Ricardo Lukas Jung
6227492
Empirische Sprachwissenschaft (B.A.)
Phonetik & Digital Humanities
15th Semester
s2458588@stud.uni-frankfurt.de

Bachelor Thesis

Lexicalizing a BERT Tokenizer

Building Open-End MLM for Morpho-Syntactically Similar Languages

Ricardo Lukas Jung

Date of Submission:
February 11, 2023

Text Technology Lab
Prof. Dr. Alexander Mehler
Dr. Zakharia Pourtskhvanidze

Erklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen Quellen oder Hilfsmittel als die in dieser Arbeit angegebenen verwendet habe.

_____

Ort, Datum

_____

Unterschrift

Abstract

This is the abstract: what is this about? what was done? what where the results?

Contents

List of Figures

List of Tables

List of Acronyms

bbgc bert-base-german-cased

BERT Bidirectional Encoders from Transformers

BPE Byte Pair Encoding

CL Computational Linguistics

GerParCor German Parliamentary Corpus

HanTa Hanover Tagger

LM Language Model

LSTM Long Short-Term Memory

ML Machine Learning

MLM Masked Language Model

NLP Natural Language Processing

POS Part of Speech

1 Introduction

2 Overview

3 Methodology

4 Results

1 Benchmark

2 Tokenization

Show specific examples of tokenization and analyze the qualitatively (maybe quantitatively)

mlm_wmt_oscar500k	Epoch 1	Epoch 2	Epoch 3	Test score
Precision	0.292614	0.446338	0.71387	0.449735
Recall	0.329531	0.552598	0.73384	0.474525
F1	0.242739	0.473851	0.69194	0.442827

Table 1: Metrics for masked language model trained on the Oscar dataset with infused Wordmap tokenization. Evaluated on sequence classification task.

mlm_wmt_gpc500k	Epoch 1	Epoch 2	Epoch 3	Test score
Precision	0.237664	0.399781	0.603534	0.441891
Recall	0.244613	0.463878	0.637516	0.440304
F1	0.163024	0.389905	0.590542	0.389116

Table 2: Metrics for masked language model trained on the GerParCor dataset with Wordmap infused tokenization. Evaluated on sequence classification task.

mlm_std_oscar500k	Epoch 1	Epoch 2	Epoch 3	Test scores
Precision	0.269615	0.422096	0.596987	0.395879
Recall	0.351077	0.501901	0.657795	0.446388
F1	0.266260	0.412598	0.604824	0.405168

Table 3: Metrics for masked language model trained on the GerParCor dataset with bbgc tokenization. Evaluated on sequence classification task.

mlm_std_gpc500k	Epoch 1	Epoch 2	Epoch 3	Test scores
Precision	0.297466	0.517110	0.656808	0.439873
Recall	0.359949	0.544994	0.676806	0.439924
F1	0.267111	0.480420	0.626593	0.392490

Table 4: Metrics for masked language model trained on the GerParCor dataset with bbgc tokenization. Evaluated on sequence classification task.

bbgc	Epoch 1	Epoch 2	Epoch 3	Test scores
Precision	0.646150	0.768675	0.860180	0.622436
Recall	0.709759	0.804816	0.883397	0.637262
F1	0.660588	0.778371	0.868166	0.624789

Table 5: Metrics for masked language model baseline bert-base-german-cased[1]. Evaluated on sequence classification task.

Summary	bbgc	std+oscar	std+gpc	wmt+oscar	wmt+gpc
Precision	0.622436	0.395879	0.439873	0.441891	0.449735
Recall	0.637262	0.446388	0.439924	0.440304	0.474525
F1	0.624789	0.405168	0.392490	0.389116	0.442827

Table 6: Test score summary for all evaluated models.

5 Discussion

6 Conclusion

7 Testchapter

1 Citing

Abrami et al. 2022

2 Quoting

> “This is a quote by textquote” (DeepL 2021) “This is a quote by enquote”

3 Referencing

Short reference ??
Long reference ??
monofont for code or string `monofont`

Bibliography

Abrami, Giuseppe, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler (June 2022). "German Parliamentary Corpus (GerParCor)". In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1900–1906. URL: https://aclanthology.org/2022.lrec-1.202.

Aepli, Noëmi and Rico Sennrich (2021). *Improving Zero-shot Cross-lingual Transfer between Closely Related Languages by injecting Character-level Noise*. DOI: 10.48550/ARXIV.2109.06772. URL: https://arxiv.org/abs/2109.06772.

