An NLP Approach for Detecting a Link Between Text Coherence and the Speed of a Bill Promulgation

1

Abstract. *We present in this paper an analysis of semantic distances in bill documents aiming to generate features for speed of law approval prediction. The Word Mover's Distance (WMD) was used as a distance measure between sentences in a bill document. From these distances, descriptive statistics were obtained to be used as features in a xgboost prediction model that predicted how long it would take for a bill to read promulgation.*

1. Introduction

The legislative system in Brazil is, albeit transparent in accordance to the law, very keen in obfuscating changes in Law Projects due to Lobbying or attempts to make the project be approved faster. Through the use of NLP we hope to produce features that can describe the legislative process and its link to the speed of the Bill's approval.

By using the Word Movers Distance (WMD) proposed [Kusner et al. 2015] we hope to measure semantic dissimilarity between sentences in a bill document. After generating the distances we use descriptive statistics features such as the average distance, variance, standard deviation, sum of distances and squared mean distance for each document as input for a prediction model.

The WMD algorithm does not use hyper-parameters as input, however a Skipgram model must be provided. The Skipgram model was introduced by [Mikolov et al. 2013] and uses a neural network to learn vector representation for words. Most freely available Skipgram models are trained using news articles, papers or product reviews, since our problem concerns texts with a specific vocabulary we decided to train our own Skipgram model on law documents and bills.

The resulting features were used as inputs for a XGBoost model which is a tree boosting model proposed by [Chen and Guestrin 2016]. It is widely used among Data Scientists at machine learning competitions involving store sales prediction, motion detection and malware classification.

1.1. Legislative process and its problems

The legislative process in Brazil is bicameral, which means a bill must be approved by the Senate and a house of representatives. The process under which a bill is submitted to be approved is, although well described, not simple. After going through many steps the newly approved law is posted on the Union's Journal.

The problem lies within how long a bill takes to be approved, where some take years and some months due to internal proceddings in each house. As the projects go through each house they may be altered during discussions and the reasoning behind those changes are not clear on the publicly available documents.

2. Methodology

As seen in [Kusner et al. 2015] most freely available Skipgram models are trained using news articles, academic papers or product reviews. Since our problem concerns an area with a specific vocabulary we decided to train our own Skipgram model on law documents and bills.

These texts were collected using the câmara dos deputados API where 6.634 text files were obtained. These text files had stopwords, uppercase and punctuation removed and were dumped on a single text file that was used as input for the Skipgram Algorithm implementation provided by Mikolov et al.

Brazilian law texts are structured as articles, paragraphs, subsections and itens. This tree like structure can be read as a normal sentence when flattened. By flattening the structure with a parser we were able to create sentences that can be easily compared with the WMD. These sentences had uppercase, punctuation and stopwords removed. For example:

1st. This law deals with:

1 - The troubles of dealing with excruciatingly:

I. Long contexts

a. in the case of LSTM architecture.
b. in the case of CNN architecture.

II. Long sentences

a. in the case of word2vec architecture.
b. in the case of wmd comparisons.

Has it's context better seen as:

(1) This law deals with: The troubles of dealing with excruciatingly: Long contexts in the case of LSTM architecture.
(2) This law deals with: The troubles of dealing with excruciatingly: Long contexts in the case of CNN architecture.
(3) This law deals with: The troubles of dealing with excruciatingly: Long sentences in the case of word2vec architecture.
(4) This law deals with: The troubles of dealing with excruciatingly: Long sentences in the case of WMD comparisons.

Figure 1. Proposed flattened text

In the next step the WMD algorithm was used to generate semantic distances between each line inside a Legislation Project with flattened text. We saved the value and position of each distance, in order to generate the following features: average distance, variance and standard deviation, sum of all distances, and squared mean distance.

Since our goal is to predict how long it would take for a project to reach the promulgation phase we extracted the project's publication date from the Original Project File, and the date of the Publication of the Law from the final PDF file. To test the prediction model we subclassified legislations in regards to the number of political mandates it took for it's approval. The political mandate is of 4 years in Brazil. In that way every class was defined as:

$$class\_doc = INT(num\_years\_approval/4) \tag{1}$$

3. Coherence Feature Examples

In order to provide examples of the effectiveness of the feature we selected some sentences from the greatest and shortest distances according to the comparison method described above. The original sentences are translated in [A]

3.1. Short Distances

3.1.1. Correct Assessment

- ”The resources of FUNCAP will be maintained in a federal financial institution and managed by a Board of Directors, composed of: three representatives of the Union;” [1]
- ”The resources of FUNCAP will be maintained in a federal financial institution and managed by a Board of Directors, composed of: a representative of the Municipalities.” [2]

The model was correctly able to assess that the sentences above are correlated. As we can see, the only differences are a plural, a number and a change of public entity.

