Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210328

Human Evaluation Experiment of Legal Information Retrieval Methods

Tereza NOVOTNÁ [a,1]
[a] *Institute of Law and Technology, Masaryk University, Brno, Czech Republic*

Abstract. In this article, I present the results of the human evaluation experiment of three commonly used methods in legal information retrieval and a new ”multilayered” approach. I use the doc2vec model, citation network analysis and two topic modelling algorithms for the Czech Supreme Court decisions retrieval and evaluate their performance. To improve the accuracy of the results of these methods, I combine the methods in a ”multilayered” way and perform the subsequent evaluation. Both evaluation experiments are conducted with a group of legal experts to assess the applicability and usability of the methods for legal information retrieval. The combination of the doc2vec and citations is found satisfactory accurate for practical use for the Czech court decisions retrieval.

Keywords. human evaluation, court decisions retrieval, doc2vec, citation analysis, LDA, multilayered approach

1. Introduction and Related Work

In this article, I summarize the results of two year long research on the application of different NLP methods to the Czech court decisions and human evaluation of these methods. In the first phase, I use semantic similarity doc2vec algorithm, citation network analysis and two topic modelling methods (Latent Dirichlet allocation as ”LDA”, non-negative matrix factorization as ”NMF”) to tackle different court decisions retrieval tasks. Afterwards, I evaluate all of the methods in the human evaluation experiment. The results of the first part of the research are rather average, therefore in the second phase I develop a new *multilayered approach* to achieve more accurate results. I again evaluate this approach in the human evaluation experiment and compare the results with the first phase evaluation results. I present here the results of the human evaluation experiments and their comparison.

The general research question that I try to answer is how accurate different commonly used methods for processing court decisions are for lawyers, who frequently perform court decisions research. The second (and more specific) research question is whether the combination of these methods leads to more accurate results. The third question is whether these results are good enough so that these methods could be the basis for practical court decisions search tools, which is a long-term goal of my research.

For the court decisions processing, I use the doc2vec method which was introduced by Le and Mikolov in [6]. In combination with the cosine similarity measure, it was

[1] E-mail: tereza.novotna@law.muni.cz.

successfully used in [7] to retrieve similar statutes or precedents to an in-hand document. Secondly, I use citation network analysis which examines the role of references among the set of legal documents, such as statutes, regulations or court decisions from which it creates the network. It is often applied to case law to observe citation patterns in [8], to improve the performance of a legal information retrieval system in [9] or for ranking of the importance of court decisions for court decision retrieval in [10]. LDA is a topic modelling algorithm introduced by Blei et al. in [1]. Legal documents clustering and summarization is a common application of topic modelling, as was shown in [2,3].

This article is structured as follows. Section 2 briefly describes the source data, the methods and a multilayered application of the methods. Section 3 contains the description of the evaluation experiment design and the evaluation group of lawyers. In Section 4, I summarize the most important results of the evaluation experiment and I discuss and compare the results with the first phase evaluation results. I conclude the article in Section 5.

2. Methodology

I conducted a two-phase evaluation experiment of three legal information retrieval methods. In the first phase, I used doc2vec, citation network analysis metrics and the topic modelling methods LDA and NMF. Afterwards, I asked legal experts to evaluate different tasks performed by these methods. Based on the evaluation results, I conducted a second phase evaluation experiment of the multilayered approach of these three methods.

2.1. Data

I used a dataset of the Czech Supreme court decisions available in the frame of the Czech Court Decisions Corpus (CzCDC 1.0) from [4]. This corpus is the only freely available set of court decisions of the Czech Supreme, Supreme Administrative and Constitutional Court. It contains raw texts of court decisions with basic metadata (date of publication, docket number, court). The Supreme Court subset of decisions contains 111 977 court decisions dated from 1994 to 2018. Nevertheless, I used the subset of the Supreme Court decisions related to the Czech Copyright Act from [11] to narrow the set of decisions to choose from in the evaluation.

2.2. Semantic Similarity - doc2vec

The first of the methods is the doc2vec model for semantically similar documents retrieval. The algorithm was used in standard settings and the model was trained for the whole dataset of the Czech Supreme Court decisions as described in [5]. The model provides for vector representations of court decisions and the similarity is computed as a cosine similarity measure between two vector representations. This method was used to retrieve semantically similar court decisions based on the *cosine similarity* measure and the similarity was evaluated in the evaluation experiment.

