Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-119

Scoring Judicial Syllabi in Portuguese

Jean-Rémi BOURGUET [a,1] and Melissa ZORZANELLI COSTA [b,2]
[a] *Núcleo de Estudos em Modelagem Conceitual e Ontologias*
Federal University of Espírito Santo (UFES) – Brazil
[b] *Tribunal Regional Federal da 2ª Região*
Justiça Federal - Seção Judiciária do Espírito Santo (JFES) – Brazil

Abstract. Law professionals generally need to investigate a large number of items to make their decisions. However, the frameworks they use are often limited to a simple full-text search. In this paper, we propose to score the results of such searches investigating ontological and non-ontological solutions. We examine their applicabilities in a real use case dealing with jurisprudences of regional federal courts in Brazil.

Keywords. Jurisprudences, Full-text search, NLP, Portuguese, Similarities

1. Introduction

Nowadays, more and more Application Programming Interfaces (APIs) supporting a Natural Language Processing (NLP) are released and their mutual usages in common infrastructures can considerably enhance the results of full-text searches. As they are regularly confronted with large numbers of judged cases stored in relational data management systems, professionals of Brazilian courts need innovations to refine these results. Indeed, if the usual way to output a full-text search is an ordered list of items, few approaches are thought to display the results in different ways. Actually, the jurisprudences that are judicial decisions taken by a specific court in Brazil (e.g. regional federal tribunal), are stored in semi-structured formats in which a large part of the relevant knowledge are present in syllabi, i.e. textual explanations in Portuguese. Our proposal is then to score the results of a full-text search among judicial syllabi supported either by ontological or non-ontological solutions. On the one hand, we took up the challenge to perform some automatic translations of the syllabi into English (with Google Cloud Translation) before computing similarities from the Princeton WordNet [1]. On the other hand, we opted to proceed word embeddings of the Brazilian penal code using Word2vec for Lucene [2] before computing similarities between lemmas by cosine measures. The remainder of this paper is organized as follows: Section 2 describes the ontological and non-ontological solu-

[1]Corresponding Author: Jean-Rémi Bourguet (jean-remi.bourguet@ufes.br) is supported by the Brazilian Research Funding Agency FAPES (grant 71047522).
[2]Melissa Zorzanelli Costa (mzcosta@jfes.jus.br) would like to acknowledge the assistance of the Tribunal Regional of 2nd Region of Brazil.

tions to score such results and evaluates the applicability of our approach through a real use case and an interface, Section 3 mentions the most similar approaches and Section 4 concludes and opens some research perspectives.

2. Scoring the results

We introduce in Definition 1 a possible global score between two texts as the maximal similarity score between their components. We denote $\overline{w_k}$ the lemmatization of a word w_k. We arbitrarily avoided taking into account the score with words present in a stop words set denoted θ.

Definition 1. *Let two texts p, q:*

$$SIM\,(p,q) = \max_{p_i \in p \setminus \theta, q_j \in q \setminus \theta} sim(\overline{p_i}, \overline{q_j})$$

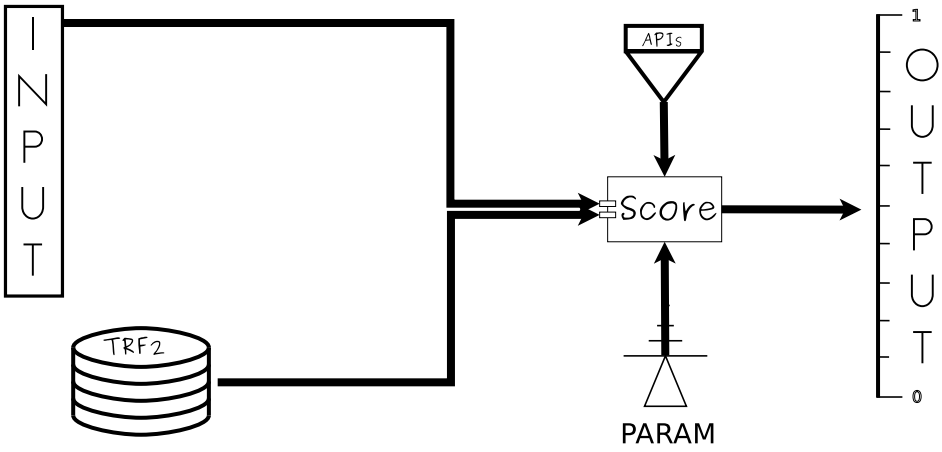

Figure 1. System description

Looking at the system description of Figure 1, several modules supports the computation of the similarity scores: `INPUT` takes a text in Portuguese, `TRF2` is a data set of jurisprudences, `APIs` is a set of NLP-based APIs supporting the computations of the similarity scores, `PARAM` manages the choice of the local measures, the stop words list, etc., `SCORE` orchestrates the computation of the similarity scores and `OUTPUT` is devoted to display them.We transformed the XML file provided by a regional Brazilian federal tribunal in an RDF file using an XSLT transformation. After that, we stored this knowledge base in a triple store supported by the Virtuoso's infrastructure [3]. We used the Jena API to query Virtuoso possibly from a Java interface developed with the native Swing API.

