LEGAL-BERT: The Muppets straight out of Law School Supplementary material

Ilias Chalkidis [†‡] **Manos Fergadiotis** [†‡]
Prodromos Malakasiotis [†‡] **Nikolaos Aletras** [*] **Ion Androutsopoulos** [†‡]
[†] Department of Informatics, Athens University of Economics and Business
[‡] Institute of Informatics & Telecommunications, NCSR "Demokritos"
[*] Computer Science Department, University of Sheffield, UK
[ihalk,fergadiotis,rulller,ion]@aueb.gr
n.aletras@sheffield.ac.uk

A Legal NLP datasets

Bellow are the details of the legal NLP datasets we used for the evaluation of our models:

- EURLEX57K (Chalkidis et al., 2019b) contains 57k legislative documents from EURLEX with an average length of 727 words. All documents have been annotated by the Publications Office of EU with concepts from EUROVOC.[1] The average number of labels per document is approx. 5, while many of them are rare. The dataset is split into *training* (45k), *development* (6k), and *test* (6k) documents.
- ECHR-CASES (Chalkidis et al., 2019a) contains approx. 11.5k cases from ECHR's public database. For each case, the dataset provides a list of *facts*. Each case is also mapped to *articles* of the Human Rights Convention that were violated (if any). The dataset can be used for binary classification, where the task is to identify if there was a violation or not, and for multi-label classification where the task is to identify the violated articles.
- CONTRACTS-NER (Chalkidis et al., 2017, 2019d) contains approx. 2k US contracts from EDGAR. Each contract has been annotated with multiple contract elements such as *title*, *parties*, *dates of interest*, *governing law*, *jurisdiction*, *amounts* and *locations*, which have been organized in three groups (*contract header*, *dispute resolution*, *lease details*) based on their position in contracts.

B Implementation details and results on downstream tasks

Below we describe the implementation details for fine-tuning BERT and LEGAL-BERT on the three downstream tasks:

[1] http://eurovoc.europa.eu/

EURLEX57K: We replicate the experiments of Chalkidis et al. (2019c), where a linear layer with L (number of labels) sigmoid activations was placed on top of BERT's [CLS] final representation. We follow the same configuration for all LEGAL-BERT variations.

ECHR-CASES: We replicate the best method of Chalkidis et al. (2019a), which is a hierarchical version of BERT, where initially a shared BERT encodes each case fact independently and produces N fact embeddings ([CLS] representations). A self-attention mechanism, similar to Yang et al. (2016), produces the final document representation. A linear layer with softmax activation gives the final scores.

CONTRACTS-NER We replicate the experiments of Chalkidis et al. (2019d) in all of their three parts (*contract header*, *dispute resolution*, *lease details*). In these experiments, the final representations of the original BERT for all (sentencepiece) tokens in the sequence are fed to a linear CRF layer.

We again follow Chalkidis et al. (2019c,a,d) in the reported evaluation measures.

C Efficiency comparison for various BERT-based models

Recently there has been a debate on the over-parameterization of BERT (Kitaev et al., 2020; Rogers et al., 2020). Towards that directions most studies suggest a parameter sharing technique (Lan et al., 2019) or distillation of BERT by decreasing the number of layers (Sanh et al., 2019). However the main bottleneck of transformers in modern hardware is not primarily the total number of parameters, misinterpreted into the number of stacked layers. Instead Out Of Memory (OOM) issues mainly happen as a product of wider models

in terms of hidden units' dimensionality and the number of attention heads, which affects gradient accumulation in feed-forward and multi-head attention layers (see Table 1). Table 1 shows that LEGAL-BERT-SMALL despite having 3× and 2× the parameters of ALBERT and ALBERT-LARGE has faster training and inference times. We expect models overcoming such limitations to be widely adopted by researchers and practitioners with limited resources. Towards the same direction Google released several lightweight versions of BERT.[2]

Model.	Params	T	HU	AH	Max BS	Training Speed		Inference Speed
						$BS = 1$	$BS = \max$	$BS = 1$
BERT-BASE	110M	12	768	12	6	1.00×	1.00×	1.00×
ALBERT.	12M	12	768	12	12	1.26×	1.21×	1.00×
ALBERT-LARGE	18M	24	1024	12	4	0.49×	0.37×	0.36×
DISTIL-BERT	66M	6	768	12	16	1.66×	2.36×	1.70×
LEGAL-BERT	110M	12	768	12	6	1.00×	1.00×	1.00×
LEGAL-BERT-SMALL	35M	6	512	8	26	2.43×	4.00×	1.70×

Table 1: Comparison of BERT-based models for different batch sizes (BS) in a single 11GB NVIDIA-2080TI. Resource efficiency of the models mostly relies on the number of hidden units (HU), attentions heads (AH) and Transformer blocks T, rather than the number of parameters.

References

I. Chalkidis, I. Androutsopoulos, and A. Michos. 2017. Extracting Contract Elements. In *Proceedings of the International Conference of AI and Law*, London, UK.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317—-4323, Florence, Italy.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019d. Neural Contract Element Extraction Revisited. In *Proceedings of the Document Intellgence Workshop collocated with NeurIPS 2019*, Vancouver, Canada.

[2] https://github.com/google-research/bert

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR*, abs/1909.11942.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *ArXiv*, abs/2002.12327.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.