2.3. Citation Network Analysis

I used citation data from the freely available dataset of citation data of the Czech courts described in [12]. This data was used to explore several theoretical legal institutes, such as the precedent binding of court decisions of the Czech highest courts or citation practice of the Czech courts. I used *authority score* to indicate the domain importance of decisions and this importance was evaluated by legal experts in the following experiment.

2.4. Topic Modelling - LDA and NMF

I used LDA and NMF methods in the third experiment. Both methods are based on the assumption that the whole dataset consists of a set of latent topics and each document in the dataset is represented by these topics and their probabilities. These topics are characterized by a distribution over words. We again applied them to the dataset of the Supreme Court decisions and used the automatic coherence score metric to select the number of topics that the model should retrieve. The best models were the 30-topic LDA model and the 20-topic NMF model as described in [13]. I used the *three most probable topics* assigned by both models to court decisions and the relevance of the topics to the legal issues in presented decisions was evaluated by legal experts.

2.5. Multilayered Approach

Based on the evaluation results from the first phase of our research, I concluded that none of the three methods is simply applicable as such since the accuracy is not high enough. At the same time, mainly the doc2vec model and citation network analysis measures have the potential to be used when refined. Therefore, I applied the methods in a multilayered approach in the second phase of this research. The assumption behind the idea is that if the methods are applied in sequence, retrieved decisions (or metrics related to them) are refined and the strengths of the methods should be emphasized. I used doc2vec model, as it had the highest evaluation results, as a basic method, and I combined it with *1) citation network analysis* and *2) the 30-topic LDA model* in two partial experiments:

Ad 1) In the first partial experiment, the existence of a citation link between the decisions is used as a subsequent method to refine the doc2vec model. It is assumed that a pair of semantically similar decisions connected with a citation is more similar than a pair of semantically similar decisions without a mutual citation.

Ad 2) In the second partial experiment, the 30-topic LDA model is used as a method for the refinement of results because it is more accurate then NMF (Section 3.2.). It was assumed that a pair of semantically similar decisions with the same topic assigned to them is more similar than a pair of decisions with different topics assigned. Here, it is assumed that refining the most semantically similar decisions with the data on the same assigned topic should lead to a higher rating in evaluation.

As the doc2vec model is the basis of the multilayered approach, therefore the similarity of legal issues and background of court decisions is the evaluated characteristics.

3. Evaluation Experiment Design

The methods and data described in the previous Section are evaluated by the group of legal experts in the evaluation experiments. I look for data on the accuracy of all of the methods and potential improvement of the results of the multilayered approach compared to the other three methods. The general evaluation experiment design is based on asking legal experts to evaluate the accuracy of the methods via evaluation questionnaires. The questions on accuracy of different methods targets the specific goal of the individual experiment. That means, the similarity of retrieved decisions is evaluated for doc2vec model, the domain importance of decisions is evaluated for the citation analysis and the relevance of topics assigned to decisions is evaluated for topic modelling methods. For the multilayered approach, the similarity of decisions is evaluated because this approach is based on the doc2vec model.

3.1. Evaluation Group

The evaluation group in this experiment consists of 46 experts. Legal experts here are practicing lawyers from different legal fields as a high expertise in law is one of the key requirements for the evaluation participants. I asked judges (and court assistants) and lawyers as both of these categories work intensively with court decisions. Although both of the categories are not represented equally, I find it important to evaluate the methods by experts from different legal fields. In the first phase of our research, 26 legal experts participated in the evaluation. In the second phase, 20 legal experts participated in the evaluation.

3.2. Methodology of Evaluation

The evaluators were presented with court decisions to read through and evaluate in the form of online Google Form questionnaires. The evaluation experiment was conducted in two phases in accordance with the schedule described in the Section 2.

In the first phase, legal experts were asked to evaluate the doc2vec model, citation analysis and the two topic modelling method (LDA, NMF). For the doc2vec model, legal experts evaluated the similarity of legal issue and the factual background of the pairs of court decisions. They evaluated 26 pairs of the decisions with the smallest cosine distance (the highest similarity) and 26 pairs with the 10th smallest cosine distance (the 10th highest similarity) for comparison. The results are in the third and fourth column in Table 1. The evaluation scale was from 1 to 6 (1 means the least similar, 6 means the most similar).