Performing similarities in a language other than English can be a challenge. Then, we will present both ontological (in Section 2.1) and non-ontological (in Section 2.2) approaches using semantic similarities or word embeddings to score the results of a full-text search.

2.1. Using semantic similarities

Semantic similarities generally use distances between semantic units (called synsets) in a ontologically founded lexicon. According to Sartor et al. [4], a synset can be defined as a set of one or more uninflected word forms (or lemmas) called word-senses belonging to the same part of speech (denoted pos) e.g. noun, verb or adjective. Each synset is encoded with its *lemma*, *pos* and a number *nb* discriminating it among some other possible senses (lemma#pos#nb). Finally, the relation of hyponymy (denoted $<^h$) is a binary relation that partially orders the synsets. Currently, the Global WordNet Association indexes a large set of open-source adaptations of lexicons in approximatively 40 languages. Concerning the Portuguese language, two repositories are referenced: ONTO.PT [5] a Portuguese repository automatically built from heterogeneous textual sources, and OpenWN-PT [6] a Brazilian Portuguese repository built using Wikipedia and alignments with the Princeton WordNet (conserving its abstract relations), manual revisions and gloss translations. Unfortunately, no API exists in the market to equip ONTO.PT or OpenWN-PT with a support for the computation of semantic similarities. Therefore, we opted for an automatic translation of each syllabus (and the input) in English using Google Cloud Translation. After that, we were able to perform semantic similarities with WordNet:Similarity (developed by Pedersen et al. in [7] and redesigned in Java by Shima [8]). One of the most common measures is the path-length relatedness founded on a node-counting scheme concerning the smallest specified role counting between two synsets (see [7]). The path-length based relatedness score (`plr`) is equal to the inverse of the shortest path length between two synsets. Nevertheless, other historical similarity measures considering the maximum depth in taxonomy (`lch`), the depth of the least common subsumer (`wup`) or the supported information content (`jcn` and `lin`) can also be computed with WordNet:Similarity.

2.2. Using word embeddings

Word embeddings is a well-known approach based on deep learning (using a neural natural language model) recently retailored by Mykolov et al. [9]. These similarities arise from a large set of word vectors produced after a learning step performed on a textual corpus ideally in a particular context of interest. This corpus may receive a pretreatment (e.g. tokenization, lemmatization, stop word filtering) in order to decrease the noise of nonsense textual information. Thus, the learning step encodes the general context of words in dense vectors. The similarity between two lemmas is obtained through the cosine of their vectors. We opted to use the Brazilian penal code[3] to perform the learning step. We used an API called Word2vec for Lucene [2] and proceeded to a lemmatization of the text (using Lematizador [10]). After that, we filtered the corpus with stop words, stop signs and numbers giving a train file of 11593 lemmas. We performed word embeddings using a size 200 for the vectors, a window (max skip length between words) of 5, discarding words that appear less than 5 times. We finally obtained a vocabulary size of 589 lemmas.

[3] http://www.planalto.gov.br/ccivil_03/decreto-lei/Del2848compilado.htm

2.3. Evaluation

We performed an evaluation of our framework looking for the word *banco* (i.e. bank in English) to quickly browse the large repository of jurisprudences by outputting an affordable set. Afterwards, we scored the answer by using semantic similarities (PLR) or by using word embeddings (EMB) with the word *dano* (i.e. damage in English). In the Table 1, we related the 4 most similar jurisprudences with the 3 highest local similarities scores obtained by semantic similarities and word embeddings. The first column describes a part of the Syllabi with the most similar words written in bold while the other columns show for each approach the most similar words, their scores and their translations between parenthesis.

