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" (in Portuguese 'Ordem dos Advogados do Brasil' or OAB) Bar exam and especially on its questions on Ethics, a subset of the multiple questions this national bar exam deals with. Using the corpus formed by a collection of multiple-choice exams, three norms related to the Ethics part of the OAB exam (the law 8906, OAB Ethics Code and OAB's General Regulation of the law 8906) and some language resources (Princeton WordNet and OpenWordNet-PT) and tools (AntConc and Freeling), we began to investigate the concepts missing from our repertory of concepts and words in Portuguese, as expounded in our basic lexical resource, the knowledge base OpenWordNet-PT. We add these concepts to a special glossary and hence obtain a representation of these texts that is mostly "contained" in the lexical knowledge base. These eventually can be used to calculate entailment and contradiction between passages, which would support the answer the questions automatically.

Keywords. wordnet, law, legal informatics, lexical resources

1. Introduction

Becoming a lawyer is a widely varied process around the world. Common to all jurisdictions are requirements of age and competence; some jurisdictions also require documentation of citizenship or immigration status. However, the most varied requirements are those surrounding the obtaining of the license to practice, whether it includes finishing a law degree, passing an exam, or serving an apprenticeship. Basic requirements vary from country to country.

In Brazil, 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 (e.g. Ethics, Human Rights, Philosophy of Law, Constitutional Law, Administrative Law, Civil Law, etc). The candidate must score at least 40 questions correctly to proceed to the second part of the exam, which requires answering four essay questions and a drafting project (writing a motion, opinion or claim document). Success in the examination allows one to practice in any court or jurisdiction of the country. However only 20% of those who take the exam are successful.

Our working hypothesis is that to answer correctly the multiple choice questions of the OAB Bar examination, candidates need to construct a model of the world, which consists of concepts that they learn when reading the text of the laws and jurisprudence associated with the questions. We would like to develop a computer system capable of mimicking some of these human abilities, using Natural Language Processing (NLP) tools.

In the next section, we describe some NLP tools, including our main lexical resource, the OpenWordNet-PT (abbreviated as OpenWN-PT or simply OWN-PT). Then we discuss extensions to deal with legal domains and the methodology we developed to deal with this kind of specific, very sophisticated domain.

2. The OAB Bar Examination

The OAB Bar examination provides a sensible benchmark to evaluate a system attempting to provide question-answering facilities, based on Brazilian laws and regulations. An ideal legal question-answering 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, i.e., which sections of which norms or laws provide support for the answer and why this is so. As stated this is too broad and too hard: we hope to provide a sample corpus (a subset of `LawCorpus`) with a single detailed law, to see how far we can get the processing to go. We then go on to describe a methodology for continuously improving our processing of the vocabulary of the law.

Previous work on the 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 obtained as PDF files from the official source at `http://oab.fgv.br/`, cleaned and prepared for processing with some scripts [7]. In [7] we also described a simple question answering system targeting the exams, based on shallow NLP methods. In [8] we improved the system by incorporating wordnet data to its analysis process, and started doing a very preliminary effort of expanding OWN-PT to the legal domain.

The expansion of a wordnet with legal terms was also investigated by [19] where legal vocabulary was added to the Italian Wordnet (ItalWordNet). Unfortunately, we could not get access to the final resource.

This work follows. It is clear from inspection (and previous work) that the legal domain has many concepts and words that are only used within the legal profession. If they are to be used to reason about the Law, these concepts and words need to be added to OWN-PT, our basic lexical resource, described below.

3. OpenWordNet-PT

The OpenWordnet-PT [6] is an open access wordnet for Portuguese, originally developed by Valeria de Paiva, Alexandre Rademaker and Gerard de Melo as a syntactic projection of Universal WordNet (UWN) [4]. When the project on OWN-PT started there was no open wordnet-like resource for Portuguese.

The process of building the OWN-PT used machine learning to construct relationships between graphs representing information coming from several versions of Wikipedia, as well as from open dictionaries. Starting as a projection at the level of the lemmas in Portuguese and their relationships, the OWN-PT has been constantly improved through linguistically motivated additions, manual and semi-automatic, making use of large corpora. This kind of construction, automatically started, but manually curated and improved, is the hallmark of our work using OWN-PT.

