Linguistic Legal Concept Extraction in Portuguese

Alessandra Cid [1,4]
Alexandre Rademaker [1,2]
Bruno Cuconato [1,2]
Valeria de Paiva [3]

JURIX, Groningen, 2018

[1] Applied Mathematics School, FGV, Brazil
[2] IBM Research, Brazil
[3] Nuance Communications, US
[4] Law School, FGV, Brazil

Topic

OAB Exams

OpenWordnet-PT

Jurix 2018

Conclusion

OAB Exams

- The OAB (Ordem dos Advogados do Brasil) Exam is the BAR Exam in Brazil
- Interesting problem for explore NLU (Natural Language Understanding) techniques.
- It provide an excellent benchmark for the performance of legal information systems, passing would signal capacity of legal reasoning comparable to human lawyers.
- Initial (shallow) processing, JURIX 2017, was just an starting point for further “deep” language processing.

OAB Exams

- Only in 2010 were the exams nationally unified.
- Two stages. We are working on the first stage, multiple choice questions. It has 80 multiple choice questions and each question has 4 options.
- In order to be approved, candidates need at least a 50% performance.
- Every year, there are 3 applications of the exam in the country.
- The exam has a global 80% failure rate. The most recent exam, July 2017, had the highest failure rate, 86% of the candidates failed.

OAB Exams

Questions per subject area and their performance rates:

area	#	(%)	**area**	#	(%)
Ethics	10	65	Constitutional Law	7	42
Consumer's Law	2	56	Civil Procedures	6	40
Children's Law	2	54	Philosophy	2	40
Criminal Procedures	5	47	Labor's Law Proc.	6	40
Regulatory Law	6	47	Criminal Law	6	38
Human Rights	3	47	International Law	2	37
Civil Law	7	44	Business Law	5	33
Environmental	2	43	Taxes	4	42
Labor's Law	5	42			

Data at `http://github.com/oab-exams`: 27 exams with 2220 questions from 2010 to 2018.

Approach i

- QA system. Given a question `Q` in NL and a corpus of legal documents in a given jurisdiction `LawCorpus`, return both a correct answer (easier if using multiple choice) and its legal foundation, i.e., which sections of which norms/laws provide support for the answer.
- As stated this is too broad and too hard: we provide a sample corpus (a subset of `LawCorpus`) with a single detailed law.

Approach ii

- Previous work on the corpus constructed from multiple choice questions attests (JURIX 2017) to the suitability of the data obtained from the OAB Bar questions. We also described a simple question answering system targeting the exams, based on shallow NLP methods.
- In GWC 2018, 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. Similar work in the Italian Wordnet too.
- This work. Legal domain has many concepts and words that are only used within the legal profession. If we want a system to reason about the Law, these concepts and words need to be added to OWN-PT, our basic lexical resource.

Topic

OAB Exams

OpenWordnet-PT

Jurix 2018

Conclusion

lexical resources for Portuguese?

- 6th mostly spoken language in the world or 7th Wikipedia
- Very few open source resources for PT, almost no connections at all

Linguistic resources are very easy to start working on, very hard to improve on and extremely difficult to maintain, as funding usually only works for new resources.

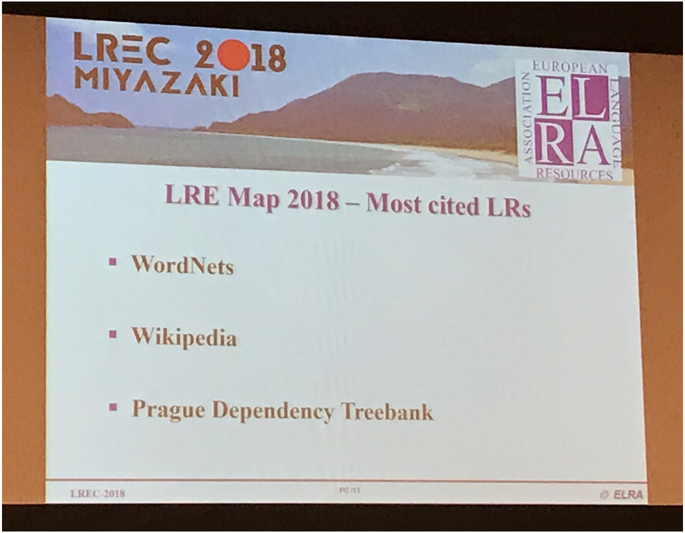
Trying to buckle the trend. Review of work in the last 8 years...

openwordnet-PT (own-PT)

- Wordnet is the most paradigmatic resource for English NLP
- Want a Portuguese Wordnet that is open access, downloadable and updateable, so that it can be improved by the community
- Especially interested in NLP for KR and automated deduction (our team)
- But also word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, question answering, etc.

