Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT: Optimization of text tokenization for deep learning

SUPERVISOR: dr hab. Łukasz Dębowski, Institute of Computer Science PAS

CO-SUPERVISOR: dr Aleksander Wawer, Institute of Computer Science PAS

DESCRIPTION:

Vector representations of discrete text units such as words or their parts [6] are a fundamental building block of presently used deep learning algorithms in natural language processing, such as transformers [11]. The aim of this doctoral project is to explore systematically whether performance of these algorithms can be improved if one considers more sophisticated methods for tokenization of input data. The standard tokenization methods, such as byte pair encoding (BPE) [9], were historically motivated by unsupervised learning of morphology of natural language [12, 3] and grammar-based data compression algorithms [4, 1]. Although a recent study [8] has claimed that a good tokenization procedure cannot be reduced to simple data compression, one can hypothesize that improved tokenization algorithms should be informed by both information theory [2] and linguistic theories of word morphology [10]. One can also draw inspirations from some results in finding compound terms in terminology extraction [5].

This project includes the design and exploration of alternative approaches to tokenization and their evaluation using large language models on both perplexity and downstream tasks performance. The emphasis will be put on statistical modeling of Polish language [7] and other morphologically rich languages. The prospective candidate is invited to explore these issues in depth experimentally while being aware of the relevant theoretical background and the cost of pretraining and retraining large language models.

The candidate should hold M.Sc. in Computer Science, Computational Linguistics, or Engineering, be knowledgeable in Machine Learning, and possess sufficient computing skills to effectively implement and analyse proposed methods. Scientific curiosity and eagerness to learn are essential.

The candidate should contact the authors of the proposal before a formal submission of documents ({ldebowsk,axw}@ipipan.waw.pl).

BIBLIOGRAPHY:

[1] M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. IEEE Transactions on Information Theory, 51:2554–2576, 2005.

[2] T. M. Cover and J. A. Thomas. Elements of Information Theory, 2nd ed. Wiley & Sons, 2006.

[3] C. G. de Marcken. Unsupervised Language Acquisition. PhD thesis, Massachussetts Institute of Technology, 1996.

[4] J. C. Kieffer and E. Yang. Grammar-based codes: A new class of universal lossless source codes. IEEE Transactions on Information Theory, 46:737–754, 2000.

[5] M. Marciniak, A. Mykowiecka, and P. Rychlik. TermoPL — a flexible tool for terminology extraction. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources

and Evaluation (LREC’16), pages 2278–2284, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1361.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 3111–3119, 2013.

[7] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik. HerBERT: Efficiently pretrained transformer-based language model for Polish, 2021.

[8] C. W. Schmidt, V. Reddy, H. Zhang, A. Alameddine, O. Uzan, Y. Pinter, and C. Tanner. Tokenization is more than compression. https://arxiv.org/abs/2402.18376, 2024.

[9] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In K. Erk and N. A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics, 2016.

[10] P. Smit, S. Virpioja, S.-A. Grönroos, and M. Kurimo. Morfessor 2.0: Toolkit for statistical morphological segmentation. In S. Wintner, M. Tadić, and B. Babych, editors, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 21–24. Association for Computational Linguistics, 2014.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017.

[12] J. G. Wolff. Language acquisition and the discovery of phrase structure. Language and Speech, 23:255–269, 1980.