Aikhenvald, Alexandra and R Dixon (2017). *The Cambridge Handbook of Linguistic Typology*. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. DOI: 10.1017/9781316135716.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett, eds. (2017). *Morphological Complexity*. Cambridge studies in Linguistics 153. Cambridge University Press. ISBN: 1107120640. DOI: 10.1017/9781316343074. URL: www.cambridge.org/9781107120648.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić (Dec. 2016). "A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 142–153. URL: https://aclanthology.org/W16-4117.

Bouckaert, Remco et al. (2012). "Mapping the Origins and Expansion of the Indo-European Language Family". In: *Science* 337.6097, pp. 957–960. DOI: 10.1126/science.1219669. eprint: https://www.science.org/doi/pdf/10.1126/science.1219669. URL: https://www.science.org/doi/abs/10.1126/science.1219669.

Colman, Andrew M. (2009). *morpheme*. DOI: 10.1093/acref/9780199534067.013.5219. URL: https://www.oxfordreference.com/view/10.1093/acref/9780199534067.001.0001/acref-9780199534067-e-5219.

Çöltekin, Çağrı and Taraka Rama (2022). "What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity". In: *Linguistics Vanguard*. DOI: doi:10.1515/lingvan-2021-0007. URL: https://doi.org/10.1515/lingvan-2021-0007.

Comrie, Bernard (1989). "Morphological Typology". In: *Language universals and linguistic typology*. 2nd ed. University of Chicago Press, pp. 42–56.

Creutz, Mathias and Krista Lagus (2002). "Unsupervised Discovery of Morphemes". In: *CoRR* cs.CL/0205057. URL: https://arxiv.org/abs/cs/0205057.

DeepL (Nov. 2021). *How does deepl work?* URL: https://www.deepl.com/en/blog/how-does-deepl-work. Last accessed: 28.12.2022.

Ehret, Katharina, Alice Blumenthal-Dramé, Christian Bentz, and Aleksandrs Berdicevskis (2021). "Meaning and Measures: Interpreting and Evaluating Complexity Metrics". In: *Frontiers in Communication* 6. ISSN: 2297-900X. DOI: 10.3389/fcomm.2021.640510. URL: https://www.frontiersin.org/articles/10.3389/fcomm.2021.640510.

Glück, Helmut and Michael Rödel, eds. (2016). *Metzler Lexikon Sprache*. ger. 5th ed. Springer eBook Collection. Stuttgart: J.B. Metzler, pp. 141–142. ISBN: 978-3-476-05486-9. DOI: 10.1007/978-3-476-05486-9. URL: http://dx.doi.org/10.1007/978-3-476-05486-9.

Hofmann, Valentin, Hinrich Schuetze, and Janet Pierrehumbert (May 2022). "An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 385–393. DOI: 10.18653/v1/2022.acl-short.43. URL: https://aclanthology.org/2022.acl-short.43.

Lehman, Christian W. (Aug. 2022). *Indogermanisch*. URL: https://www.christianlehmann.eu/ling/sprachen/indogermania/RomGesch/idg.php.

Peters, Ben and Andre F. T. Martins (July 2022). "Beyond Characters: Subword-level Morpheme Segmentation". In: *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 131–138. DOI: 10.18653/v1/2022.sigmorphon-1.14. URL: https://aclanthology.org/2022.sigmorphon-1.14.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training". In: URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych (2020). *How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models*. DOI: 10.48550/ARXIV.2012.15613. URL: https://arxiv.org/abs/2012.15613.

Schuster, Mike and Kaisuke Nakajima (2012). "Japanese and Korean voice search". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152. DOI: 10.1109/ICASSP.2012.6289079.

Schwartz, Lane et al. (2020). *Neural Polysynthetic Language Modelling*. DOI: 10.48550/ARXIV.2005.05477. URL: https://arxiv.org/abs/2005.05477.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

Toraman, Cagri, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik (2022). *Impact of Tokenization on Language Models: An Analysis for Turkish*. DOI: 10.48550/ARXIV.2204.08832. URL: https://arxiv.org/abs/2204.08832.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

Wu, Yonghui et al. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. DOI: 10.48550/ARXIV.1609.08144. URL: https://arxiv.org/abs/1609.08144.