Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts

This is the submitted version (pre peer-review, preprint) of the following publication:

Published Version:

Availability:
This version is available at: https://hdl.handle.net/11585/905768 since: 2023-05-08

Published:
DOI: http://doi.org/

Terms of use:
Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)
Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts

Sezen Perçin\textsuperscript{1,2,3} and Andrea Galassi\textsuperscript{1,3}\textsuperscript{*} and Francesca Lagioia\textsuperscript{4} and Federico Ruggeri\textsuperscript{3} and Piera Santin\textsuperscript{4} and Giovanni Sartor\textsuperscript{4} and Paolo Torroni\textsuperscript{3}\textsuperscript{*}

\textsuperscript{1}Department of Electrical and Electronics Engineering, Boğaziçi University, Turkey
\textsuperscript{2}Technische Universität München, Munich, Germany
\textsuperscript{3}DISI, University of Bologna, Bologna, Italy
\textsuperscript{4}CIRSFID-Alma AI, University of Bologna, Bologna, Italy

\textsuperscript{*}a.galassi@unibo.it

Abstract

Creating balanced labeled textual corpora for complex tasks, like legal analysis, is a challenging and expensive process that often requires the collaboration of domain experts. To address this problem, we propose a data augmentation method based on the combination of GloVe word embeddings and the WordNet ontology. We present an example of application in the legal domain, specifically on decisions of the Court of Justice of the European Union. Our evaluation with human experts confirms that our method is more robust than the alternatives.

1 Introduction

Many of the state-of-the-art Natural Language Processing (NLP) techniques are based on deep learning methods with millions of parameters (Devlin et al., 2019; Vaswani et al., 2017), and therefore they usually require vast amounts of data to be trained. Even if a lot of progress has been made in the development of unsupervised or semi-supervised methods, many high-level tasks are still addressed in a supervised fashion, especially when they concern complex tasks or very specific domains, such as predictions on legal documents (Drawzeski et al., 2021; Poudyal et al., 2020; Zhong et al., 2020). At the same time, creating corpora for such applications is particularly challenging and expensive since this process requires the collaboration of domain experts for the labeling process. One possible way to address this problem is data augmentation (Shorten et al., 2021), which exploits existing data to generate new synthetic ones. These synthetic samples must be different enough from the original ones to provide a valuable contribution to the training. Still, at the same time, their semantic content must remain similar enough not to invalidate their labels. In NLP, one possibility is to replace some words or sentences of the original samples with other ones that hold the same semantic meaning. This can be done by exploiting similarities between sub-symbolic representations of text, such as word and sentence embeddings, or exploiting relationships in symbolic representations, such as WordNet (Fellbaum, 2010).

Inspired by works regarding semantic relatedness (Lee et al., 2016; Vasanthakumar and Bond, 2018), we propose to merge graph-structured and embedding-based augmentation by combining the use of WordNet and similarity between word embeddings. In particular, we create new synthetic samples by replacing some terms with words with similar semantic meaning. We exploit WordNet to compute a set of candidate words and then choose the most similar one according to its GloVe word embedding (Pennington et al., 2014).

We present an example of the application of such a method in the legal domain. Our context is a task of sentence classification, where we want to automatically predict whether a sentence extracted from a judgment is representative of a principle of law. Since the distribution between the negative and positive classes is heavily unbalanced, we need to rely on data augmentation. We compare different techniques and ask a team of legal experts to evaluate the new synthetic data. Their evaluation confirms that the quality of the synthetic data generated through our method is superior to data generated exploiting only WordNet or GloVe embeddings. Our contribution is three-fold:

- (i) we propose a novel method to perform textual data augmentation by mixing the use of WordNet and Word Embeddings;
- (ii) we perform a qualitative evaluation on legal documents, where human domain experts assess the efficacy of our method with respect to alternatives;
• (iii) we perform a preliminary quantitative evaluation, using neural language models to measure the similarity between the augmented texts and the original ones.

We make our code, data, and evaluation publicly available.¹

2 Related Works

Data augmentation is a frequently used strategy in NLP to introduce diversity in the datasets that will help models overcome phenomena such as overfitting (Shorten et al., 2021). In particular, paraphrasing-based data augmentation techniques (Li et al., 2022) aim to create new synthetic data preserving the meaning of the original source.

One popular family of augmentation methods relies on knowledge graphs, thesauruses, and lexical database such as WordNet. WordNet (Fellbaum, 2010) is a lexical database where words are grouped into sets of cognitive synonyms called "synsets". Serving as a relational network, it is widely used as a source of synonyms and for the measurement of similarity between terms. For example, Mosolova et al. (2018) use WordNet to retrieve a list of synonyms of a word, and replace it with one chosen randomly. Xiang et al. (2020) expand such approach by constraining candidates according to Part of Speech (POS) tags by selecting them based on a similarity measure, and test their approach on various text classification tasks. Wang and Yang (2015) follow a different approach and instead they rely on semantic embeddings, embedding words with Word2Vec and replacing candidate words with their nearest neighbour.