`https://github.com/own-pt/openWordnet-PT`

Trends in LREC

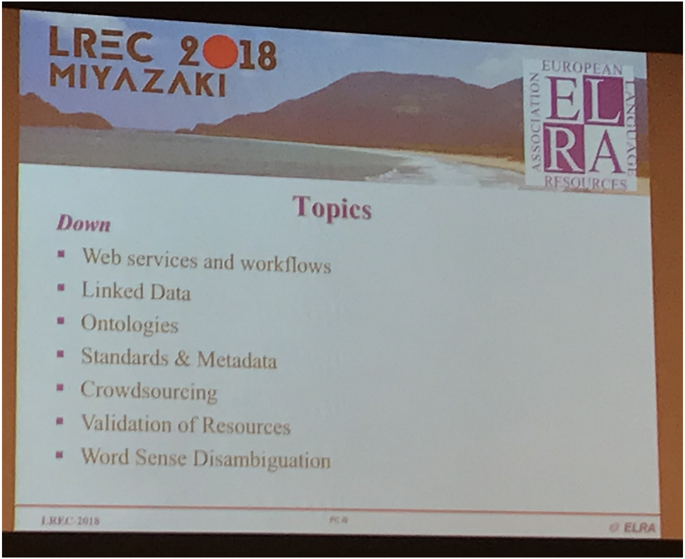


LRE Map 2018 – Most cited LRs

- WordNets
- Wikipedia
- Prague Dependency Treebank

LREC-2018 PC/11 © ELRA

Trends in LREC



LREC 2018 MIYAZAKI

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

Topics

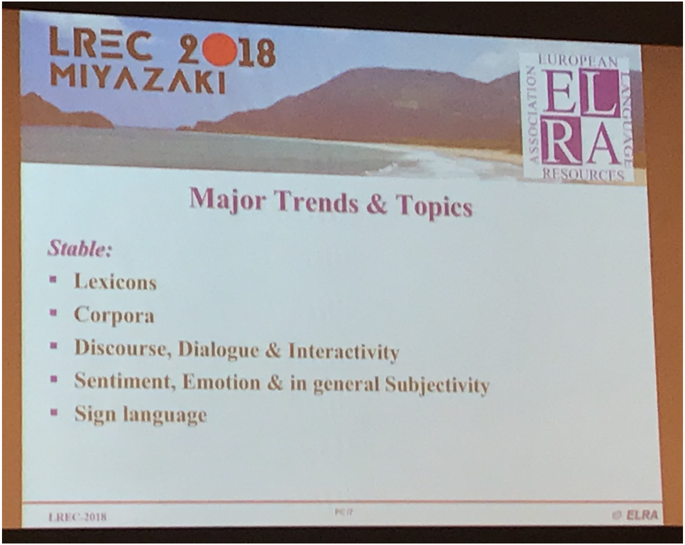
Down

- Web services and workflows
- Linked Data
- Ontologies
- Standards & Metadata
- Crowdsourcing
- Validation of Resources
- Word Sense Disambiguation

LREC-2018

© ELRA

Trends in LREC



LREC 2018
MIYAZAKI
EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

Major Trends & Topics

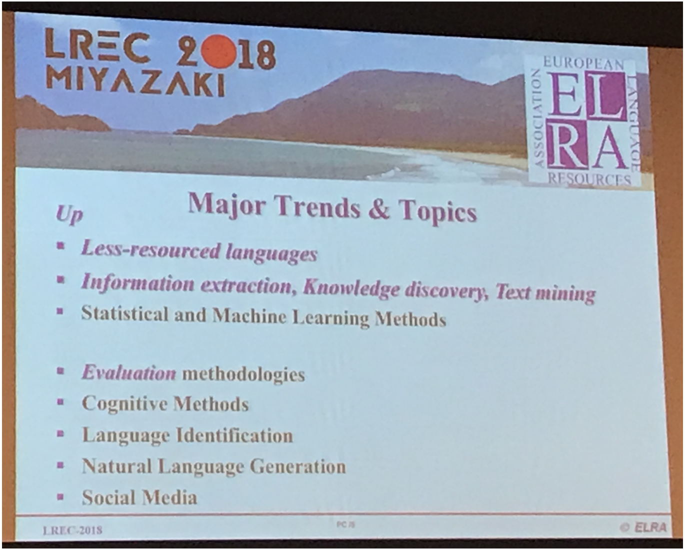
Stable:

- Lexicons
- Corpora
- Discourse, Dialogue & Interactivity
- Sentiment, Emotion & in general Subjectivity
- Sign language