Secondly, they evaluated the domain importance of 13 decisions with the highest and 13 decisions with the lowest (zero) authority scores. The evaluation scale was again from 1 to 6 (1 means the least important, 6 means the most important). The decisions with the highest authority score have an average rating of **3.42** and zero authority score decisions have an average rating of **2.73**, which is a significant difference.

Thirdly, they evaluated the relevance of assigned topics to the decisions. They evaluated a totally of 76 decisions with the set of three most probable topics retrieved by each model (LDA and NMF) for each decision. The evaluation scale was again from 1 to 6 (1 means the least relevant, 6 means the most relevant). The mean rating results of both topic modelling methods were rather poor: **2.38** for the LDA model and **2.32** for the NMF model (on the scale from 1 to 6, where 6 means the most relevant).

In the second phase, legal experts evaluated the combination of the doc2vec model and citation analysis data and the doc2vec model and the LDA topics. They were asked to evaluate the similarity of legal issue and the factual background of the pairs of court decisions. Firstly, they evaluated 40 pairs of the decisions with the smallest cosine distance connected with mutual citation. Secondly, they evaluated 20 pairs of the decisions with the smallest cosine distance and with the same topic assigned by the LDA model and 20 pairs with the smallest cosine distance and with the different topic assigned by the LDA model for comparison. The evaluation scale was from 1 to 6 (1 means the least similar, 6 means the most similar).

4. Results and Discussion

I present the results of the evaluation experiments here and I compare the results of first phase experiments (doc2vec, citation analysis, topic modelling) with a multilayered approach (doc2vec and citation analysis, doc2vec and LDA topic model).

4.1. Means and Frequency of Ratings of doc2vec Model and Citations

Firstly, I consider the difference of the mean rating value of the two most similar decisions (third column of Table 1) and of the two most similar decisions connected with a citation (second column of Table 1) significant. Secondly, it is necessary to take into consideration the fact, that the evaluation group was high in legal expertise, however very domain diverse. Evaluated court decisions were decisions related to the Czech Copyright Act. The evaluation group on the other hand consists of lawyers from different legal fields and legal professions. That means that an expert in Copyright law will probably evaluate the same pair of presented decisions differently than a judge of criminal law. Regarding these reasons and regarding the fact, that the model shouldn't generally serve only a domain limited group of lawyers, but it should be used as widely as possible, these results are found sufficient.

The frequency of different rating values in Figure 1 only supports this conclusion. A vast majority of higher similarity ratings leads to the conclusion, that a vast majority of retrieved decisions with a mutual citation link are somehow relevant, even though lawyers find some differences either in the legal issue or in the background. Therefore, I find these results sufficient enough to create a base for the Czech Supreme Court decisions retrieval tool. The possible future steps and limitations of this idea will be discussed in the last Section.

	The most similar with a citation	The most similar	The 10th most similar
Mean value	**4.4**	3.58	3.12

Table 1. Means of evaluation ratings of similarity of court decisions with/without a citation link

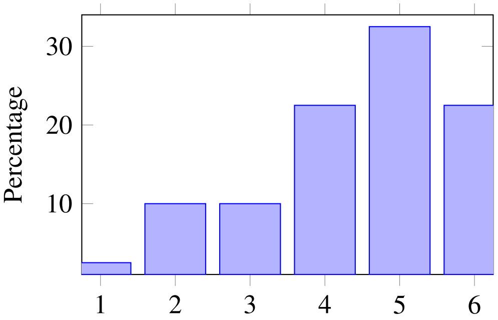

Figure 1. Frequency of evaluation ratings of similarity of court decisions with a citation link

4.2. Means and Frequency of Ratings of doc2vec Model and Topics

The topic modelling methods were the weakest in the first phase of research, LDA had better result than NMF. On the other hand, their potential is great in case the retrieved topics would be accurate enough. As the assigned topics could potentially mean another metadata layer for court decisions or even a very simple summarization of the text. Therefore, I decided to try to apply it in combination with the more accurate doc2vec model to see whether this combination could mean a way forward with LDA. The assumption here is that information on the most probable topic assigned by the 30-topic LDA model could discard potential false positives retrieved by the doc2vec model, i.e. court decisions retrieved with the highest cosine similarity but not relevant. This way, the combination could make the doc2vec model more accurate.

