Linguistic Legal Concept Extraction in Portuguese

Alessandra Cid [a] Alexandre Rademaker [b] Bruno Cuconato [c] Valeria de Paiva [d]

[a] *FGV/Direito Rio and FGV/EMAp*
[b] *IBM Research and FGV/EMAp*
[c] *FGV/EMAp*
[d] *Nuance Communications*

Abstract. This work investigates legal concepts and their expression in Portuguese, concentrating on the "Order of Attorneys of Brazil" Bar exam. Using a corpus formed by a collection of multiple-choice questions, three norms related to the Ethics part of the OAB exam, language resources (Princeton WordNet and OpenWordNet-PT) and tools (AntConc and Freeling), we began to investigate the concepts and words missing from our repertory of concepts and words in Portuguese, the knowledge base OpenWordNet-PT. We add these concepts and words to OpenWordNet-PT and hence obtain a representation of these texts that is mostly "contained" in the lexical knowledge base.

Keywords. wordnet, law, legal informatics, lexical resources

1. Introduction

The "Order of Attorneys of Brazil" (in Portuguese 'Ordem dos Advogados do Brasil' or OAB), the Brazilian Bar association, administers a bar examination nationwide three times a year. The exam is divided in two stages. The first consists of 80 multiple choice questions covering several disciplines. The candidate must score at least 40 questions correctly to proceed to the second part of the exam. Success in the examination allows one to practice in any court or jurisdiction of the country.

We would like to use Natural Language Processing (NLP) tools to develop a computer system capable of providing question-answering facilities, based on Brazilian laws and regulations. An ideal legal system would take a question `Q` in natural language and a corpus of all legal documents in a given jurisdiction `LawCorpus`, and would return both a correct answer (easier if using multiple choice) and its legal foundation. However, this is too broad and too hard: we hope to provide a sample corpus (a subset of `LawCorpus`) with a single processed law, to see how far we can get the processing done.

Previous work [6] on a corpus constructed from multiple choice questions, attests to the suitability of the data obtained from the OAB Bar questions. The data from OAB's previous exams and their answer keys were cleaned and prepared for processing, and a simple question answering system, targeting the exams, based on shallow NLP methods

was described. The work in [7] improved the system by incorporating wordnet [1] data to its analysis, and started a preliminary effort to expand OpenWordnet-PT (OpenWN-PT or simply OWN-PT), our basic lexical resource, to the legal domain.

The expansion of a wordnet with legal terms was also investigated by [15] where legal vocabulary was added to the Italian Wordnet (ItalWordNet). Unfortunately, we could not get access to the final resource.

This work follows [7]. It is clear from inspection that the legal domain has many concepts and words that are only used within the legal profession. These concepts and words need to be added to OWN-PT, described below if this is to be used to reason about the law.[2]

2. OpenWordNet-PT

OpenWordnet-PT [5] is an open access wordnet for Portuguese, originally developed as a syntactic projection of Universal WordNet (UWN) [3]. OWN-PT has been constantly improved through linguistically motivated additions, manual and semi-automatic. Most of this work has been based on grammatical functions: we improved the verb lexicon [4], we provided nominalizations and their links to verbs [10], and we increased demonyms and gentilics [13], which was meant to break down the huge class of adjectives into smaller subsets. Regarding specialized domains, we did a preliminary study of Geological Eras [11], and here we are tackling another sophisticated and specialized field, the Law domain.

In order to deal with legal texts we need to expand OWN-PT with legal terms and *multiword expressions* (MWEs), that describe the field, but this is known to be a hard problem in linguistics [14]. Some of these are in Latin, such as *habeas corpus* or *data venia*. But most others are simply common Portuguese words, used in fixed expressions, which have more specific meanings. For example the expression *defensor público* could be used for someone who defends the public or someone who defends something in public, but it is mostly used to describe the attorney, appointed by the Estate to defend the interests of poor citizens, who are not able to pay for a lawyer. Some recent work, especially on English noun compounds [8], makes the point that MWEs can be compositional or non-compositional, conventionalized and not conventionalized.

