A Study on Summarization as a Pre-Processing Step for Legal Judgment Prediction

Michelle Kim
kimmic16@msu.edu
Michigan State University, 428 S. Shaw Ln., East Lansing, MI, USA

Abstract

We propose to use *summaries* of legal documents, instead of entire documents, as input data for the task of legal judgment prediction, i.e., the task of ruling in favor or against a legal decision. We also introduce MeritsSum, a dataset of the US Supreme Court Merits cases. MeritsSum, which contains information on judgment decisions and summaries, can be used for two different tasks of legal judgment prediction and text summarization. This paper shows that Recurrent Neural Networks (RNNs) can solve a judgment prediction task with higher accuracy when summaries are used.

1 Introduction

Understanding and processing of legal documents require a significant amount of time and labor of experienced experts. The automation of legal document processing improves the workflow efficiency for experts and gives access to legal information for lay-people. With the recent advances in natural language processing technology, such automation seems accomplishable.

A straightforward approach is to adapt off-the-shelf pre-trained models such as BERT (Devlin et al., 2018) or Seq2Seq (Sutskever et al., 2014) and fine-tune these models on application-specific legal documents of interest. However, this straightforward approach often fails because end-to-end models cannot adequately capture the unique characteristics of legal documents.

One characteristic of legal documents that hinders the adaptation of off-the-shelf models is the lengthiness of documents compared to other domains. As the computational requirements of a Transformer(Vaswani et al., 2017)-based model rise quadratically with length, training legal documents on those models is expensive and time-consuming.

In this paper, we propose to process concise summaries instead of lengthy documents for computational efficiency and better performance. Through extensive experiments on legal judgment prediction, we investigate the potential benefit of summaries on the performance of different off-the-shelf models. Furthermore, we introduce MeritsSum, a publicly available legal dataset [1] to encourage research on automating legal document processing.

The main contributions of this work are as follows: (1) A new perspective of using summarization as a pre-processing method for legal documents is presented. (2) We introduce MeritsSum, a dataset for both legal text summarization and legal judgment prediction tasks. (3) We conduct experiments on MeritsSum to predict legal judgment.

2 Related Work

Legal Judgment Prediction Legal judgment prediction is one of the legal tasks in which Artificial Intelligence (AI) is most actively applied. The goal of the task is to predict the judgment results according to the facts and the statutory articles in the Civil Law and Common Law system (of the United States). Both AI researchers and legal professionals have been actively developing legal judgment prediction algorithms, as evidenced by a growth in recent legal-based workshops and conferences such as Law and Machine Learning ICML 2020 Workshop.

Recently, many researchers have successfully applied state-of-the-art deep learning architectures to solve the legal judgment prediction task. Chen et al. (2019) used the gating mechanism to predict prison terms. Pan et al. (2019) applied the attention technique to incorporate the information from multiple cases. TopJudge (Zhong et al., 2018) predicts multiple subtasks such as charges and penalty terms by formalizing subtasks dependencies with

[1] https://github.com/cozymichelle

a neural encoder and a Directed Acyclic Graph (DAG). Lastly, Wan et al. (2019) applied an audio segmentation technique, typically used for speech recognition, to classify legal documents.

Legal Datasets With the recent increase of online legal information, large-scale datasets such as C-LJP (Xiao et al., 2018), EURLEX57K (Chalkidis et al., 2019), and BillSum (Eidelman, 2019) were published. C-LJP is a legal prediction dataset, containing more than 2.6 million criminal cases from the Supreme People's Court of China. EURLEX57K is an English legal prediction dataset with cases published by the European Court of Human Rights. BillSum is a dataset for summarization of US Congressional and California state bills.

Text Summarization and Classification A few researchers (Kolcz et al., 2001; Shen et al., 2004; Jeong et al., 2016) have integrated text summarization and classification techniques into their studies. Kolcz et al. (2001) uses a summary to reduce the number of features and applies an extraction-based technique. Jeong et al. (2016) improves text classification with the feature-weighting method for text summarization. However, Kolcz et al. (2001) and Jeong et al. (2016) deal with pure text categorization. Shen et al. (2004) proposes a domain-specific algorithm that uses Web-page summarization techniques for preprocessing in Web-page classification. To the best of our knowledge, this is the first work to utilize summarization as a form of preprocessing for solving legal Artificial Intelligence problems.

