Legal-BigBird: An Adapted Long-Range Transformer for Legal Documents

Preprint · December 2021

DOI: 10.13140/RG.2.2.14172.92802

1 author:

Loïc Kwate Dassi
École Nationale Supérieure d'Informatique et de Mathématiques
3 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Legal-BigBird: An Adapted Long-Range Transformer for Legal Documents

Loic Kwate Dassi
Department of Computer Science
INP-Ensimag
Grenoble, 38400
Loic.Kwate-Dassi@grenoble-inp.org

Abstract

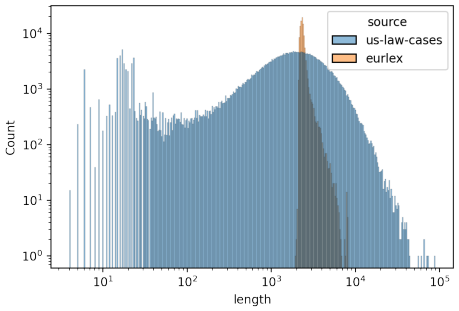
The legal domain is attracting considerable attention in natural language processing (NLP) due to the number of legal documents generated (contracts, business deals, etc.) throughout professional activities and the logical business processing required on that documents. Treat legal documents is particularly cumbersome due to the context-specific knowledge and its extensive length. BigBird has achieved significant performance both on the computational side and on learning representation in the long-range arena. Few researchers have investigated the ability of long-range Transformer models to tackle the knowledge representation problem in the legal domain. We present in this work an adaptation of the long-range Transformer-based model BigBird on legal domain complemented with a use case in legal case retrieval. We continued the training of BigBird with the self-supervised learning task masked language modeling on legal corpora. Without fine-tuning, we tested the pre-trained models on legal case retrieval. We showed that adapting BigBird on legal corpora improves the knowledge representation of documents and outperforms by 5 in accuracy score the vanilla BigBird on the same task.

1 Introduction

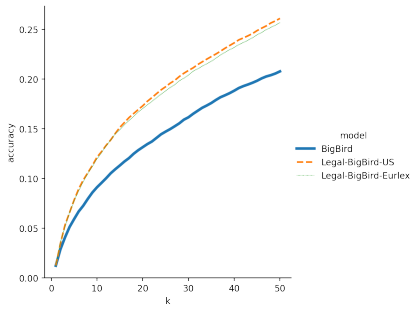
Legal data mining is essential in the legal profession. The constitutional regulations are stated and written by legislation actors in several forms. For example, statutory (created by legislators), cases law (created by judges). The advice or decision-making process made by attorneys should be based on that legal documents and therefore, required a logical search among them. Based on the logical legal processing, various NLP-based tasks have been drawn and a vast amount of researches have been led to address those tasks. The most important NLP-based legal tasks include topic-based legal document classification [5], legal case retrieval [8], legal case entailment [6], statute law retrieval [3], legal textual entailment data corpus [9], legal question answering [11, 2]. The shared part of the foregoing tasks is that they require an insightful vector representation of the documents treated. As those tasks are shaped in natural language, the approaches developed in natural language understanding are concomitantly suitable to address the underpinning document encoding task. Transformer made a breakthrough in the development of approaches for text encoding. [1] proposed a tailored BERT intended to assist legal NLP research. Likewise BERT, Legal-BERT is subject to the same computational limitations due to the full self-attention mechanism while processing long texts. In order to address those limitations, we propose in this work an adjusted BigBird [10] for long legal document encoding. We leveraged the potential of long-range Transformers (BigBird in that case) to scale down the computational cost of legal document processing and extend the legal technology applications. As a use-case, we used Pyserini (a toolkit for reproducible information retrieval research) [4] for the long legal case retrieval task on the dataset provided by [7]. Comparing the retrieval accuracy of each model trained, we effectively showed that the adjustment of BigBird on

NeurIPS 2021 Black in AI Workshop

legal corpora improves the learning representation output. One pillar that sustains this work is the fact that legal documents (legal cases) are especially lengthy as it is shown in figure 1a. Thus the basic Transformers with the quadratic self-attention are computational expensive whilst processing those long documents. Address both the quadratic sequence length complexity and the logical understanding is the main incentive of the use of BigBird.



(a) Document length distribution



(b) top-k accuracy

2 Model

Reduce the quadratic complexity while preserving the encoded features of the full attention has been the central point in the development of the long-range encoder model. BigBird, on average leads the board on the long-range arena task and will be the subject of our work. We trained two versions of BigBird. The first, named Legal-BigBird-us, was trained on the publicly available US legal cases. The second, Legal-BigBird-eurlex, was trained on EURLEX, a large-scale multi-label classification of EU laws. Both models were trained using the following settings: *epochs*: 2; *lr* = 1e-5; *optimizer*: *Adam*; *batch size*: 32, *lr scheduler*: *ReduceOnPlateau*; *lr decay*: 0.75. To investigate the ability of the pretrained to capture insights in legal-context we tested our models on the legal case retrieval task, that is, we indexed the US legal cases of the database [7] with Pyserini using [CSL]'s representation as to the vector representation of each case. For each case, we picked the top k (k ranging from 2 to 50) related cases according to the indexation algorithm of Pyserini. We used the provided edge list of the mapping graph between the legal cases as the ground-truth.

3 Results and discussion

The results showed in figure 1b corroborate our hypothesis that adapting a language model on a specific context will intrinsically improve the knowledge representation herein. Whilst the vanilla BigBird was trained on large-scale and general-purpose corpora, there remain specific domains where transfer learning or and adaptation learning is necessary to tune the pre-trained-model parameters appropriately. Without a specific fine-tuning on the retrieval case task, the two pre-trained versions outperform by 5 the vanilla BigBird.

4 Outlook

On the way of addressing the knowledge representation of long documents in the legal domain, we are extending this work not only on the retrieval tasks but also on the text entailment. Besides predicting the best-related cases given a query, we are working on the challenging task of figuring out why two cases might be tied, that is, we are looking for a way to draw out the two segments of text responsible for the relatedness of two cases. A computer-aided system could be therefore derived to improve the recommendation system used by professionals in the legal domain.

References

[1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[2] Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3] Mi-Young Kim, Juliano Rabelo, and Randy Goebel. Statute law information retrieval and entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL, page 283–289, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations, 2021.

[5] Mariana Y. Noguti, Eduardo Vellasques, and Luiz S. Oliveira. Legal document classification: An application to law area prediction of petitions to public prosecution service. *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul 2020.

[6] Yunqiu Shao, Bulou Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. Thuir@coliee-2020: Leveraging semantic understanding and exact matching for legal case retrieval and entailment, 2020.

[7] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. Legal Document Retrieval using Document Vector Embeddings and Deep Learning. In *Science and information conference*, pages 160–175. Springer, 2018.

[8] Vu Tran, Minh Le Nguyen, and Ken Satoh. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, Jun 2019.

[9] Sabine Wehnert, Sayed Anisul Hoque, Wolfram Fenske, and Gunter Saake. Threshold-based retrieval and textual entailment detection on legal bar exam questions, 2019.

[10] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.

[11] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. Jec-qa: A legal-domain question answering dataset, 2019.