Generally, the results are better, but not great. The mean rating values are in Table 2, the results of the similarity of court decisions for comparison are in Table 1. The combination of methods is even slightly less accurate than the doc2vec applied solely. This combination of methods does not make the retrieval more accurate. On the other hand, when compared to the results of the LDA method itself in Section 3.2., the mean of evaluation ratings is significantly higher. However, this conclusion only supports the accuracy of the doc2vec model rather than the usability of the LDA topic model algorithm.

	The most similar with a same topic	The most similar with a different topic
Mean value	**3.25**	2.4

Table 2. Means of evaluation ratings of similarity of court decisions with assigned topics

5. Conclusion

The multilayered approach - the combination of the doc2vec model and a citation link - showed decent results when compared to the stand-alone application of the methods. The doc2vec model is a generally applicable algorithm with satisfactory results also in different domains, thus it is not a surprise. On the other hand, the citation data are originally created by judges and court assistants, i.e. subjectively and by highly qualified legal experts. Therefore, it is again not a surprise that the citations make the text processing algorithm such as the doc2vec model more accurate and more relevant. On the other hand, the subjectivity, the context of a citation and last but not least, the time relevance of such citations need to be taken into consideration. Secondly, in line with expectations, the LDA method does not show sufficient results to be used in practice. The assigned topics, either alone or even in combination with the doc2vec model, were assessed as significantly less accurate in relation to the presented decisions.

Nevertheless, I find the presented results satisfactory as another step forward to a court decisions retrieval system in the Czech Republic. Additionally, I also consider these results to be important in terms of understanding how the methods used in this article work in practice when applied to legal sources. Although it is still a long way to go, I hope that this paper will lead to the practical application of methods and bridge the gap between legal informatics and daily legal practice in the Czech Republic.

6. Acknowledgment

I acknowledge the support of the ERDF project “Internal grant agency of Masaryk University” (No. CZ.02.2.69/0.0/0.0/19_073/0016943). I would like to thank Jakub Harašta for consultations and ideas for this research.

References

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3(4–5): p. 993–1022.

[2] Kumar VR, Raghuveer K. Legal Document Summarization using Latent Dirichlet Allocation. International Journal of Computer Science and Telecommunications. 2012 July; 3(7): p. 114-117.

[3] Lu Q, Conrad JG, Al-Kofahi K, Keenan W. Legal Document Clustering with Built-in Topic Segmentation. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management; c2011; New York, NY, USA; p. 383–392.

[4] Novotna T, Harasta J. The Czech Court Decisions Corpus (CzCDC): Availability as the First Step. 2019. arXiv preprint arXiv:1910.09513.

[5] Novotna T. Document Similarity of Czech Supreme Court decisions. Masaryk University Journal of Law and Technology. 2020; 14(1): p. 105-122.

[6] Le, Q., Mikolov, T. Distributed Representations of Sentences and Documents. International Conference on Machine Learning; 2014; p. 1188–1196.

[7] Renjit, S., Idicula S. M. CUSAT NLP@AILA-FIRE2019: Similarity in Legal Texts using Document Level Embeddings. Overview of the FIRE 2019 AILA track: Artificial Intelligence for Legal Assistance. Proc. of FIRE; 2019; p. 12-15.

[8] Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., Wahlbeck, P. J. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis; 15(3); 2007; p. 324–346.

[9] Kumar, S. Similarity analysis of legal judgments and applying ‘Paragraph-link’to find similar legal judgments (Doctoral dissertation, Ph. D. thesis, International Institute of Information Technology Hyderabad); 2014.

[10] Geist, A. The Open Revolution: Using Citation Analysis to Improve Legal Text Retrieval. European Journal of Legal Studies; 2008; 2(3); p. 137–145.

[11] Harašta, J. Srovnávací studie právních informačních systémů: Rozdíly mezi systémy při využití různých vyhledávacích strategií. Revue pro právo a technologie; 2020; 11(22); p. 219–260.

[12] Harašta, J., Novotná, T., Šavelka, J. Citation Data of Czech Apex Courts. arXiv:2002.02224; 2020.

[13] Novotná, T., Harašta, J., Kól J. Topic Modelling of the Czech Supreme Court Decisions. In: ASAIL 2020 Automated Semantic Analysis of Information in Legal Text; 2020; p. 1.5.