There are two main approaches to automatic summarization: extractive and abstractive. Extractive approach (Luhn, 1958; Edmundson, 1969; Vodolazova et al., 2013) creates summaries by extracting relevant sentences from a given text. The importance is measured by statistical scores such as term frequency (TF), inverse term frequencies (ITF), and inverse sentence frequencies (ISF). Abstractive approach (Ganesan et al., 2010; Barzilay and McKeown, 2005) generates new sentences through learning semantics features with Natural Language Processing (NLP) techniques.

3 Legal Summarization Dataset

MeritsSum, a legal summarization dataset contains documents of 701 cases and their Summaries of Supreme Court Merits cases from 2007 to 2019. 8,849 files were collected from the online blog of the Supreme Court of the United States (SCOTUS) [2]. The case files in pdf were converted into texts via Adobe and PDFMiner. The Opinion documents and their Summaries were collected from Casetext[3], a legal research website. Summaries are short phrases or sentences written by judges, and the average word length of the summaries is 26 words. Table 1 presents a summary of the statistics of our newly compiled dataset.

	Mean	Max	Min
Document	2987	19461	37
Summary	26	637	3

Table 1: Word length distributions of the MeritsSum Dataset.

Table 2 shows an example of what a typical legal document representation looks like in the MeritsSum Dataset. The first section lists meta-information such as the case title, docket number, date of hearing, and final judgment. The second section contains the *Opinion* document, which is much longer than shown here. Finally, the bottom of Table 2 shows the overall *Summary* of the document.

For the legal judgment prediction experiments, Opinion documents and summaries are used to predict the final judgment of the court. The judgment prediction was simplified into a binary classification problem of whether a case is "Affirmed." For example, judgment "Affirmed" and "Affirmed and remanded" are classified as "Affirmed, while "Reversed and remanded," "Vacated and remanded," and "Reversed" are not.

4 Experiments

We compare the performance of legal task prediction using the collected MeritsSum Dataset. Given either an Opinion document or its summary, the model predicts whether the case is affirmed or not. 80% of the data is used for training, 10% is used for validation, and the remaining 10% is used for testing. PyTorch 1.0.0 was used as the backend framework.

4.1 Results

We implemented classical text classification models, including Convolutional Neural Network (CNN), Deep Pyramid Convolutional Neural Network (DPCNN), Long Short-Term Memory

[2]https://www.scotusblog.com/
[3]https://casetext.com/

Case:	Upper Skagit Indian Tribe v. Lundgren
Docket Number:	No. 17–387.
Date:	05-21-2018
Judgement:	Vacated and remanded
Opinion: Justice GORSUCH delivered the opinion of the Court. Lower courts disagree about the significance of our decision in Compare. Ancestors of the Upper Skagit Tribe lived for centuries along the Skagit River in northwestern Washington State. But as settlers moved across the Cascades and into the region, the federal government sought to make room for them by displacing native tribes. In the treaty that followed with representatives of the Skagit people and others, the tribes agreed to "cede, relinquish, and convey" their lands to the United States in return for $150,000 and other promises. *Treaty of Point Elliott, Jan. 22, 1855, 12 Stat. 927*; see Today's dispute stems from the Upper Skagit Tribe's efforts to recover a portion of the land it lost. ...	
Summary: In Upper Skagit, which the Supreme Court decided after we issued our opinion in *Cayuga I*, neighboring landowners filed an adverse possession action against the Upper Skagit Tribe, seeking to quiet title to a disputed strip of land as to which both groups lay claim.	

Table 2: An example of the SCOTUS dataset.

