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

Loic Kwate Dassi[1]
[1]National School of Computer Science and Applied Mathematics of Grenoble, France

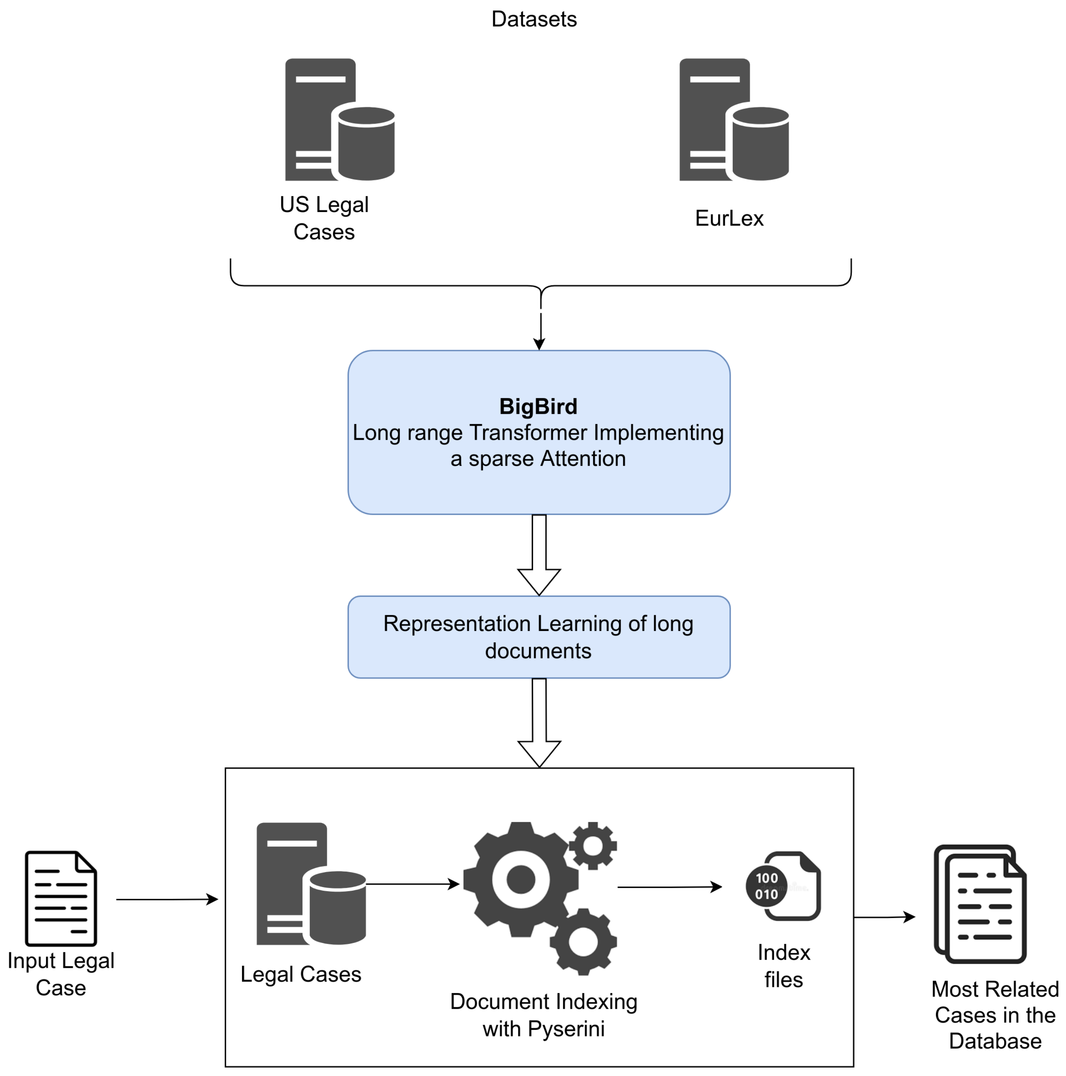

Figure 1: Pipeline Summarisation

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.

