Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts

Sezen Perçin[1,2,3] and **Andrea Galassi**[3] ✉
Francesca Lagioia[4] and **Federico Ruggeri**[3] and **Piera Santin**[4]
Giovanni Sartor[4] and **Paolo Torroni**[3]
[1]Department of Electrical and Electronics Engineering, Boğaziçi University, Turkey
[2]Technische Universität München, Munich, Germany
[3]DISI, University of Bologna, Bologna, Italy
[4]CIRSFID-Alma AI, University of Bologna, Bologna, Italy
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;

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 47 - 52
December 8, 2022 ©2022 Association for Computational Linguistics

- (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.

Other possible data augmentation strategies include rule-based approaches (Wei and Zou, 2019), syntactic alterations (Şahin and Steedman, 2018), interpolation approaches (Zhang et al., 2018), generative data augmentation and back-translation (Sennrich et al., 2016), and random manipulation of words (Yan et al., 2019). Additional information can be found in the surveys by Shorten et al. (2021) and Li et al. (2022).

3 Method

Our augmentation method **augWN+GV** combines the use of the lexical database WordNet (WN) with the properties of the vector space defined by GloVe pre-trained word embeddings (GV).

Given a sample sentence, composed of a list of words $\{w_1, ..., w_n\}$, we randomly choose one word to be replaced among those that are adjectives, nouns, or adverbs. We do so by computing the POS tags of each word POS_{w_i} through the NLTK library and considering only the words for which $POS_{w_i} \in \{NN, NNS, NNP, NNPS, JJ, JJR, JJS, RB, RBR, RBS, RP\}$.[2] Then, given a word w_j to replace, we proceed as follows:

1. we retrieve from WordNet the synsets with a meaningful relationship and the related lemmas;
2. we create a list of 10 candidate lemmas, excluding the original word and giving priority to the synsets whose WordNet POS tag corresponds to POS_{w_j};[3]
3. we encode the word w_j and each candidate through pre-trained GloVe (Pennington et al., 2014) embeddings of size 100;
4. we select the candidate w_k that is most similar to w_j and perform the replacement through cosine similarity.

We compared our method against four baselines:

- **augWN** follows our method for the selection of candidates, but then the choice is not based on GloVe but rather on random selection;
- **augWN+POS** is similar to the previous baseline, but additionally only candidates w_k

[1] https://github.com/adele-project/maxims

[2] We included RP words since they can be used as adverbial particles.

[3] For 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 $\{+1, 0, -1\}$. We have chosen to use a 3-values scale to identify not only replacements that are completely correct (+1) and those that are incorrect (-1), but also those that are imprecise or too informal for our specific domain (0). 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. Figure 1 and Table 1 respectively report an example of an augmented sample and 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

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.

The need to take account of requirements relating to environmental protection, however legitimate, cannot excuse the expulsion of selective measure, even particular ones such as environmental impose, from the scope of clause 87(1) EC, as report may in any result usefully be taken of the environmental objective when the compatibility of the department of state assistance measure with the usual marketplace is being assessed pursuant to clause 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.

Word	**Replacement**	**Score**	**Word**	**Replacement**	**Score**
justify	excuse	+1	event	result	+1
exclusion	expulsion	0	objectives	objective	+1
measures	measure	+1	State	deparment of state	-1
specific	particular	+1	aid	assistance	0
levies	impose	0	common	usual	-1
article	clause	0	market	marketplace	0
account	report	+1			

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.

	Human Evaluation											**Avg LB**
Method	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**Total**	**similarity**
baselines												
augWN	-3	-5	-2	**1**	-2	-6	-2	-7	-1	-1	-28	0.763
augWN+POS	-2	-1	2	-2	4	-1	1	3	**2**	**9**	15	0.779
augGV	-3	**0**	-4	-1	6	0	-3	1	0	7	3	0.879
augLB	2	-1	**3**	-1	**10**	**6**	1	**8**	1	-4	25	0.886
our proposal												
augWN+GV	**8**	-1	2	-1	8	5	**2**	4	0	8	**35**	**0.894**

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 **augGV** 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 underrepresented. 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

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of EMNLP*, pages 2898–2904, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. ACL.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *NLLP@EMNLP*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christiane Fellbaum. 2010. *WordNet*, pages 231–243. Springer Netherlands, Dordrecht.

Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *COLING*, pages 915–926, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2022. Mining legal arguments in court decisions. *CoRR*, abs/2208.06178.

Yang-Yin Lee, Hao Ke, Hen-Hsen Huang, and Hsin-Hsi Chen. 2016. Combining word embedding and lexical database for semantic relatedness measurement. In *WWW (Companion Volume)*, pages 73–74. ACM.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Anna Mosolova, Vadim Fomin, and Ivan Bondarenko. 2018. Text augmentation for neural networks. In *AIST (Supplement)*, volume 2268 of *CEUR Workshop Proceedings*, pages 104–109. CEUR-WS.org.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *ArgMining*, pages 67–75, Online. Association for Computational Linguistics.

Federico Ruggeri, Francesca Lagioia, Marco Lippi, and Paolo Torroni. 2022. Detecting and explaining unfairness in consumer contracts through memory networks. *Artif. Intell. Law*, 30(1):59–92.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *J. Big Data*, 8(1):101.

E Umamaheswari Vasanthakumar and Francis Bond. 2018. Multilingual Wordnet sense ranking using nearest context. In *GWC*, pages 272–283, Singapore. Global Wordnet Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP/IJCNLP (1)*, pages 6381–6387. Association for Computational Linguistics.

Rong Xiang, Emmanuele Chersoni, Yunfei Long, Qin Lu, and Chu-Ren Huang. 2020. Lexical data augmentation for text classification in deep learning. In *Canadian Conference on AI*, volume 12109 of *Lecture Notes in Computer Science*, pages 521–527. Springer.

Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. 2019. Data augmentation for deep learning of judgment documents. In *IScIDE (2)*, volume 11936 of *Lecture Notes in Computer Science*, pages 232–242. Springer.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *ICLR*. OpenReview.net.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.