November 2019

Extracting terminologies in the legal domain: a syntactic pattern-based approach for Spanish

Mariano Rico [1], Pablo Calleja , Patricia Martín and Elena Montiel
Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM), Spain

Abstract. In this preliminary work we have adapted an English regular expression into Spanish and have run it over a corpus of legal documents in the labour law domain to automatically identify legal terms. The syntactic patterns of the terms obtained have been compared against the syntactic patterns of a human-made glossary in a closely-related domain (the TERMCAT collective agreements glossary). Results suggest that a small number of patters are able to retrieve 91% of all terms in the corpus.

Keywords. automatic term extraction, law domain, syntactic patterns, Spanish

1. Introduction

Automatic Term Extraction (ATE) is a Natural Language Processing (NLP) task aimed at identifying the core vocabulary of a specialized domain [4] by means of the computerized analysis of text corpora. ATE is also known as Terminology Extraction, Terminology Mining, Term Recognition, Glossary Extraction, Term Identification and Term Acquisition. Multi-word terms are usually referred to as *keyphrases*. ATE is a challenging area with applications in information retrieval and document-related tasks such as categorization, clustering and summarization [1].

This area, as many other Computer Science areas, has had a revolution due to the emergence of *deep learning*, a machine learning approach based on artificial neural networks which achieves results similar to humans (or even better) in areas like *voice to text* or *image recognition*. However, despite the excellent results achieved [9] by deep contextual language models such as ELMo [8] and BERT [2] on keyphrase extraction, they also suffer from the Achilles heel of deep learning: results cannot be explained. These systems are trained with lots of positive and negative cases, and given a new text, they produce a set they produce a set of terms, but there is no explanation for the results. For this reason, in this preliminary work we have decided to focus on traditional techniques, so that we can move later on to deep learning methods.

[1] Corresponding Author: Mariano Rico, Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid (UPM), Campus de Cantoblanco sn, Boadilla del Monte, 28660 Madrid, Spain; E-mail: mariano.rico@upm.es.

The traditional pipeline for ATE is (1) compilation of a corpus representative of the subject of study, (2) identification of multi-word units as a unique concept (known as *unithood*), (3) measurement of the likelihood that the extracted units are valid terms for the domain (known as *termhood*), (4) detection of different linguistic realizations for the same concept (*term variants*), and (5) evaluation and validation of the terms extracted to compare them to the ones identified by a human expert.

Following this pipeline, we have used a corpus of the Spanish legal domain [6], containing 0.5 million words (tokens), compiled in the context of the Lynx project (`http://www.lynx-project.eu/`), a European innovation action that aims at creating a knowledge graph in the legal domain. In order to identify the *unithood* we have based on the evidences that point out that **noun phrases**, that is, phrases that have nouns as their head or perform the same grammatical function as such phrases, are the predominant grammatical structure in terms. Additionally, these studies report on the fact that these noun phrases have 2 or more words in 85% of the terms extracted in a technical domain [7]. Thus, we have focused on noun phrase patterns that follow a certain syntactical structure (described in section 2), and make a proposal for the Spanish language. The reminder tasks in the traditional ATE pipeline (termhood, term variants detection and validation) are out of the scope of this short paper.

To validate the proposed patterns for Spanish we have used a human made terminology glossary for the legal domain (specifically, collective negotiations) and compared the results (as explained in section 3).

For the sake of reproducibility, we have provided a web page with detailed technical information and data at `http://nlp.linkeddata.es`

2. Syntactic Patterns for Noun Phrases in Spanish

Much has been written about noun phrase patterns in English [5,10,3] but, to the best of our knowledge, there are no syntactic pattern studies of the sort for Spanish.

Some of the first studies on noun phrase patterns in English are the ones from Justeson & Katz (1995) [5] which have the form of this regular expression: $(A|N)^{*}N(PD^{*}(A|N)^{*}N)^{*}$.

Regular expressions allow to comprise a set of patterns. This regular expression can be verbalized as “any number (including zero) of A (adjective) or N (noun) followed by a N and any number (including zero) of the block P (preposition) followed by any number (including zero) of D (determiner) followed by any number (including zero) of A or N, followed by a N”. The regular expression comes from a grammar with the following production rules:

$$BaseNP \rightarrow (Adj|Noun)^{*} \;\; Noun \tag{1a}$$

$$PrepPhrase \rightarrow Prep \;\; Det^{*} \;\; BaseNP \tag{1b}$$

$$NounPhrase \rightarrow BaseNP \;\; PrepPhrase^{*} \tag{1c}$$

This grammar is easily interpretable and can generate patterns of several lengths. For example, for length 1 we have the pattern *N* (e.g. salary), for length 2 we have *AN* (e.g. competitive salary) and *NN* (e.g. framework agreement), etc. For example, limiting the length to 5, this grammar produces 55 patterns. If we consider longer term lengths, the number of patters grows exponentially.

In this work we have adapted this English grammar to the Spanish rules and generated a new regular expression. Basically, we have modified production rule 1a, the one for BaseNP. This rule accounts for the basic structure of noun phrases in Spanish, which consists of the head noun going in the first position commonly followed by an adjective, noun, or prepositional phrase. The remaining rules do not change. The resulting grammar for Spanish is:

$$\begin{aligned} BaseNP &\rightarrow Noun \quad (Adj|Noun)^* \\ PrepPhrase &\rightarrow Prep \quad Det^* \quad BaseNP \\ NounPhrase &\rightarrow BaseNP \quad PrepPhrase^* \end{aligned} \tag{2}$$

The regular expression for grammar 2 becomes into $N(A|N)^*(PD^*N(A|N)^*)^*$. For length 1 we have N (e.g. salario) for length 2 we have NA (salario competitivo) and NN (acuerdo marco), etc. We have some long patterns with high frequency, like ”boletín oficial de la comunidad de madrid” (length 7).