The OWN-PT has been developed since 2010 with the main objective of eventually serving as a lexicon for a proposed NLP system focused on logical reasoning, based on knowledge representations coming from language. The philosophy of OWN-PT is to maintain a close alignment with the original Princeton WordNet (PWN) [10], but to remove the biggest mistakes created by automated methods, using language tools and skills for this cleaning up task. One positive consequence of the close connection between PWN and OWN-PT is the latter ability to minimize the impact of lexicographical decisions on the separation or grouping of senses in a synset, as these decisions are, for the time being, delegated to PWN, to a large extent. We strive for precision, rather than coverage, as far as OWN-PT is concerned, and precision is surely needed when dealing with the legal domain.

4. Analyzing the legal vocabulary

We want to make sure that the basis created with OWN-PT is broad enough to allow us to deal with specific domains such as Law or Geology. Clearly these specific domains have specific vocabulary, both in terms of words that are not part of the everyday vocabulary, but specially in the use of expressions. In Law, there are several expressions that are not, as yet, part of the OWN-PT. Many common nouns are missing. A significant number of these are nominalizations such as *impetração* (a kind of filing), *postulação* (postulation), where the verbs *impetrar* and *postular* (to file, to postulate) are already in the OWN-PT lexicon.

Some of the missing expressions are adjectives, like *fundacional* (foundational) and *constitutivo* (constituent) coming from nouns *fundação* (foundation) and *constituição* (constitution), where sometimes PWN prefers not to list all the morphologically derived expressions. Still others are nouns that are nominalizations of adjectives, like *nulidade* (nullity), derived from *nulo* (null), where, again, morphology could play am important role.

Expressions such as the name of a law, e.g. *Estatuto da Advocacia* or the name of the professional association of lawyers in Brazil, the “Ordem dos Advogados do Brasil” (OAB) are essential synsets that needed to be created. These are expected, since the named Brazilian legal entities are clearly different from the American ones. We need synsets corresponding to the ones for *President of the United States* and *U.S. Congress*, for instance.

There are also other, more general legal expressions, called *multiword expressions* (MWEs), that really describe the field, but are harder to deal with. 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 on these multiword expressions, especially on English noun compounds [9,18], makes the point that multiword expressions can be compositional or non-compositional, conventionalized and not conventionalized. In [9], it is said in the introduction:

> The lack of practical data sets that can be used in the training and evaluation of multiword expression (MWE) related systems is a notorious problem [13,12]. It is partly due to the heterogeneous nature of MWEs, partly due to their frequency, and partly due to the unclear boundaries between MWEs and regular phrases. These issues have made the compilation of useful MWE data sets challenging, and any effort to create them invaluable.

It is abundantly clear that specific domains like Law require a bigger 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 [9] 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.

It is also clear that non-compositional MWEs are easier to spot: for example *má fé* (bad faith) has nothing to do with *fé* (faith) in its most used meaning of 'religious belief'. It simply means "in a deceiving way" and it is not specific to Law, but common currency both in Portuguese and in English, where the synset {`00753240-n`: *for double-dealing, duplicity*} has a gloss which reads *acting in bad faith; deception by pretending to entertain one set of intentions while acting under the influence of another*.

5. Experiments

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

In our first experiment, we investigated the English terms in the PWN synsets that were classified by the synset[1] {`08441203-n`: *jurisprudence, law*} in PWN/OWN-PT.[2]

[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] 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.

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 some of the terms that were on the topic, verifying the quality of the translations of the terms that were already translated 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 last three norms are: the law 8906 of July of 1994, the 'Código de Ética da OAB' (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. Since AntConc works over the raw text, without using any linguistic annotation, we had to filter n-grams that were clearly not MWEs. Two annotators filtered the list independently and we combined the results ending up with 430 candidates of MWEs.

Following [9], instead of deciding which n-grams are true MWEs as opposed to simple collections of words that happen to 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?[3] 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 expressions is finally added to a new synset, hyponym of the synset where its head word was added.[4]

In our third experiment, we investigated the lexical units of the law 8906, one of the norms used in the second experiment. Law 8906 describes the rights and obligations of lawyers and how they can advocate for their clients. 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 starting to analyze the law itself and before trying to relate the OAB ethics questions to their answers.

