Using OpenWordnet-PT for Question Answering on Legal Domain

Pedro Delfino (FGV/EMAp and Direito Rio)
Bruno Cuconato (FGV/EMAp)
Guilherme P. Passos (IBM Research and UFRJ)
Gerson Zaverucha (UFRJ)
Alexandre Rademaker (IBM Research and FGV/EMAp)
The Problem

• The OAB (Ordem dos Advogados do Brasil) Exam is the BAR Exam in Brazil

• The OAB exams provide an excellent benchmark for the performance of legal information systems, passing the exam would signal capacity of legal reasoning comparable to human lawyers.

• Interesting problem for explore NLU (Natural Language Understanding) techniques.

• The current (shallow) processing is just an starting point for further “deep” language processing. Mainly results so far: data exploration and curation!

• No ML approach! KB and reasoning! We really want to understand the problem and explain the results using GOFAI.
The OAB Exam

• Only in 2010 were the exams nationally unified.
• Two stages. We are working on the first stage, multiple choice questions.
• 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.
Questions per subject area and their performance rates

<table>
<thead>
<tr>
<th>area</th>
<th>#</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethics</td>
<td>10</td>
<td>65</td>
</tr>
<tr>
<td>Consumer’s Law</td>
<td>2</td>
<td>56</td>
</tr>
<tr>
<td>Children’s Law</td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>Criminal Procedures</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>Regulatory Law</td>
<td>6</td>
<td>47</td>
</tr>
<tr>
<td>Human Rights</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>Civil Law</td>
<td>7</td>
<td>44</td>
</tr>
<tr>
<td>Environmental</td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>Labor’s Law</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>Constitutional Law</td>
<td>7</td>
<td>42</td>
</tr>
<tr>
<td>Civil Procedures</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>Philosophy</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>Labor’s Law Proc.</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>Criminal Law</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>International Law</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>Business Law</td>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td>Taxes</td>
<td>4</td>
<td>42</td>
</tr>
</tbody>
</table>
The OAB Exam: ethics

• In the exam, Ethics means questions about the rights, the duties and the responsibilities of the lawyer. Ethics questions also have a high performance rate.

• This is the simplest part of the exam with respect to the legal foundation of the questions.
  • Almost all the questions on Ethics are based on the Federal law 8906 from 1994 (89 articles) and well designed normative document.
  • A minor part of the questions on Ethics is related to two other norms: (i) OAB General Regulation (169 articles) and (ii) OAB ethics code (66 articles).

• It is important to note that these two norms are neither legislative nor executive norms. Indeed, they are norms created by OAB itself.
Data preparation

- Exams in PDF files.
- From PDF to text using Apache Tika.
- Manually revised the obtained text, and introduced markup to signal the beginning of questions, alternatives and some meta data about them.
- Simple Parser to XML
- The final data comprises 25 exams totaling 2061 questions.
- Making experiments reproducible!

https://github.com/own-pt/oab-exams
Norms preparation

• How to encode the norms related with the golden set? Federal law 8906 from 1994, the OAB general regulation and the OAB ethics code.

• The LexML is a joint initiative of the Civil Law legal system countries seeking to establish open standards for the interchange, identification and structuring of legislative and court information.

• We used the LexML (XML schema called “LexML Brasil”). Also related to Akoma Ntoso and EUR-Lex schemas.

https://github.com/lexml (parser from docx to XML) and http://www.lexml.gov.br
Golden Set

• We sampled 30 questions on Ethics for analysis (from the 210 questions in all exams)
• one of the authors manually identified the articles in the laws that justify the answer, creating our golden data set.
• usually, one article on the federal law 8906 was enough to justify the answer to the questions (15 questions).
• Less often, the justification is in OAB General Regulation (3 questions), or on the OAB Ethics Code (8 questions).
• 3 questions were justified by two articles in law 8906, and another in a jurisprudence from the Superior Court of Justice about an article from the law 8906.
Experiments: previous work (Jurix, Dec 2017)

Main simplification: a question is justifiable by one article.

1. The system receive a question statement and its multiple alternatives, and we wanted it to retrieve the right answer along with its justification in a given legal norm.

2. Given a question and its correct answer find the article that justify it in a given norm.

3. Given a question, find the correct answer and the article that justify it in all norms.
Experiments: previous work

question statement

statement

question items

A

B

C

D

art1

art2

art131-1

...
Experiments: previous work

• legal norms are preprocessed by performing tasks such as converting text to lower case, eliminating punctuation and numbers and, optionally, removing stop-words.

• After that, the articles of the norms are represented as TF-IDF vectors in a Vector Space Model (VSM).

