Ricardo Lukas Jung
6227492
Empirische Sprachwissenschaft (B.A.)
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:
January 2, 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

Contents

List of Figures

List of Tables

List of Acronyms

ML Machine Learning

CL Computational Linguistics

MLM Masked Language Model

BERT Bidirectional Encoders from Transformers

1 Introduction

This chapter covers the background, intentions and scope of the thesis. Explain what the thesis is about.

This work shows how crucial the intervention in subsystems of machine learning is. Fundamental processes like tokenization carry

1.1 Motivation

There is an ongoing urge in the Computational Linguistics (CL) community to understand natural language. Research in the past decades shows use of frequentist and statistical methods (such as ZITATION) to their advantage, leading to the emergence of the first machine learning (ML) models. It became apparent that these ML models are the best currently available approach to an automated understanding of natural language. The structural parallels of machine learning to human learning have often been drawn (ZITATION)) to demonstrate how similar and more importantly: how different both can be. A powerful feature of Machine Learning (ML) (as opposed to human learning) is the possibility of actively controlling the the learning parameters in a supervised environment. To test the efficiency of ML parameters a variety of tasks (ZITATION) are designed and applied. A trained model will yield performance scores based on the quality of its training, much like humans on language tests. But the automated modeling of language is not the first instance language modelling in a broader sense. Traditional linguistics (DEFINITION has produced fundamental research the prior to the discovery of ML architectures and their implementation. While generic ML frameworks seem appealing in the presumption that they require less work to reach somewhat satisfactoy results, they are far from complete or perfect. The integration of aforementioned traditional linguistic knowledge into learning processes for machine learning is the underlying motivation of this thesis.

Language learners usually build up a lexicon consisting of lexemes which they will have to analyze accurately in order to be productive in that target language. A ML model relies on a tokenizer to create such a vocabulary (**ZITATION**). It is programmed to segment tokens into subwords (if possible) and provide a vocabulary comprising all the components needed to analyze a given string. Ideally those subwords will be part of the functional vocabulary in the target language, so called morphemes **ERKLÄRUNG**. A morpheme is defined as the smallest unit carrying meaning in a language. The morphemes of a language and its generated tokenizer vocabulary rarely coincide. Typically, tokanizer vocabularies will contain a lot of noise and linguistically nonsensical segmentations or words. Following the guiding principle that **input quality is ouput quality** not only in language learning, the morpheme vocabulary is identified as the point of leverage in the upcoming section. Note: explain why i use tokens and words, they are interchaneable right? holistic, need less attention to produce satisfactory

1.2 Overview

This section provides an overview of the current state of the art and attempts that have been explored in the past.

The need for performant language models caused the modeling approaches to fan out, following previous findings of linguistics and computational sciences. This part will focus on summarizing relevant findings concerning the most relevant part of this thesis: tokenization. Tokenization is the process of splitting tokens into further segments. Two main algorithms have emerged to tackle this task. One being byte pair encoding (Sennrich, Haddow, and Birch 2016; Gage 1994)

1.3 Scope and Structure

1.4 Hypotheses

2 Methodoloy

2.1 Machine Learning Model

Explain BERT and MLM tweak des tokenizers

2.2 Data

The field of NLP (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 (DeepL 2021)
citeast ()2021)

> This is a quote

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

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

Gage, Philip (Feb. 1994). "A New Algorithm for Data Compression". In: *C Users J.* 12.2, pp. 23–38. ISSN: 0898-9788.

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.

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.