It is clear that specific domains like Law require a big set of MWEs, both compositional (or not) and conventionalized (or not). Briefly we can say that *semantic non-compositionality* is the property of a compound whose meaning can not be readily interpreted from the meanings of its components. *Conventionalization* refers to the situation where a sequence of words that refer to a particular concept is commonly accepted in such a way that its constituents cannot be easily substituted for near-synonyms, because of some cultural or historical conventions. A large fraction of compounds are to some extent conventionalized, however we are interested in clear and well-known conventionalizations, which [8] refer to as "marked conventionalization". We assume that non-compositional compounds are by definition conventionalized, hence it only makes sense to consider conventionalization (or not) of compositional compounds.

[1] A wordnet is a lexical database that groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

[2] A longer version of this article was deposited in the `arXiv.org` [2].

3. Experiments

In order to identify relevant legal terms and MWEs and to analyze how this legal vocabulary can be incorporated to the OWN-PT, we describe three small experiments.

In our first experiment, we investigated the English terms in the Princeton WordNet (PWN) [9] synsets that were 'classified by' the synset {`08441203-n`: *jurisprudence, law*} in PWN and OWN-PT.[3] Our hypothesis was that, by translating the English terms that were already classified as legal vocabulary, we would incorporate important legal terms in Portuguese to the OWN-PT. We analyzed the terms that were on the topic, verifying the quality of the translations and seeing if we could translate the ones that were not. We reached the conclusion that the synsets related to {`08441203-n`: *jurisprudence, law*} were very specific to American Law and that by adding their translations to the OWN-PT we were not expanding it with relevant words for legal vocabulary in Portuguese. For example, we found several expressions for specific types of laws in English, such as "Gag Law" or "Blue sky law", that are not used in Portuguese. Given this situation, we moved on to our second experiment.

Our second experiment deals with a wholesale construction of a glossary of legal terms extracted from the OAB questions and from the three norms that are the base of the Ethics questions of the OAB exam. These three norms are: the Law 8906 of July of 1994, the 'Código de Ética da OAB' (the Ethics Code of the OAB) and the 'Regulamento Geral da OAB' (OAB's General Regulation). We analyzed these documents using AntConc [1], a corpus analysis toolkit for concordancing and text analysis. Using AntConc, we obtained a list of 6,890 bi-grams and tri-grams on the texts that occur more than 9 times. AntConc works over the raw text, without any linguistic annotation, and we have to filter n-grams that were clearly not MWEs. Two annotators filtered the list independently and we combined the results ending up with 430 MWEs candidates.

Instead of deciding which n-grams are true MWEs as opposed to simple collections of words that occur together, we used a simple test to classify each candidate as compositional or non-compositional and conventional or non-conventional. Is the meaning of this expression explained by the meanings of its parts? If not, then we think we have a non-compositional MWE. If the meaning of the expression is compositional, is it a title of an article in the Portuguese Wikipedia?[4] If yes, we reckon this is sufficient evidence to characterize a conventional MWE. If it is not a Wikipedia title, it may be that Wikipedia should have one such page and is missing it. Therefore, our process is an oversimplification that could be improved in the future. Finally, we identified the head words from expressions and added them to the proper synsets in OWN-PT, when they exist. If a head word suggest a concept that does not exist, we create a new synset in OWN-PT, placing it in the right position of the network of concepts, and assign the word to it. In both cases, the expression is finally added to a new synset, hyponym of the synset where its head word was added.[5]

---

[3] PWN contains many semantic relations between synsets, besides the most well-known `hyponym`, `hypernym`, and `antonym`, we also have the relation `classifiedByTopic` for grouping synsets into domains. All OWN-PT synsets have 1-1 mappings to Princeton WordNet synsets. Data is available at `http://wnpt.brlcloud.com/wn/`.

[4] We obtained a list of titles of all Portuguese pages from Wikipedia at `https://dumps.wikimedia.org/other/`.

[5] The list of MWEs and all data from the experiments will be made available at `http://github.com/own-pt/`.

	total	unique	no sense
Nouns	2629	727	190
Adjectives	634	234	60
Verbs	1167	330	16
Adverbs	268	77	32

Table 1. Analysis of Law 8906 by Freeling

In our third experiment, we investigated the lexical units of the Law 8906, one of the norms used in the second experiment. Since the Ethics part of the Bar examination is one of the most straightforward sections of the exam, it makes sense for us to make sure that the whole law is processed correctly and that all the required vocabulary is in place, before trying to relate the OAB ethics questions to their answers and justifications.