• Every document d is represented by a vector whose size is the vocabulary size of the corpus D. The value of each component t of the vector corresponding to d is given by Equation

$$\text{TFIDF}_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \log \left( \frac{|D|}{|d \in D : t \in d|} \right)$$
Experiments: previous work

• A directed graph with a node for each article of a norm in the corpus.
• When provided a question-answer pair, our system preprocesses the question statement and the alternatives in the same way as it does to the articles in the base graph, turns them into TF-IDF vectors using IDF values from the document corpus.
• The statement node is connected to every article node, and each article node is then connected to every alternative node.
• The edges are given weights whose value is the inverse cosine similarity between the connected nodes’ TF-IDF vectors.
• The system then calculates the shortest path between question statement and answer item using Dijkstra’s algorithm
• and returns the article that connects them as the answer justification.
• our graph structure does not allow for more than one node connecting statement and alternative => our simplification!
Experiment 1 results

• Find the correct answer and its justification in a given norm.
• Although it chose the correct alternative 10 times (out of 30), it only provided the correct justification for 8 of these.
• Numbers are not important, we are trying to create a baseline for further processing.

The young adults Rodrigo (30-year-old), and Bibiana (35-year-old), who are properly enrolled in an OAB section […] Considering the situation described, choose the correct alternative:
A) Only Bibiana meets the eligibility criteria for the roles.
B) Only Rodrigo meets the eligibility criteria for the roles.

[ […] ]
(2016 OAB exam, 19th edition, question 7)

[question statement]
A) does not compel him to pay the agreed upon legal fees.
B) does compel him to pay the agreed upon legal fees.

[ […] ]
(2015 OAB exam, 18th edition, question 1)
Experiments 2 and 3 results

• Experiment 2 (find the justification in a given norm) : the system retrieved the correct article in 21 out of 30 question-answer pairs.

• Experiment 3 (find justification in all 3 norms, all for the same topic, increasing the difficulty) : it scored the right article in 18 out of the 30 question-answer pairs.
Questions Types

• Following


• We start to identify the questions characteristics
Example of Question with a narrative and answers as full statements (google translation)

Paulo, a Bachelor of Laws, held relevant positions in the Executive Power of the three spheres of Government, acquiring in-depth knowledge of the internal activities of Public administration. After retiring, without requiring inscription in the tables of the OAB, establishes legal advice, having raised several clients from the opening of its activity.

According to the narrated and observed the norms statutory, mark the correct statement.

OPTIONS
A) The lawyer's exclusive activities include legal advice and legal advice, but not the consulting services.
B) The retired law bachelor has not prohibited any practice of legal activity, even if not entered in the tables of the OAB.
C) (CORRECT) The lawyer acts in the judicial activity struggling for the defense of the interests of its clients and in the legal
D) The private activities of the lawyer include advice juridical management, and the performance of special judges.
João postulated, through a lawyer's representation, an action conviction in the face of the company Cacos e Cacos Ltda. obtaining a favorable sentence, condemning the defendant to the payment of the amount of R$ 100,000.00, plus R$15,000.00 of attorney's fees. After the final decision of the judicial decision, João and his lawyer Pedro are scientists that society is bankrupt, and their credits must undergo an authorization procedure. In that case, the nature of the claims corresponding to attorneys' fees, under the terms of the considered as

OPTIONS
A) unsecured.
B) real.
C) (CORRECT) privileged.
D) natural.
The following alternatives present some of the powers of the Federal Council of the Portuguese Bar Association of Brazil, with the exception of one. Check it out.

OPTIONS
A) To represent, in or out of court, the collective interests of the lawyers.
B) Ensuring the dignity, independence, prerogatives and valuation of advocacy.
C) (CORRECT) Represent, without exclusivity, Brazilian lawyers in international advocacy bodies and events.
D) To edit and amend the General Regulations, the Code of Ethics and Discipline, and any Proceedings deemed necessary.
Simple text retrieval from Law texts without clue about the norm (google translation)

The power to prosecute and originally judge Governor of State for common crime is

OPTIONS
A) Supreme Federal Court.
B) (CORRECT) Superior Court of Justice.
C) Special Body of the Court of Justice.
D) Criminal Court of the capital where the Court of Justice of the respective State.
In view of the recent popular demonstrations, reported on TV that certain state deputies of State of the Federation were using the money from the budget for health for their own benefit ... In attention to the disciplined in Law n. 4,717 / 65, which deals with the Popular Action, tick the **wrong alternative**.

OPTIONS

A) Marta, a Brazilian citizen, resident and domiciled in the same State, you can qualify as a litisconsorte of Marcos.

B) B: CORRECT) In the same line as the write of mandado de segurança, the right to sue her falls in 5 (five) years.

C) C) The State, in the opinion of its legal representative, in public interest, it could act along side of Marcos in the conduct of the action.

D) D) Being dismissed as unfounded the action brought by Marcos, may appeal, in addition to the Public Ministry and any other citizen.
The possibilities for WSD/WN

• Constitutive acts and contracts of legal persons, in order to be registered regarding the legal practice statute, must: [. . . ]
  • C) contain the lawyer’s [. . . ] signature (*visto*).

(17th ed. OAB exam, question 2)

• §2 The constitutive acts and contracts of legal persons can only be registered in the competent bodies, under a penalty of invalidity, when signed (*visar*) by lawyers.