Our approach stems from Xiang et al.’s and follows the intuition of Wang and Yang. We rely on WordNet to select a pool of candidate words, but we choose the replacement by measuring the similarity between their GloVe word embeddings (Pennington et al., 2014). However, we provide a simpler definition of the candidate list considering the synsets collected from the WordNet opening room for syntactic differences while preserving the semantic integrity of the sentences. Moreover, we address the challenging domain of legal documents, in which retaining domain-specific validity while introducing textual diversity is a critical factor. Finally, we provide an evaluation of synthetic samples involving human experts.

³We included RP words since they can be used as adverbial particles.

³For example, the WordNet POS tag n correspond to the NLTK POS tags NN, NNS, NNP, NNPS.
whose $POS_{w_k}$ correspond to $POS_{w_j}$ are considered; in this way we enforce two POS constraints: one on the synsets level, and one on the lemmas level;

- **augGV** does not rely on WordNet, but only on the vector space properties of the pre-trained GloVe word embeddings, replacing the original word with the most similar one among those present in the vocabulary.

- **augLB** is a neural augmentation method (Shorten et al., 2021) based on Legal-BERT language model (Chalkidis et al., 2020): firstly the candidate word is replaced with a mask token, then the sentence is inputted to the neural language model, and finally, the model generates a novel word in place of the mask token.

4 Evaluation

To perform a preliminary evaluation of our method, we generated a small set of synthetic samples and then asked domain experts to judge them. We also measure the difference between the augmented sentences and the original ones in terms of similarity between their embeddings.

We generated the synthetic starting from a given textual sentence, randomly selecting one suitable candidate word in it, and applying one augmentation method to it. The original sample and the synthetic one thus obtained would therefore differ only for one term. This process was then applied multiple times to the synthetic sample, replacing other words and generating new samples. We repeated this process until we replaced about 60% of the candidate terms of the original sentence.

4.1 Data

We conducted our experimentation on segments of texts in English language extracted from decisions of the Court of Justice of the European Union (CJEU) on fiscal state aid. In particular, we have chosen sentences that are representative of a principle of law (legal maxims or rationes decidendi).

Such sentences are used to highlight the decisive principle of law contained in each judgement, that will be useful to assure the uniform interpretation of the law with respect to the courts of first or second instance. Out of the 334 segments extracted by domain experts from 41 documents, we randomly selected 10 of them. We have chosen to work with CJEU decisions because they usually contain a rich and diverse set of legal principles established in a case that determine the judgment.

4.2 Metrics

For the human evaluation, two domain experts have analyzed each single augmentation step, assigning a value between $\{+1, 0, -1\}$. In particular, we proceeded as follows: for each sentence we assigned a score to each replaced word, then we sum the single values to obtain an overall score for the sentence at stake. We have chosen to use a 3-values scale to identify replacements that are:

- completely correct (+1), i.e., words or expressions having the same meaning in the considered legal contest;

- imprecise for our specific domain (0) , i.e., words or expressions having either a similar or nearly the same meaning but pertaining to an informal or nonlegal linguistic register;

- completely incorrect (-1), i.e., words or expressions having a different meaning with regard to either the common sense or the legal contest.

To illustrate consider the following legal maxim and a corresponding augmented sample.

A State measure which benefits all undertakings in national territory without distinction cannot therefore constitute State aid.

A department of state bar which profit all attempt in subject soil without preeminence cannot therefore constitute state aid.

While benefits and profit are synonyms playing a perfectly fungible function, distinction and preeminence may recall a slightly different meaning, the former indicating a difference or contrast, the latter referring to superiority, supremacy or excellence, even though the replacement does not change the sentence meaning. Finally, if we consider the word pair national territory and subject soil, the replacement is uncorrected. While the former refers to a legal concept, i.e., the surface area, subsurface, waters and atmosphere comprising the territory of the country and its exclusive economic zone, the
The need to take account of requirements relating to environmental protection, however legitimate, cannot justify the exclusion of selective measures, even specific ones such as environmental levies, from the scope of Article 87(1) EC, as account may in any event usefully be taken of the environmental objectives when the compatibility of the State aid measure with the common market is being assessed pursuant to Article 87(3) EC.

Figure 1: Example of one legal maxim and a synthetic sample obtained after the application of multiple augmentation steps.

Table 1: Human evaluation of single word replacements, with respect to the context.