The experiment was carried out using Freeling [15], a well-known NLP library to analyze Brazilian Portuguese. We processed the Law 8906, investigating the results of

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

[4] 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

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 Princeton WordNet, does not cater for pronouns, determiners or prepositions, it did not have a meaning assignment for these cases. Freeling 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 introduce some errors. For example, the Portuguese word *juizado* (court) was not in the dictionary, so its lemmatization 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, so that tokens that are part of an MWE are joined enabling the word sense disambiguation module to associate the whole expression to an OWN-PT synset. Other bugs are still under investigation, but the results we 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 the pair (lemma,PoS tag), and we only considered that a word was missing a sense if it was tagged as the right PoS. Law 8906 comprise 87 articles summing up 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.

As a way of completing OWN-PT, one can discover easy synsets that needed completion. For example, the synsets {`01987341-a`: *reserved (marked by self-restraint and reticence; "was habitually reserved in speech, withholding her opinion")*} and *01988324-a*reserved (set aside for the use of a particular person or party), are very easy to complete. The word *reservado* is almost exactly the same as the English one and has the associated adverb *reservadamente* corresponding to the also empty synset {`00441649-r`: *reservedly (with reserve; in a reserved manner)*}. There are plenty of occasions where English and Portuguese conceptualize things differently and these require plenty of effort. Finding the easy synsets, where not only concepts, but even words are almost the same and making sure that these are complete in the lexical base is one of our main goals for OWN-PT.

6. Conclusions

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.

This initial work does not require a huge effort of evaluation, since we are mostly discovering word forms and senses that the lexical resource does not have, yet. We can count how many word forms were added to synsets, how many MWEs we had to create, but we have not found baselines to compare our work to, so far.

As maintainers and curators of OWN-PT this kind of work provides a way of dicing up a nice subset of synsets to look up and correct. Most of the work we have done so far on improving OWN-PT has been based on grammatical functions: we had a push to improve the verb lexicon [5], another push to provide nominalizations and their links [11], an effort to increase demonyms and gentilics [16], which was meant to break down the huge class of adjectives into smaller subsets. We have not done as much in terms of topics or semantic domains. We did a preliminary study of Geological Eras [14], but this was mostly to check the feasibility of merging an external ontology to the subjacent hierarchy of OWN-PT noun synsets. We also did a preliminary assessment of temporal related concepts [17], using HeidelTime [20]. But this is our first time tackling a sophisticated and specialized field like Law.

We are aware of several difficulties ahead. Firstly the judicial system in Brazil (based on Roman law) is very different from the 'Common Law' system in use in the US and UK, where most of the lexical resources we want to make use of, originate. This difficulty is discussed in [3], who wanted to use FrameNet [2] to create a corpus encompassing the whole judicial system in Brazil. They say that "This[their proposed] corpus is being planned to be representative of the entire legal production of Brazilian courts and Brazilian legislative houses, such as the laws published by the Brazilian Senate and the judicial decisions of the Federal Courts". We are less ambitious, we hope to produce a corpus of laws and regulations that allow us to answer the Ethics questions of our collection of OAB exams, at least to begin with.

Secondly we would like to produce a large glossary of legal terms that could be used for students actively taking the OAB exam, focusing on the multiple choice questions. There are several good juridical dictionaries in Portuguese and in English, but it seems to us that an open-source one, with relations of synonymy and antonymy, mostly based on the past OAB entrance examinations would be useful to students and professors of Law. Thirdly, we reckon that this project, although ambitious, would allow us to push on the direction we really want to work on, that is, the direction of reasoning with the contents of the legal texts. We plan to start using shallow methods, but also want to try our hand at deep logical representations, hybridized together with learning approaches, to try and detect entailment and contradiction between pieces of legal text.

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] Collin F. Baker and Hiroaki Sato. The FrameNet data and software. In Yuji Matsumoto, editor, *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, page 161–164, 2003.

[3] Anderson Bertoldi and Rove Luiza de Oliveira Chishman. The limits of using framenet frames to build a legal ontology. In *Proceedings of Ontobras*, 2011.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 132–141. Association for Computational Linguistics, 2012.

[13] Diana McCarthy, Bill Keller, and John Carroll. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 73–80. Association for Computational Linguistics, 2003.

[14] 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.

[15] 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.

[16] 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.

[17] Livy Real, Alexandre Rademaker, Fabricio Chalub, and Valeria de Paiva. Towards temporal reasoning in portuguese. In *Proceedings of 6th Workshop on Linked Data in Linguistics*, Miyazaki, Japan, May 2018.

[18] 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.

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

[20] Jannik Strötgen, Julian Zell, and Michael Gertz. Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 15–19, 2013.