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 12, 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

BWE Bert WordPiece

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

TTLab Text Technology Lab

WM Wordmap

arabic

1 Introduction

2 Overview

Morphological tokenization can be understood as the process of identifying segments in text that are a productive in a given language, carrying meaning and hence also fitting the definition of a morpheme. describe the most recent findings on morphologically pretrained models in machine learning literature

2.1 Related Works

In the past, many efforts towards morphological tokenization have been made. This thesis was mainly inspired by the FLOTA

Earlier generalized attempts like morfessor (Creutz and Lagus 2002) have been outperformed by Sequence based models that also use linguistic morphology Peters and Martins 2022 . Notably, top-down generation of subword vocabularies has shown promising results for tokenization in fusional languages. This aligns with the notion that standard BPE (Sennrich, Haddow, and Birch 2016) or WordPiece Wu et al. 2016 tokenization effectivity suffers from complex morphology causing a big vocabulary. The overall comparison Peters and Martins 2022, p. 134 shows an increase in performance for languages of similar morphological complexity. It is interesting to see that this form of tokenization performs less well for English, a language that has seen a decline in morphology. Much better benchmarks are reached applying its agglutinative fusional peers, e.g. Italian, Latin, Spanish, Russian Toraman et al. find that the vocabulary size plays a special role in morphological tokenization and even define a ratio vocabulary size ratio between 20 to 40% to the number of model parameters depending on the type of tokenizer (Toraman et al. 2022, pp. 11–12). Since tokenization and vocabulary are obviously interdependent, the vast amount of typological variety seen in languages raises the question: is there a right way of tokenizing? This issue is addressed by Rust et al., where a mid-scale investigation was done to see whether different languages actually need more specific tokenizers compared to generalized tokenizers. They report an improvement of model accuracy and F-score across all tasks and languages (Rust et al. 2020). While this sketches a commission for Natural Language Processing (NLP) to always consider choosing a method tailored for single languages, the answer to the problem of performance versus maintenance in models might not be as elaborate as treating every language

singularly. Every language is undoubtedly unique, but that does not rule out simplification by means of further classifying and grouping target languages. In an effort to explain the provenience and relatedness of languages many tools in the domain of typology, NLP and indo-european studies have been constructed.

Whether it be identification of morphological features (Comrie 1989, pp. 42–56), complexity measures (Çöltekin and Rama 2022) or connection through reconstruction (Bouckaert et al. 2012), the different linguistic disciplines suggest observable regularities by which to morphologically group languages. Leveraging the relatedness of languages in NLP is not a new idea in tokenization or Part of Speech (POS) -tagging, but is seeing mixed results up to this day, even with augmentation methods (Aepli and Sennrich 2021). The options seem to branch out quickly, but the mechanism of clearly separating lexemic and functional information seems inherent to most languages. The way they differ is in they combine grammatical functions in morphemes (fusion) and bind them to lexemic morphemes (synthesis). This may be why approaches with stemming, lemmatization or other morphological analyses are very relevant to building good tokenizers for all languages on the isolating to synthetic spectrum (Schwartz et al. 2020, pp. 51–53).

2.2 Target Languages

This section identifies target languages that share common morphological features with German. It is assumed that languages of the same morphological type will behave similarly when analyzed morphologically. German was selected to serve as example language within the family of fusional languages. The aim is not to propose yet another case study of German, but to introduce German as a surrogate to further the scope of application on other languages of similar morphological complexity.

German (ISO639-3: `deu`) is a west-germanic language and the official language of Germany, Austria, Switzerland, Liechtenstein and Luxemburg (Glück and Rödel 2016). It is an inflectional synthetic language with approximately 130 million speakers[1]. German is largely researched and is still paid much attention to in the domain of (computational) linguistics.

On one side, linguistic typology has come up with many useful classifications for languages. On the other, in the pursuit of reconstructing languages Indo-European studies have established a widely accepted phylogenetic model of the diachronic dependecy of Indo-European languages. Both disciplines contribute to language classicifications that are used

[1] https://de.statista.com/statistik/daten/studie/1119851/umfrage/deutschsprachige-menschen-weltweit/ Last accessed: 09.01.23

in this subsection.

