 (2014)
3. Espinosa-Anke, L., Ronzano, F., Saggion, H.: Weakly supervised definition extraction. In: Angelova G, Bontcheva K, Mitkov R, editors. International Conference on Recent Advances in Natural Language Processing 2015 (RANLP 2015); 2015 Sept 7-9; Hissar, Bulgaria. Stroudsburg: ACL (Association for Computational Linguistics); 2015. p. 176-85. ACL (Association for Computational Linguistics) (2015)

4. Iftene, A., Trandabăţ, D., Pistol, I.: Grammar-based automatic extraction of definitions and applications for romanian. In: Proceedings of RANLP workshop" Natural Language Processing and Knowledge Representation for eLearning environments. pp. 19–25 (2007)
5. Jacquemin, C.: Spotting and discovering terms through natural language processing. MIT press (2001)
6. 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)
7. Marín, R.L.H.: Definiciones en el derecho. Anuario de filosofía del derecho (11), 367–380 (1994)
8. Przepiórkowski, A., Degórski, Ł., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kubon, V., Wójtowicz, B.: Towards the automatic extraction of definitions in slavic. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. pp. 43–50 (2007)
9. Saggion, H.: Identifying definitions in text collections for question answering. In: LREC (2004)
10. Walter, S.: Linguistic description and automatic extraction of definitions from german court decisions. In: LREC (2008)
11. Westerhout, E.: Definition extraction using linguistic and structural features. In: Proceedings of the 1st Workshop on Definition Extraction. pp. 61–67 (2009)