LREC-2018

© ELRA

Trends in LREC



LREC 2018
MIYAZAKI
EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

Major Trends & Topics

Up

- *Less-resourced languages*
- *Information extraction, Knowledge discovery, Text mining*
- Statistical and Machine Learning Methods

- *Evaluation* methodologies
- Cognitive Methods
- Language Identification
- Natural Language Generation
- Social Media

LREC-2018 PC /5 © ELRA

a bit of history

- Initially a transformation and extension of data from the Universal Wordnet/MENTA (UWN/MENTA)
- machine learning to construct relationships between graphs from Wikipedia in several languages + machine readable dictionaries
- continuously improved through linguistically motivated additions and removals, either manual or semi-automatic, making use of large Portuguese corpora (DHBB, Bosque, etc.)
- two-tiered methodology: high precision for popular words, high recall for long tail
- Could be used for other languages: best for languages well represented on the internet and with reasonably large Wikipedia.

nominalizations: the nomlex-pt

- useful for linguistic research as well as for information extraction, basic example destruction/destroy
- an extension (incorporated) of OWN-PT, with links connecting deverbal nouns with their corresponding verbs.
- Bootstrap via manually created 2,000 entries via translation of the English NOMLEX. Next with corpora for expansions
- Useful to check issues with the coherence and richness of own-pt, e.g. “aviltar” (not common, missing in the “humilhar” synset) and “aviltamento”
- Further incorporation of the Princeton **morphosemantic** links that classify the **derivational** links.

Web (social) interface

- new social and collaborative interface implemented and deployed in 2016, “Seeing is Correcting”
- Simple interface → content perspicuous
- OWN-PT is also part of Open Multilingual Wordnet, and Global WordNet Foundation
- Many experiments, described in the website including
- Verb lexicon improvements, gentilics, morpholinks (many not finished, yet)
- OWN-PT part of FreeLing, Google Translate, BabelNet, Onto.PT

`http://openwordnet-pt.org`

applications

- Freeling
- Sentiment analysis using SentiWordnet for tweets for football
- DHBB, recently open-source. Biographical data requires good NER
- Comparison of wordnet-like resources for PT
- Linked open Data
- Universal Dependencies

An overview of Portuguese WordNets (2016)

Name	Creation		Update	Usage
	Synsets	**Relations**		
WN.PT	manual	manual	manual	closed
WN.BR	manual	transitivity	manual?	free synsets
MWN.PT	manual? trans.	transitivity	?	paid license
Onto.PT	RE,*clustering*	RE,*clustering*	automatic	free
OpenWN-PT	UWN project.	transitivity	semi-autom	free
UfesWN.BR	MT	transitivity	?	free
PULO	triangulation	transitivity	semi-autom	free

- Fully automatic construction approach leads to a larger resource.
- An intrinsic trade-off between the size of a wordnet and the accuracy and usefulness of the resource under scrutiny.

Linked open data: OWL and RDF

- Verifying Integrity Constraints of a RDF-based WordNet (2016).
- OWN-PT if freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
- Verifying integrity constraints of our openWordnet-PT against the ontology for Wordnets encoding.
- Consistency check of OWL and Integrity Constraints in RDF
- A DL knowledge base is comprised by two components, TBox and ABox.
- The TBox contains intensional knowledge (terminology).
- The ABox contains extensional knowledge (assertional).

universal dependencies

- It is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 60 languages.
- release UD 2.2
- github contribution, individual repositories in a single organization
- central documentation
- issues and discussions on docs.
- tools for validation, annotations etc.

`http://universaldependencies.org` and slides and UD-Portuguese-Bosque.

Next steps

Collaboration on the ILI from GWA.

Legal domain expansion following previous work on JurWordnet (Italian). The context is the project “Passing the Brazilian OAB (Bar) Exam”.

Corpora annotation. Some ideas from Francis Bond on “Teaching through Tagging – Interactive Lexical Semantics” – Detecting Meaning with Sherlock Holmes

Lexical semantics vs Lexical Resources?

HPSG/ERG predicates to PWN synsets. Work by Francis Bond. Can it be applied to Portuguese (if/when we have a PT Grammar)? Long term, motivated by IE in GeoScience/EN.

Topic

OAB Exams

OpenWordnet-PT

Jurix 2018

Conclusion

Legal Vocabulary i

- Several expressions are not in OWN-PT. Many common nouns are missing and 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 adjectives, like fundacional (foundational) and constitutivo (constituent) coming from nouns fundação (foundation) and constituição (constitution), where PWN also miss some morphologically derived expressions. Still others are nouns that are nominalizations of adjectives, like nulidade (nullity), derived from nulo (null).