Syllabus	PLR	EMB
DIREITO CIVIL. ATIVIDADE BANCÁRIA [...] Tal responsabilidade somente fica descaracterizada na **ocorrência** de uma das hipóteses do § 3º do referido art. 14, o que não **ocorreu** na espécie. 2 - O princípio da reparabilidade do **dano** moral foi expressamente reconhecido [...] 5 - Em face da responsabilidade **civil** contratual, aplicável a inversão do ônus da prova prevista no artigo 6º [...] para elidir sua responsabilidade civil, comprovar que o **fato** derivou da culpa do cliente ou da força maior ou caso fortuito [...]	1.0 dano (damage) 0.25 ocorrência (event) 0.25 ocorreu (occur)	1.0 dano (damage) 0.33 lei (law) 0.22 fato (fact)
PROCESSO CIVIL. BANCO CENTRAL DO BRASIL [...] **Lei** 8.112/90 à hipótese em tela pois [...] o **crédito** em discussão decorre da relação trabalhista que outrora existia entre os litigantes [...] podendo ocasionar sérios embaraços ao orçamento do agravante, a caracterizar o fundado receio de **dano**. 9. Não há violação ao **direito** constitucionalmente assegurado de acesso ao Judiciário, eis que o recorrente poderá ajuizar ação ordinária, sendo-lhe vedado tão-somente **constituir** Certidão da Dívida Ativa [...]	1.0 dano (damage) 0.33 direito (right) 0.25 constituir (constitute)	1.0 dano (damage) 0.33 lei (law) 0.28 crédito (credit)
ADMINISTRATIVO. CADERNETA DE POUPANÇA. CORREÇÃO MONETÁRIA [...] I- A competência da Justiça Federal in ratione personae encontra-se disposta no art. 109, inciso I, da **Lei** Fundamental. [...] Reverência ao princípio constitucional da irretroatividade da **lei** para **prejudicar** o **direito** adquirido e ato jurídico perfeito [...] em respeito ao **direito** adquirido e ao ato jurídico perfeito, não calhando a alegação de negativa de vigência do art. 17 da **Lei** no. 7.730/89. [...]	0.5 prejudiciar (impair) 0.33 direito (right) 0.33 direito (right)	0.33 lei (law) 0.33 lei (law) 0.33 lei (law)
PROCESSO CIVIL - CRUZADOS BLOQUEADOS [...] POR FORÇA DA **LEI** 8024/90 [...] **RESPONDEREM** PELA CORREÇÃO MONETÁRIA [...] AS QUAIS FORAM PRIVADAS DA DISPONIBILIDADE DO **DINHEIRO** [...] NORMA POSTERIOR QUE ALTERE O ÍNDICE DE **CORREÇÃO INCIDENTE** SOBRE TAL MODALIDADE DE INVESTIMENTO [...] RECURSO DO BANCO CENTRAL IMPROVIDO E RECURSO DO BANCO DO BRASIL **PARCIALMENTE** PROVIDO.	0.5 correção (change) 0.25 incidente (incident) 0.2 responderem (respond)	0.33 lei (law) 0.26 dinheiro (money) 0.23 parcialmente (partially)

Table 1. Evaluation of the approaches on a real user case

We remarked on two important limits: i- concerning semantic similarities, since lemma maps to one or more word senses, the original translated noun *damage* is considered as a verb scoring a similarity of 0.5 due to the order: change#v#1$<^h$damage#v#1$<^h$impair#v#1; ii- concerning word embeddings the word *lei* (law) is systematically recognized as the most similar after *dano* due to the relatively smaller set of vectors obtained from the Brazilian penal code.

2.4. Interface

An illustration of the interface for the application of the framework `LooPings` [11] to display the semantic similarities is presented in Figure 2 for the `PLR` measure. The scores are finally placed on a segment $[0, 1]$. Note that because sets of answers can have the same scores (cases of ex æquo), the interface randomly choose one ID to display beside groups of points with the same scores.

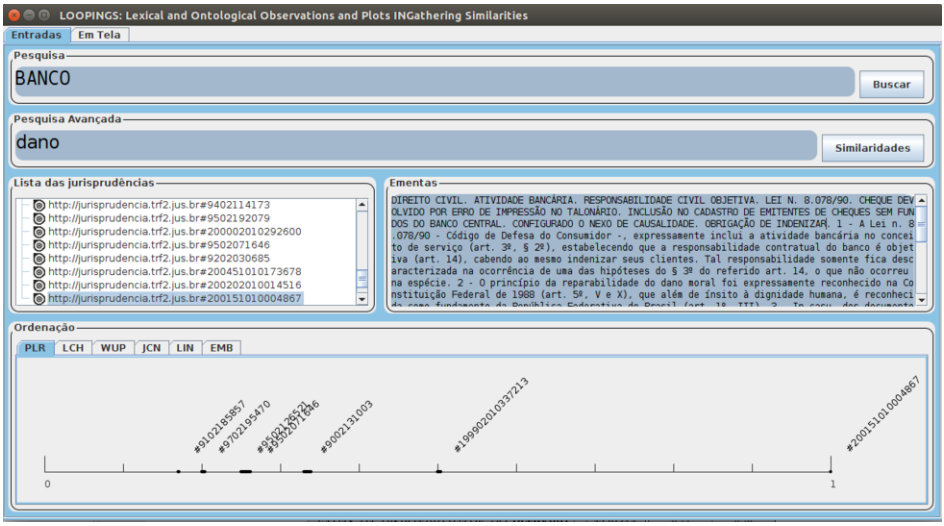

Figure 2. An interface showing the application of `LooPings`

3. Related works

The issue of ordering legal documents has already been investigated by Lu and Conrad [12] who proposed an issue-based content recommendation system with a built-in topic detection/segmentation algorithm for the legal domain.