3. Syntactic patterns from human-made terminologies

In the previous section we analyzed a simple grammar for Spanish Noun Phrases. We can see that even for simple grammars, the number of syntactic patterns grows quickly, and several questions arise: analyzing manually created terminologies, (1) do we really find this high variety of patterns in the terms identified? and (2) how long are the terms identified, that is, how many tokens do they have?, how common are such long patterns?

Aimed at answering these questions we have analyzed a manually created terminology (dataset of terms) of collective agreements in Spanish available from the Lynx project data portal (TERMCAT[2]). This terminological glossary belongs to the TERMCAT, the Catalan Center for Terminology, and was available in Spanish and Catalan as part of their ”open terminology” project. In the context of the Lynx project, this dataset was converted into RDF, and is now available online at[3]. It comprises 719 terms, with 1,718 tokens. Figure 1 shows all the patterns identified in the glossary, ordered by frequency. The vertical axis shows the accumulative sum of the instances found for each pattern. That is, the pattern *NA* is the most frequent one with 32% of the terms, the next pattern in terms of frequency is *N*, and the sum of both covers 52% of the terms. The sum of all the patterns covers 100% of the terms.

We can observe in this figure that with only 12 patterns we cover 91% of the terms in the selected terminology. Also we can see that very long patterns are possible. For instance, the pattern NPNPDNPNAA (length 10) is found for the term “contrato de trabajo para la realización de trabajos fijos discontinuos”.

[2] See `http://data.lynx-project.eu/dataset?q=termcat`

[3] See `http://data.lynx-project.eu/dataset/tcnes`

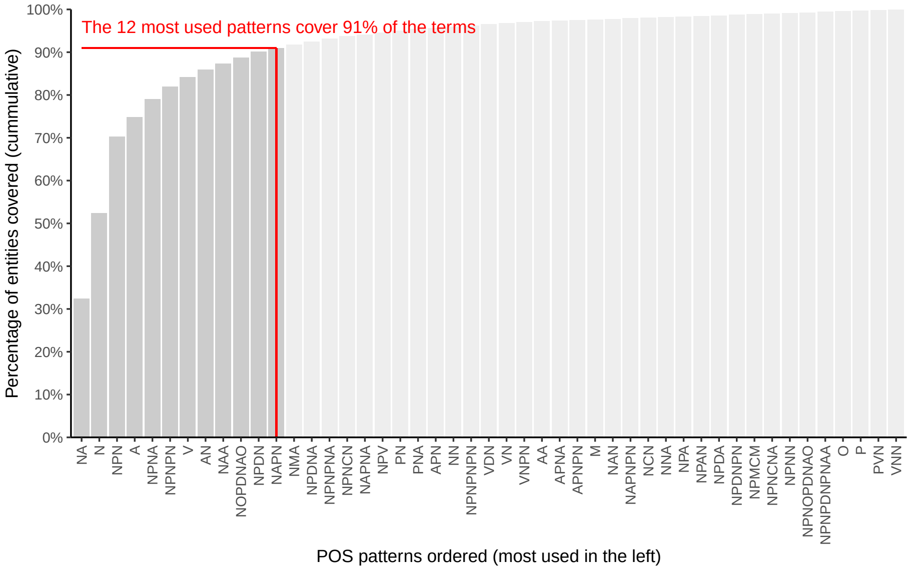

Figure 1. Syntactic patterns in the dataset (manually created) for legal domain (Spanish) found in TERMCAT.

4. Evaluating patterns against corpus of legal domain in Spanish

The next step was to use the patterns studied in the previous section against a corpus specific of the legal domain (labour law and collective agreements in Spain) [6].

The results of this study show that (1) the simpleNP for Spanish shown in section 2 generates good results but does not cover terms of length 10, (2) the union (logical or) of the first 12 patterns in the TERMCAT dataset produces very good results, but can miss some of the less frequent long patterns, such as ”boletín oficial de la comunidad de madrid” and (3) the union (logical or) of all the patterns in the TERMCAT dataset identifies infrequent long patterns, that although linguistically valid, do not correspond with real valid terms (for instance ”oficial de la comunidad de madrid”).

5. Future work

We plan to perform similar experiments in other sub-domains within the legal domain, and have results validated by terminology experts in the legal area. We also plan to extend this work using *word embeddings*, which have shown an excellent performance [9] on keyphrase extraction.

References

[1] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, 2019.

[3] A. Handler, M. Denny, H. Wallach, and B. O'Connor. Bag of what? simple noun phrase extraction for text analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 114–124, 2016.

[4] K. Heylen and D. De Hertog. Automatic term extraction. *Handbook of Terminology*, 1(01), 2015.

[5] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(1):9–27, 1995.

[6] P. Martın-Chozas and P. Calleja. Challenges of terminology extraction from legal spanish corpora. In *Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance. CEUR vol. 2309*, 2018.

[7] H. Nakagawa and T. Mori. A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology*, volume 14, pages 1–7. Association for Computational Linguistics, 2002.

[8] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[9] D. Sahrawat, D. Mahata, M. Kulkarni, H. Zhang, R. Gosangi, A. Stent, A. Sharma, Y. Kumar, R. R. Shah, and R. Zimmermann. Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings. *arXiv preprint arXiv:1910.08840*, 2019.

[10] T. Vu, A. Aw, and M. Zhang. Term extraction through unithood and termhood unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.