Legal Vocabulary ii

- Names of laws, e.g. Estatuto da Advocacia or the name of the professional association of lawyers in Brazil, the “Ordem dos Advogados do Brasil” (OAB) are synsets that needed to be created. Named Brazilian legal entities are different from the American ones. We need synsets corresponding to the ones for President of the United States and U.S. Congress, for instance.

multiword expressions (MWE) i

- really describe the field, but are harder to deal with.
- Some are in Latin, *habeas corpus* or *data venia*. But most others are simply fixed expressions, which have specific meanings. *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 English noun compounds makes the point that MWE can be compositional or non-compositional, conventionalized and not conventionalized.

multiword expressions (MWE) ii

- We assume that non-compositional compounds are by definition conventionalized.
- Non-compositional MWEs are easier to spot – *má fé* (bad faith) has nothing to do with *fé* (faith) in its most used meaning of ‘religious belief’. It means “in a deceiving way” and it is not specific to Law, see 00753240-n.

multiword expressions (MWE) iii

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. 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.

Three experiments

The goal is to identify relevant legal terms and multiword expressions and how they can be incorporated to the OWN-PT.

1. English terms in the PWN synsets that were classified by the synset {jurisprudence, law}.
2. We seleceted a corpus and we obtained a list of 6,890 bi-grams and tri-grams on the raw texts that occur more than 9 times. Two annotators filtered the list independently and we combined the results ending up with 430 candidates of MWEs.
3. Lexical units of the preprocessed 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.

First Experiments

- Our hypothesis: by translating the English terms that were already classified as legal vocabulary, we would incorporate important legal terms in Portuguese to the OWN-PT.
- Check the translations of the terms that were already translated and seeing if we could translate the ones that were not.
- Conclusion: synsets 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. Several expressions for specific types of laws in English, such as “Gag Law” or “Blue sky law”.

Second Experiments i

- Legal terms extracted from the OAB questions and the three norms related to the Ethics questions of the OAB exams. Law 8906 of July of 1994, ‘Código de Ética da OAB’ (Ethics Code of the OAB) and ‘Regulamento Geral da OAB’ (OAB’s General Regulation).
- Using AntConc we obtained 6,890 bi-grams and tri-grams on the texts that occur more than 9 times. Raw text needed to be filtered.
- Two annotators filtered the list independently and we combined the results ending up with 430 candidates of MWEs.

Second Experiments ii

- We used a simple test to classify each candidate as compositional and conventional.
- If compositional, is it a title of an article in the Portuguese Wikipedia? If yes, sufficient evidence to take it as conventional MWE. If it is not in Wikipedia, it may be that Wikipedia is missing it. 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. In both cases, the expressions is finally added to a new synset, hyponym of the synset where its head word was added.

Thrid Experiment i

- Lexical units of the law 8906 – rights and obligations of lawyers. It comprise 87 articles summing up 231 sentences and 10,242 tokens (1,508 unique types/words).
- Using Freeling to process, 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.

Thrid Experiment ii

- 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 errors: juizado (court) was not in the dictionary, so its lemmatization was wrongly ascribed as juizados.
- The MWEs 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 WSD module to associate it to an OWN-PT synset.
- Simple cases, the word *reservado* is almost exactly the same as the English associated adverb *reservadamente*.

Thrid Experiment iii

- {reserved (marked by self-restraint and reticence; “was habitually reserved in speech, withholding her opinion”)} and
- {reserved (set aside for the use of a particular person or party)}.
- {reservedly (with reserve; in a reserved manner)}.

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

Topic

OAB Exams

OpenWordnet-PT

Jurix 2018

Conclusion

Conclusion i

- We investigated legal concepts and their expression in Portuguese. Using the corpus formed by the collection of multiple-choice questions in the exams, three ethics norms.
- Complete the expansion of OWN-PT that we started constructing. Test suites and regression tests.
- Evaluation. We are mostly adding word forms and senses. 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.

Conclusion ii

- So far improving OWN-PT based on grammatical functions: verb lexicon, nominalizations, demonyms and gentilics. We have not done as much in terms of topics or semantic domains. A preliminary study of Geological Eras, mostly to check the feasibility of merging an external ontology.
- 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.
- We hope to produce a corpus of laws and regulations that allow us to answer the Ethics questions of our collection of OAB exams (Universal Dependencies?).

Conclusion iii

- Produce a large glossary of legal terms that could be used for students actively taking the OAB exam. There are several good juridical dictionaries in Portuguese and in English, but no open-source one, with relations of synonymy and antonymy.
- Reasoning with the contents of the legal texts. Deep logical representations hybridized together with learning approaches, to detect entailment and contradiction between pieces of text. OAB exams questions and their answers and justifications.

Thank you and . . .