SRBerta -BERT transformer language model for Serbian legal texts

Miloš Bogdanović[1] and Jelena Tošić[1]

[1]Faculty of Electronic Engineering, University of Niš, 18000, Niš, Serbia,
milos.bogdanovic@elfak.ni.ac.rs, tosic.jelena@elfak.rs

The proofreading of formal language in official documents is a specific challenge that requires domain knowledge regarding grammatical, lexical, and orthographic, but also formal rules used within a domain specific language. A separate part of the previous problem refers to checking the correct use of formal language when writing legislative texts. Such tasks are delegated to specialists, and domain experts, whose daily work could be facilitated by the development of software tools for these specific purposes. The ultimate goal, which is also the biggest challenge, is for a machine to understand a language. For a machine to learn a particular language, it must understand, not only the words and rules used in a particular language, but also the context of sentences and the meaning that words take on in a particular context. In the experimental development that was carried out, the goal of the language model we developer - SRBerta, was to understand the formal language of Serbian legal documents. In 2018, the Google Research AI team presented the BERT (Bidirectional Encoder Representation from Transformers) artificial neural network architecture [1], setting 2 goals: masked language modeling and next sentence prediction. Starting from the BERT architecture, in 2019 the Facebook AI team presented RoBERTa (Robustly optimized BERT pretraining approach), a network optimized for the task of masked language modeling [2]. SRBerta was created on the basis of the RoBERTa architecture, whereby the training of the SRBerta network for the task of understanding the formal language of Serbian legislation was carried out in two phases. In the first phase, the OSCAR dataset was used to train the SRBerta network. OSCAR is a large set of open data created using linguistic classification over data from the Common Crawl corpus [3]. The dataset we used consisted of 645,747 texts. The evaluation of the SRBert network was performed using 10of which consists of 60,000 input sequences, i.e., small texts in the Serbian language. A random masking of 15first stage show that the SRBerta model converges around an accuracy value of 73increases to a value of 73.7Serbian whose words are masked in 15(token) hidden behind it in 73.7In the second phase, SRBerta was fine-tuned using a larger number of available legal texts. This data was gathered from the Legal Information System of the Republic of Serbia. These legislative texts, each of which is between 12 and 15 MB of data, had to be prepared, that is, preprocessed, in a slightly more specific way, with the aim of generating as many input sequences as possible. At the end of the preprocessing process and after the creation of input sequences (tensors), created in a

similar way as in the process of initial training of the network over the Serbian language, masked input sequences for training were created, with a total size of 10,266. Four fine-tuning epochs were performed, with the best-measured value for the accuracy metric being 84.8task of masked language modeling of legislative texts and thus proved the feasibility of creating such a tool based on the previously defined principles of natural language processing. Based on all of the above, the development and testing conclude that it is possible to achieve a high level of accuracy (industrially acceptable of over 90having a sufficiently high-quality and large set of data and an appropriate physical architecture of the system on which we perform the training process.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, Association for Computational Linguistics, Volume **1** (Long and Short Papers) (2019), pages 4171–4186.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019), cite arxiv:1907.11692

[3] OSCAR project (Open Super-large Crawled Aggregated coRpus) (2023), https://huggingface.co/datasets/oscar-corpus/OSCAR-2301