( law no. 8906, article 1)

[00996485-v, sign, subscribe: mark with one’s signature] and [06404582-n,sig-nature: your name written in your own handwriting] and the morphosemantic link “result”.
UkB for WSD and OpenWordet-PT

• Freeling is an open source language processing library developed at the TALP research center.

• Freeling is a pipeline based library: tokenization, lemmatization (dict based, needed fixes), POS tagging etc. The WSD module is a UKB implementation.

• UKB applies random walks, e.g. Personalized PageRank, on the Knowledge Base (KB) graph to rank the vertices according to the given context. Graph-based WSD methods manage to exploit the interrelations among the senses in the given context. In this sense, they provide a principled solution to the exponential explosion problem, with excellent performance.

• For Portuguese, FL relies on OWN-PT, an open freely available wordnet for Portuguese.
Coverage of OpenWordnet-PT

• We did a survey on the coverage of OWN-PT for the OAB corpus
• The PWN synset [08441203-n,law/jurisprudence: the collection of rules imposed by authority.] is a general concept about law, and is linked to hundreds of synsets via the classifiesByTopic relation. Are them properly translated in OWN-PT? Legal jargon can be language and cultural dependant.
• We started from the most most common words.
OpenWordnet-PT changes

• New synsets: “cartório” (notary office)
• Missing words as in [06532763-n,*nulidade*: nullity].
• Other cases were those of relations that did not exist in OWN-PT; if present, these relations would improve the results of the UKB algorithm.
• The nominalization (morphosemantic link) between [00664276-v,comprovar: authenticate] and [06855035-n,comprovação: authentication].
• In the end, since we focused only on the possible improvements to our immediate purpose, we have added to OWN-PT 2 synsets, 8 semantic and lexical relations, and 25 lexical units.
• UKB mistakes are at least consistent: [06532095-n,ato: legal act] was assigned to the [00037396-n,act: as in action]
Experiments adaptations

• The WSD modules to assign OWN-PT synsets, with a weight value (normalized in order to sum 1), to each token or sequence of tokens.

• For an input text we have a list of key-value pairs (s,w) with a sense key and a weight value, in contrast to a simple list of tokens, as we had in the previous experiment.
Experiments adaptations

• $f_{s,w}$ is the sum of each occurrence of sense $s$ weighted by $w$.

• TF: we count the weighted occurrence as a “continuous occurrence”, instead of boolean, where the degree of occurrence is the weight of the sense.

• IDF: if the sum in a document is higher than 1, it counts as an occurrence, which is counted only once. Otherwise, it counts only according to the weight received.

$$\text{TFIDF}_{s,w,d} = \text{TF}_{s,w,d} \text{IDF}_{s,w,D}$$

$$\text{TF}_{s,w,d} = \frac{f_{s,w,d}}{\sum_{s' \in d} f_{s',w',d}}$$

$$\text{IDF}_{s,w,D} = \log \left( \frac{|D|}{\sum_{d \in D} w^{I(w<1) I(s \in d)}} \right)$$
Results

- 30 questions/answers and justifications
- Comparing previous results with WSD based results
- (QA): choosing the right answer at the multiple choice problem, given the questions and the laws (all three normative documents related to the legal ethics area).
- not only correct answer, but a correct justification as well, experiment (QA+J)
- (J) the system’s task was to determine which article (considered every law it has seen) justified the (already given) answer.

<table>
<thead>
<tr>
<th></th>
<th>QA</th>
<th>QA+J</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>word system</td>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>synset system</td>
<td>14</td>
<td>11</td>
<td>17</td>
</tr>
</tbody>
</table>
Cases

• Concerning the expiration of punitive disciplinary infractions, choose the right alternative. [ . . . ] A) The punitive aim in regard to disciplinary infractions expires after five years [ . . . ] B) The punitive aim in regard to disciplinary infractions expires after three years [ . . . ]

(15th ed. OAB exam, question 4)

Fail to find the right answer, shorted-paths are almost the same. WSD of the surround words would not be affected, both hyponomous of digits.
Cases

- [question statement] [...] A) does not compel him to pay the agreed upon legal fees. [...] B) does compel him to pay the agreed upon legal fees. [. . . ]

(18th ed. OAB exam, question 1)

No improvement with WSD! No bag-of-words model will suffice.
conclusions

• We presented few simples experiments aimed to explore the data about OAB Exams.
• We have already learned some patterns of questions types
• Further work on “deep” natural language processing (semantic representations)
• TF-IDF does not solve important difficulties, such as compositional understanding, pragmatics, etc. Nevertheless, the contributions to OWN-PT can be seen as a benefit by itself and will be valuable in the future planned experiments.
• Further work with LexML for represent other legal documents, increase the corpus.
• Logic approach and the translation from syntax to predicate/arguments to semantic representation.
Obrigado! Thanks!


arademaker@gmail.com
alexrad@br.ibm.com