The experiment was carried out using Freeling [12], a well-known NLP library to analyze Brazilian Portuguese. We processed the Law 8906, investigating the results of the tokenization, lemmatization, part-of-speech (PoS) tagging and word sense disambiguation. We checked if all the content words are assigned to OWN-PT senses in the context of the articles of the law. This allowed us to evaluate how Freeling's modules could be adapted to process the law more accurately and enabled us to measure how many words belonging to the legal vocabulary were already on OWN-PT or needed to be added.

Some of Freeling's results after processing the law were expected. Since OWN-PT, just as PWN, does not cater for pronouns, determiners or prepositions, it did not have a meaning assignment for these cases. Freeling's lemmatization and PoS tagging modules are driven by a dictionary of word forms. The words that are not in Freeling's dictionary must have the lemma and part-of-speech tag guessed, which introduces some errors. For example, the Portuguese word *juizado* (court) was not in the dictionary, so the lemmatization of *juizados*, was wrongly ascribed as *juizados*. This was evidence that FreeLing's dictionary did not have it and we simply added it. The multiword expressions identified and added to OWN-PT must also be added to the Freeling locutions file. Other bugs are still under investigation, but the results obtained so far are summarized in Table 1, where we present basic statistics of Freeling's analysis of Law 8906. To obtain the unique totals we considered pairs (lemma, PoS tag), and we only considered that a word was missing a sense if it was tagged as the right PoS tag. Law 8906 comprise 87 articles summing up to 231 sentences and 10,242 tokens (1,508 unique types/words). Table 1 shows in the last column that we are still missing some words in OWN-PT.

4. Conclusion

This preliminary work investigates legal concepts and their expression in Portuguese. Using the corpus formed by the collection of multiple-choice questions in the exams, three ethics norms, language resources and NLP tools we began to investigate the concepts missing from our repertory of concepts and words in Portuguese, the knowledge base OWN-PT.

As for future work, we need to complete the expansion of OWN-PT that we started constructing. When the mappings are consistently investigated, we need to establish a

process to make sure that newer changes do not undermine the previous work, i.e. we need to establish test suites and regression tests. Finally we would like also to design and implement our own system for computing "entailment and contradiction detection" between the OAB examination questions and their answers and justifications (segments of text, or spans, in the laws that justify the answer of the question).

References

[1] Laurence Anthony. Antconc (version 3.5.7) [computer software], 2018. Available from `http://www.laurenceanthony.net/software`.

[2] A. Cid, A. Rademaker, B. Cuconato, and V. de Paiva. Linguistic Legal Concept Extraction in Portuguese. *ArXiv e-prints*, October 2018. https://arxiv.org/abs/1810.09379.

[3] Gerard De Melo and Gerhard Weikum. Towards a Universal WordNet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522. ACM, 2009.

[4] Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker. Making virtue of necessity: a verb lexicon. In *PROPOR – International Conference on the Computational Processing of Portuguese*, Tomar, Portugal, 2016.

[5] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. Published also as Techreport, http://hdl.handle.net/10438/10274.

[6] Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. Passing the Brazilian OAB Exam: Data preparation and some experiments. In Adam Wyner and Giovanni Casini, editors, *Legal Knowledge and Information Systems*, volume 302 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2017. 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017). Expanded version at https://arxiv.org/abs/1712.05128.

[7] Pedro Delfino, Bruno Cuconato, Guilherme Paulino Passos, Gerson Zaverucha, and Alexandre Rademaker. Using OpenWordnet-PT for Question Answering on Legal Domain. In *Global Wordnet Conference 2018*, Singapore, January 2018. to appear.

[8] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, 2015.

[9] Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[10] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. Extending a lexicon of portuguese nominalizations with data from corpora. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil, October 2014. Springer.

[11] Henrique Muniz, Fabricio Chalub, Alexandre Rademaker, and Valeria de Paiva. Extending wordnet to geological times. In *Global Wordnet Conference 2018*, Singapore, January 2018. to appear.

[12] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[13] Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. Gentle with gentilics. In *Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016)*, Slovenia, May 2016.

[14] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: a pain in the neck for NLP. In *Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Heidelberg, 2002. Springer Berlin.

[15] Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. Jur-wordnet. In *Proceedings of the 2nd International Global Wordnet Conference*, pages 305–310. Citeseer, 2004.