	ACCURACY		PRECISION		RECALL		F1	
	Doc	Summary	Doc	Summary	Doc	Summary	Doc	Summary
CNN	0.7324	0.6197	0.4545	0.3333	0.2778	0.125	0.3448	0.1818
DPCNN	0.6338	0.4789	0.25	0.1905	0.2222	0.1667	0.2353	0.1778
LSTM	0.6761	**0.7183**	0.3529	**0.6429**	0.3333	**0.375**	0.3429	**0.4737**
BERT	0.6761	0.662	0.3684	0.0	0.3889	0.0	0.3784	0.0
TOPJUDGE	0.7746	0.5775	0.6667	0.25	0.2222	0.125	0.3333	0.1667

Table 3: Results of Model Predictions on the Legal Summarization Dataset for Legal Judgment Prediction. "Doc" represents when full Opinion documents are used, and "Summary" indicates that summaries of the documents are used instead. Boldface text indicates when a model's performance increased when using summaries.

(LSTM), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), as well as the state-of-the-art legal judgment prediction model TopJudge (Zhong et al., 2018). A pre-trained BERT bert-base-uncased style was fine-tuned on our data with 16 epochs, learning rate of 1e-5, and BERTAdam optimizer. All other models were trained with 50 epochs, learning rate of 1e-3 and Adam optimizer. The results are shown in Table 3.

We measured accuracy, precision, recall, and F1 scores to compare the performance of the chosen models. Contrary to our expectation that incorporating summarization would help improve legal judgment prediction, all metrics of all the models except the LSTM model dropped when summaries instead of entire documents were used. Interestingly, BERT with summaries showed the worst performance; the model predicted everything as"not affirmed", which resulted in zero precision, recall, and F_1.

It is important to note that the LSTM model with summaries gave higher scores across all metrics. Another interesting observation is that LSTM had the shortest average word length of misclassified texts for both full documents and summaries, as shown in Table 4. RNNs can capture historical information but struggles with reaching far-away information. Due to these architectural characteristics, LSTM trained on summaries can better capture information and display higher scores.

Also, the LSTM model with summaries performed better than the BERT result using the full documents. Therefore, we conclude that LSTM models may better extract the necessary information when processing summaries than processing long documents, even with attention.

5 Future Work

In future studies, we plan to improve the performance of legal judgment prediction models with the help of pre-trained embeddings and models such as Law2Vec, a dataset of legal word embeddings. The potential of summarization preprocess-

	Doc	Summary
CNN	4071.3	22.0
DPCNN	3406.2	22.0
LSTM	3349.7	20.6
BERT	3632.3	22.5
TOPJUDGE	3965.4	21.0

Table 4: Average word length of misclassified texts

ing steps can also be tested on other legal tasks such as similar case matching and legal question answering. Furthermore, a potential extension of this work is to test the possibility of transfer learning with summarization on other domain-specific inputs such as medical documents.

6 Conclusion

In this paper, we have proposed a novel perspective of applying summarization as a pre-processing method for legal judgment prediction. We examined the proposed perspective on the newly collected MeritsSum, the US legal dataset. Our findings indicate that pre-trained, off-the-shelf embedding models are not ideal for this task. Instead, we conclude that an LSTM is best suited to use the short context available in summaries to achieve a more accurate prediction than models provided with full legal documents.

References

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. *arXiv preprint arXiv:1908.11521*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions.

Hyoungil Jeong, Youngjoong Ko, and Jungyun Seo. 2016. How to improve text summarization and classification by mutual cooperation on an integrated framework. *Expert Systems with Applications*, 60:222–233.

Aleksander Kolcz, Vidya Prabakarmurthi, and Jugal Kalita. 2001. Summarization as feature selection for text categorization. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 365–370.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Sicheng Pan, Tun Lu, Ning Gu, Huajuan Zhang, and Chunlin Xu. 2019. Charge prediction for multi-defendant cases with multi-scale attention. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pages 766–777. Springer.

Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, and Wei-Ying Ma. 2004. Web-page classification through summarization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 242–249.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tatiana Vodolazova, Elena Lloret, Rafael Muñoz, Manuel Palomar, et al. 2013. The role of statistical and semantic features in single-document extractive summarization.

Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. 2019. Long-length legal document classification. *arXiv preprint arXiv:1912.06905*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.