3.1.2. Incorrect Assessment

- ”The Federal Executive Branch shall support, in a complementary manner, the States, the Federal District and the Municipalities in an emergency situation or state of public calamity, through the mechanisms provided for in this Provisional Measure.” [3]
- ”The states, the Federal District and the municipalities must provide accounts of the funds drawn, in the form of the regulation.” [4]

Although there is a continuity context, our model is not supposed to capture temporal correlations, so the assessment of a low distance is incorrect given the characteristics of the model. A person, however, should be able to realize that those two sentences are correlated.

3.2. Long Distances

3.2.1. Correct Assessment

- ”The Union may anticipate quotas, in order to promote the adhesion of the different federated entities in FUNCAP.” [5]
- ”It is the Executive Branch authorized to donate public food stocks, in natura or after processing, directly to poor populations, aiming at the fight against hunger and misery, as well as to populations affected by disasters, when characterized by emergency situations or state of public calamity, through a joint proposal of the Ministry of Agriculture, Livestock and Supply, Ministry of National Integration and Civil House of the Presidency of the Republic.” [6]

The model was correctly able to assess that the sentences above are not correlated in the slightest, which is easily perceivable.

3.2.2. Incorrect Assessment

- ”It is the Public Administration, direct, indirect and foundational, authorized to institute a program that guarantees the extension of the maternity leave for its servants, in accordance with the provisions of art. 1.” [7]

- ”The extension shall be guaranteed, to the same extent, also to the employee who adopts or obtains judicial custody for the purpose of adopting a child.” [8]

Although the general subject of the two sentences is related, the involved entities are different on their nature, one is public and the other one is private. Due to that, our model should have been able to correlate the similarities of public and private agents and, therefore, keeping the distance low, however the measured distance was high. The level of subjectiveness for this assessment is very high and requires a lot of previous knowledge. This was the best example found of an incorrect assessment of great distances. This fact indicates that the model is a good incoherence detector, even though not being the best coherence detector.

4. Results

Using a very simple, non optmized, xgboost model on top of those distances, we have been able to achieve a 47% prediction accuracy to the classes as defined in 1. We had a dataset comprised of the average internal distance for 2513 Law Projects as our variable and the classes as our target.

It must be said that the data was not balanced, so we had nine unbalanced classes. If we would predict the target time with a random approach we should see a prediction of under 40% accuracy which was the ammount of objects on the first class alone. Our model was, however able to correctly assess the classes and keep the proportion of predictions aligned with the true values. Which means that somehow, even though the feature is obviously not enough to predict the time it will take for a Law Project to be approved, we were able to generate with relative ease a feature that is descriptive of the legislative process and that can be applied to many other legal analysis.

In summary, if we were to actually predict the speed for a bill promulgation, we would have to take into account many factors of economical and political order, such as the relevance of the project to the current economical situation of the country, the relevance of the political actors that are backing the project, and many other rather difficult things. Meanwhile this feature alone, by just analyzing the general cohesion of the Law Project was able to perform a little bit better than a random approach in predicting the speed of a Bill Promulgation. Therefore we hope this feature might be used in conjunction of others to describe and aid in machine learning techniques to describe the legislative process.

References

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 785–794. https://doi.org/10.1145/2939672.2939785*.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). *From word embeddings to document distances.* International Conference on Machine Learning (pp. 957-966).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Retrieved from http://arxiv.org/abs/1310.4546*. ArXiv:1310.4546.

A. Original Sentences

1. ”Os recursos do FUNCAP serão mantidos em instituição financeira federal e geridos por um Conselho Diretor, composto por: três representantes da União;”
2. ”Os recursos do FUNCAP serão mantidos em instituição financeira federal e geridos por um Conselho Diretor, composto por: um representante dos Municípios.”

3. ”O Poder Executivo federal apoiará, de forma complementar, os Estados, o Distrito Federal e os Municípios em situação de emergência ou estado de calamidade pública, por meio dos mecanismos previstos nesta Medida Provisória.”
4. ”Os Estados, o Distrito Federal e os Municípios cotistas deverão prestar contas dos recursos sacados, na forma do regulamento.”
5. ”A União poderá antecipar cotas, de forma a fomentar a adesão dos demais entes federados no FUNCAP.”
6. ”É o Poder Executivo autorizado a doar estoques públicos de alimentos, in natura ou após beneficiamento, diretamente às populações carentes, objetivando o combate a fome e a miséria bem como, às populações atingidas por desastres, quando caracterizadas situações de emergência ou estado de calamidade pública, mediante proposta conjunta do Ministério da Agricultura, Pecuária e Abastecimento, do Ministério da Integração Nacional e da Casa Civil da República.”
7. ”É a Administração Pública, direta, indireta e também fundacional, autorizada a instituir programa que garanta prorrogação da licença-maternidade para suas servidoras, nos termos do que prevê o art. 1o.”
8. ”A prorrogação será garantida, na mesma proporção, também à empregada que adotar ou obtiver guarda judicial para fins de adoção de criança.”