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

Thesis submitted in fulfilment of the requirements for the degree of Bachelor of Arts

Lexicalizing a BERT Tokenizer

Building Open-End MLM for Morpho-Syntactically Similar Languages

Ricardo Lukas Jung

Date of Submission:
December 20, 2022

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

Contents

1 Introduction

The field of NLP *natural language processing* (Glück and Rödel 2016) has been expanded ever since the emergence of the language models. Natural language processing is understood as the

cite (Glück and Rödel 2016)
citeast (2016)

cite (Tenenbaum et al. 2011)
citeast 2011)
The intent of this thesis is to inject linguistic bias into the machine learning framework of BERT to sharpen the analytical capacities of a masked language model. This is done by altering the

Bibliography

Glück, Helmut and Michael Rödel, eds. (2016). *Metzler Lexikon Sprache*. ger. 5th ed. Springer eBook Collection. Stuttgart: J.B. Metzler, Online–Ressource (XXVI, 814 S. 64 Abb., 12 Abb. in Farbe, online resource). 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.

Tenenbaum, Joshua, Charles Kemp, Thomas Griffiths, and Noah Goodman (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction". In: *Science* 331.6022, pp. 1279–1285. DOI: 10.1126/science.1192788.