The usage of conceptual layers to support searches among jurisprudences is a relative new trend of research in Brazil (see [13] for example). The syntactic similarities can also be used to assist such tasks because their computations can be transposable for the Portuguese language (see [14] for example), but very few works dealt directly with semantic similarities using a Portuguese material. One remarkable work was carried out by Aleixo and Pardo [15] in which node-length path similarities (after lemmatization and stop list treatments) are performed using a Brazilian Portuguese thesaurus in order to compute relatedness between sentences. In keeping with this trend, Baldez de Freitas et al. [16] proposed a measure extending the path length based relatedness in order to compute similarities between terms of distinct ontologies.

The usage of word embeddings to browse legal items is also relatively new. Landthaler et al. [17] recently explored a method that provided semantically similar answers for arbitrary length search queries using word embeddings.

4. Conclusion

In this paper, we described both ontological and non ontological approaches to perform similarities among judicial syllabi from a set of jurisprudences of a regional federal court in Brazil. We also proposed an interface to display the results of a full-text search. We now intend to propose an approach to browse the jurisprudences integrating NLP-based APIs and thesaurus in Portuguese.

References

[1] Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.
[2] Koji Sekiguchi. word2vec for Lucene, 2016. http://goo.gl/dgTxiz.
[3] Orri Erling and Ivan Mikhailov. RDF support in the virtuoso DBMS. In *Networked Knowledge-Networked Media*, pages 7–24. Springer, 2009.
[4] Giovanni Sartor, Pompeu Casanovas, Mariangela Biasiotti, and Meritxell Fernández-Barrera. Approaches to legal ontologies: Theories, domains, methodologies. law. *Governance and Technology series. Springer*, 2011.
[5] Hugo Gonçalo Oliveira. The creation of Onto.PT: a wordnet-like lexical ontology for Portuguese. In *Proceedings of 11th International Conference of Computational Processing of the Portuguese Language*, volume 8775 of *LNCS*, pages 161–169. Springer, 2014.
[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, 2012. The COLING 2012 Organizing Committee.
[7] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet: Similarity - measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 1024–1025. AAAI Press, 2004.
[8] Hideki Shima. Wordnet Similarity For Java (WS4J), 2015. http://goo.gl/FTAU52.
[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[10] Erick Galani Maziero. *Análise retórica com base em grande quantidade de dados*. PhD thesis, Universidade de São Paulo, 2016.
[11] Adama Sow and Jean-Rémi Bourguet. LooPings: A look at semantic similarities. In Delgado Y.H. Leiva Mederos A.A., editor, *Proceedings of the 2nd International Workshop on Semantic Web*, volume 1797 of *CEUR Workshop Proceedings*, pages 23–32, 2016.
[12] Qiang Lu and Jack G. Conrad. Bringing order to legal documents - an issue-based recommendation system via cluster association. In Joaquim Filipe and Jan L. G. Dietz, editors, *Proceedings of KEOD 2012*, pages 76–88. SciTePress, 2012.
[13] Rafael Brito de Oliveira and Renata Wassermann. Utilização de ontologia para busca em base de dados de acórdãos do STF. In Mara Abel, Sandro Rama Fiorini, and Christiano Pessanha, editors, *Proceedings of the IX Seminar on Ontology Research in Brazil*, volume 1908 of *CEUR Workshop Proceedings*, pages 147–157. CEUR-WS.org, 2017.
[14] Eloize Rossi Marques Seno and Maria das Graças Volpe Nunes. Some experiments on clustering similar sentences of texts in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 133–142. Springer, 2008.
[15] Priscila Aleixo and Thiago Alexandre Salgueiro Pardo. Finding related sentences in multiple documents for multidocument discourse parsing of brazilian portuguese texts. In *Proceedings of WebMedia08*, pages 298–303. ACM, 2008.
[16] Juliano Baldez de Freitas, Vera Lúcia Strube de Lima, and Josiane Fontoura dos Anjos Brandolt. Semantic similarity, ontologies and the portuguese language: A close look at the subject. In *Proceedings of PROPOR08*, pages 61–70. Springer, 2008.
[17] Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. Extending full text search for legal document collections using word embeddings. In Floris Bex and Serena Villata, editors, *Legal Knowledge and Information Systems - JURIX 2016*, volume 294 of *Frontiers in Artificial Intelligence and Applications*, pages 73–82. IOS Press, 2016.