<table>
<thead>
<tr>
<th>Word</th>
<th>Replacement</th>
<th>Score</th>
<th>Word</th>
<th>Replacement</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>justify</td>
<td>excuse</td>
<td>+1</td>
<td>event</td>
<td>result</td>
<td>+1</td>
</tr>
<tr>
<td>exclusion</td>
<td>expulsion</td>
<td>0</td>
<td>objectives</td>
<td>objective</td>
<td>+1</td>
</tr>
<tr>
<td>measures</td>
<td>measure</td>
<td>+1</td>
<td>State</td>
<td>department</td>
<td>state</td>
</tr>
<tr>
<td>specific</td>
<td>particular</td>
<td>+1</td>
<td>aid</td>
<td>assistance</td>
<td>0</td>
</tr>
<tr>
<td>levies</td>
<td>impose</td>
<td>0</td>
<td>common</td>
<td>usual</td>
<td>-1</td>
</tr>
<tr>
<td>article</td>
<td>clause</td>
<td>0</td>
<td>market</td>
<td>marketplace</td>
<td>0</td>
</tr>
<tr>
<td>account</td>
<td>report</td>
<td>+1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Evaluation of augmentation methods over 10 legal maxims samples. For each augmentation method we report the score obtained for each legal maxim, the sum of such scores, and the average cosine similarity between the sentence embeddings of the synthetic sentence and the original one.

<table>
<thead>
<tr>
<th>Method</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
<th>Total</th>
<th>Avg LB similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>augWN</td>
<td>-3</td>
<td>-5</td>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>-6</td>
<td>-2</td>
<td>-7</td>
<td>-1</td>
<td>-1</td>
<td>-28</td>
<td>0.763</td>
</tr>
<tr>
<td>augWN+POS</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>-2</td>
<td>4</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>15</td>
<td>0.779</td>
</tr>
<tr>
<td>augGV</td>
<td>-3</td>
<td>0</td>
<td>-4</td>
<td>-1</td>
<td>6</td>
<td>0</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>0.879</td>
</tr>
<tr>
<td>augLB</td>
<td>2</td>
<td>-1</td>
<td>3</td>
<td>-1</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>-4</td>
<td>25</td>
<td>35</td>
<td>0.886</td>
</tr>
<tr>
<td>our proposal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>augWN+GV</td>
<td>8</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>35</td>
<td>0.894</td>
</tr>
</tbody>
</table>

latter points to a geological concept, i.e., a mixture of organic matter, minerals, gases, liquids, and organisms that together compose the upper layer of earth.

The evaluation was performed by both experts together, solving disagreements through discussion. We measured which augmentation method preserves better the meaning of the original text by summing together the scores obtained at each step. To perform a fair comparison, we used the same original samples for each of our augmentation methods, and in each step, we replace the same term. In Figure 1 we show a further example of an augmented sample, while Table 1 reports the related evaluation.

As an additional evaluation, we also measured how much the synthetic samples differ from the original ones in terms of distance between their embeddings. We used Legal-BERT (Chalkidis et al., 2020) to generate the sentence embeddings of the two samples and then measured their cosine similarity.
4.3 Results

As can be seen in Table 2, our method seems to be the more robust. Indeed, in the evaluation of the single sources it obtains a negative score only two times, its performance is close to the best method in each case, and it outperforms the alternatives in the total score. Nonetheless, the performance on different legal maxims is highly variable, with scores ranging from +8 to -1.

The performance of augLB is comparable to augWN+GV in most cases, with the remarkable exception of document #10, where the difference between the two scores is above 10 points. Another difference between the two methods is that the substitutions performed through augLB tend to preserve the grammatical rules of the sentences, while the same can not be said for augWN+GV.

The worst performing method is augWN and it is also the only one to obtain a negative total score. The introduction of additional constraints in augWN+POS greatly improves the previous method by about 40 points. augGV does not perform well, obtaining a positive score only in 3 cases.

For what concerns the similarities between embeddings, our method outperforms all the others. However, it is important to remark that the difference between augWN+GV, augLB, and augGN amounts to a few decimals. Surprisingly, augWN+POS does not perform well, obtaining a score about 0.1 lower than augGV.

5 Conclusion

We presented a data augmentation method that leverages both the symbolic information available in knowledge graphs and the sub-symbolic information provided by word embeddings. We have applied this technique to the challenging domain of legal documents and asked a team of experts to evaluate each replacement. The results confirm the quality of our method with respect to alternative approaches, yet they emphasize that more work is needed to obtain satisfactory results. We relied on GloVe since is a popular and widely adopted representation with a low computational footprint. Nonetheless, our proposal can be adapted to other embeddings.

In future work, we plan to further test this technique in a task-based setting where we train a machine learning model to recognize the sentences that contain a principle of law. Moreover, we will apply it to other legal tasks where data is difficult to produce or where some classes are greatly under-represented. Examples of these tasks are argument mining (Poudyal et al., 2020; Habernal et al., 2022; Grundler et al., 2022) and identification of unfair clauses in contracts (Galassi et al., 2020; Drawzeski et al., 2021; Ruggeri et al., 2022).

Acknowledgements

This work has been partially funded by the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement 101017142 (StairwAI), by the European Union’s Justice programme under grant agreement 101007420 (ADELE), by the European Horizon 2020 ERC project CompuLaw (Computable Law) under grant agreement 833647, and by the Italian Ministry of Education and Research’s PRIN programme under grant agreement 2017NCPZ22 (LAILA).

References


International Conference on Computational Linguistics.