Problematic

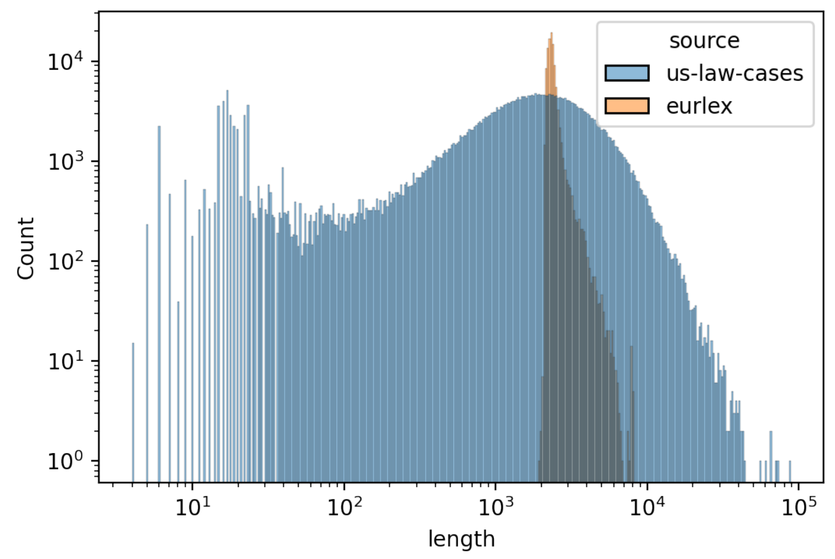

Figure 2: Statistics on Document Length

Methodology

Transformer made a breakthrough in the development of approaches for text encoding. Chalkidis et al., 2020 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 Zaheer et al., 2021 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) Lin et al., 2021 for the long legal case retrieval task on the dataset provided by Sugathadasa et al., 2018. Comparing the retrieval accuracy of each model trained, we effectively showed that the adjustment of BigBird on 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 **??**.

Model

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 Sugathadasa et al., 2018 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.

Experiments and Results

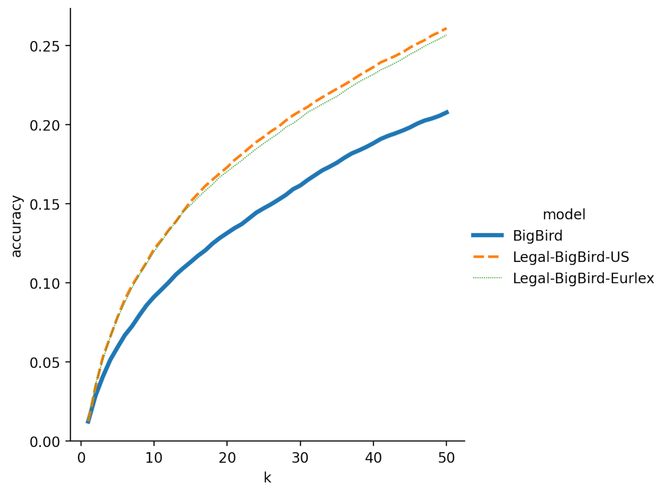

Figure 3: Top-k Accuracy on the Retrieval Task

1 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 et al. *LEGAL-BERT: The Muppets straight out of Law School*. 2020. arXiv: 2010.02559 [cs.CL].

[2] Jimmy Lin et al. *Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations*. 2021. arXiv: 2102.10073 [cs.IR].

[3] Keet Sugathadasa et al. "Legal Document Retrieval using Document Vector Embeddings and Deep Learning". In: *Science and information conference*. Springer. 2018, pp. 160–175. DOI: 10.1007/978-3-030-01177-2_12.

[4] Manzil Zaheer et al. *Big Bird: Transformers for Longer Sequences*. 2021. arXiv: 2007.14062 [cs.LG].