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:
January 5, 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

BERT Bidirectional Encoders from Transformers

CL Computational Linguistics

LM Language Model

LSTM Long Short-Term Memory

ML Machine Learning

MLM Masked Language Model

NLP Natural Language Processing

1 Introduction

This thesis shows the use of specific intervention in tokenization subsystems of machine learning 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 WHY IS THIS SUBJECT RELEVANT This chapter covers the background, intentions and scope of the thesis. Explain what the thesis is about.

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 interchangeable right? holistic, need less attention to produce satisfactory

1.2 Hypotheses

The following research questions will be formulated for testing:

HYP1: Adjustments to tokenization have significant impact on an Language Model (LM)s performance.

How to achieve this hypothesis?

HYP2: Providing lexical information to a tokenizer increases benchmark accuracy on MLM tasks.

How to achieve this hypothesis?

1.3 Scope and Structure

What is covered and what not? What is the shape of this thesis and what order does it have?

2 Overview

2.1 State of the Art

describe the most recent findings on morphologically pretrained models in machine learning literature

3 Methodoloy

in this section the whole methodoloy is covered. what do i use in this thesis, why do i use it and lastly, how? make sure the why covers methodological implications. (vergiss nicht alle pakete als quelle im Anhang)

3.1 Requirements

A series of tools will help to achieve lexicalized tokenization. They will be explained in this chapter along with their methodological edge.

3.1.1 Machine Learning Model

Bidirectional Encoders from Transformers (BERT) is a language learning transformer model designed for Natural Language Processing (NLP) tasks (Vaswani et al. 2017). Upon release it achieved higher performance scores compared to previously used Long Short-Term Memory (LSTM) models (Devlin et al. 2018). Two main model characteristics can be observed for BERT. Firstly, it is the first LM to implement simultaneous attention heads, allowing for bidirectional reading. The methodological implication of reading to the left and right of a token is to include more information about the language in single embeddings. Secondly, BERT introduced the (at the time novel) Masked Language Model (MLM) method for training. The method involves masking a specified (default 15%) amount of random tokens in the input sequence. Masked tokens are guessed by the model which can then update its weights according to success or failure.

The NLP community has since developed BERT and adapted it to the needs of contemporary NLP problems (roberta, germanbert, mbert CITATION). Its wide support, comparability and versatility make BERT the model of choice for this thesis. Another notable feature in BERT is the implementation of the WordPiece tokenizer module (QUELLE?). Default BERT WordPiece tokenization is predominantly heuristic by combining strings based on a precalculated score. A variety of pre-trained tokenizers are available, although they come with a caveat. Once a tokenizer is trained on a dataset it is specific to that dataset. This means the application of a tokenizer on another dataset may result in out-of-vocabulary issues and different token/subtoken distributions.

Particularly relevant to this thesis is the option to train an own tokenizer from the base module. Usually, WordPiece generates its own set of subtokens called *vocabulary*. Tokens are then WORDPIECE ALGORITHMUS ERKLÄREN By providing an algorithmically generated vocabulary to WordPiece and then training it on a new dataset the tokenization behavior is changed.

3.1.2 Data

explain the data that is used

3.1.3 Benchmark

explain olmpics

3.2 Implementation

Tatsächliche Anwendung der Methoden auf die Daten

3.2.1 Tokenizer

Generating a custom pre-training vocabulary

Tokenizer Training

How did I train the tokenizer, how did it go? Which problems arose? What went well?

3.2.2 Masked Language Model

Model implementation and parameters, runtimes?

3.2.3 oLMpics Benchmark

tweak des tokenizers: segmentation ist eine frage der interpretation. The Ultimately, segmentation is a matter of interpretation. As mentioned in 3.1.1, the default WordPiece Tokenizer lacks A linguistically informed

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

4 Results

4.1 Tokenization

5 Testchapter

5.1 Citing

cite (Glück and Rödel 2016)
cite asterisk (2016)

5.2 Quoting

> This is a quote

5.3 Referencing

Short reference 3.1.1
Long reference subsection 3.1.1

Bibliography

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

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.

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