Morphological complexity is a term to describe how languages use paradigms to connect grammatical information with lexemic information (Baerman, Brown, and Corbett 2017). Mind that morphological complexity is a nominal category to describe gradients of function-to-morpheme correspondence and measure of morphematic agreement, not a qualitative assessment. The common denominators that make languages morphologically complex are their morphological features. Those languages that use affixation, fusion, composition and derivation (among others) are all fit candidates compared to german. A summary of morphological typology is provided in 2017, pp. 78–93).

Due to the scope of this thesis, German will be the exemplary target language for the experimental setup. Its morphological complexity can be compared to other related or non-related languages as shown in Table 1. There still is no universally accepted measure the complexity of a language due to , but groupings exist on different parameters:

Language	**Similarity**	**ISO 639-3**
Norwegian	Closely Related	isl
Danish	Closely Related	nor
Dutch	Closely Related	nld
English	Closely Related	eng
Icelandic	Closely Related	isl
Romanian	Morphology	ron
Spanish	Morphology	spa
Finnish	Morphology	fin
Italian	Morphology	ita
Hungarian	Morphology	hun

Table 1: Listing of languages similar to German given the type of similarity based off Lehman (2022) and Ehret et al. (2021). ISO identifiers provided at WALS[2].

There have been interpolations between human judgements and statistical measures (on similarity of languages) which can be taken into consideration (Bentz et al. 2016). The point to be taken is that while there is no definite proof of concept for tokenizations being effective when connecting target languages through morphological parameters, there is a strong suggestion in data and intuition of researchers that tokenization for morphologically similar languages should profit from these similarities.

With an arguable exception to English, the languages in Table 1 treat their lexemes with similar morphological processes. The upcoming interlinear glossings (as per Leipzig Glossing

Rules[3]) provide examples for inflectional morphology in verbs within this group of languages. To outline word formation processes, a glossing from Nikanne (2017, p. 71) is considered:

(1) Tytöt istu-i-vat tuolilla
girls sit-PST-3PL chair.ADE
'The girls sat on the chair.'

This Finnish sentence is a textbook example of agglutinative inflectional morphology. The verb {istu} is inflected by suffixing two morphemes marking the past tense {i} and the third person plural {vat} (curled brackets denote morpheme boundaries). In this case every morpheme expresses one grammatical function, apart from {istu} which contains the lexical information for the verb "to sit ". The functional morphemes in example (1) follow the word to be inflected. With many more functional morphemes present in Finnish, Nikanne reports that there is an order in which inflectional morphemes usually appear. In consequence, analyzing Finnish verbs results in different but reoccuring patterns depending on the degree of inflection. The lexemic morpheme {istu} can be modified by {i} alone to just express past tense and still be productive. Following up on the idea of agglutination, Hungarian applies an slightly different strategy to achieve inflection:

(2) Tegnap meg-hallgattunk egy lányt.
yesterday PRF-listen.PST.1PL a girl.ACC
'Yesterday we interviewed a girl.'

Hungarian is also classifed as an agglutinative language for its frequent use of affixes. It is additionally known to combine several grammatical functions into one morphene as can be seen in example (2) as given by Kiss and Hegedus (2021, p. 262), making it a hybrid of agglutinative and fusional. The analysis of the verb in (2) does not allow canonical segmentation to the stem although there is an underlying form {hall;∗} meaning "hear ". Instead, it exists in inflection paradigms like the given example {hallgatunk} combinding the grammatical categories past tense and first person plural. The morphemes addressed so far where either suffixed or not segmentable. As lexical part {hallgatunk} receives a prefix {meg} expressing perfect tense. This use of morphemes can also be seen in germanic or romance languages, like Italian (adapted from Iacobini and Masini (2005, p. 163)):

[3]https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

(3) far=se=la sotto
do-REFL.PRT-PRON.PRT under
‘To quake in one's boots’’

Here the {se} and {la} carry two functions each and are suffixed to {far}, showing that there are morphological types in between agglutinating and fusioning. Arguably, {se} being a clitic pronoun that will appear in different positions acting as indirect object, but never independent of the verb.

After a partial look on the classified languages two important empirical descriptive caveats remain: there are exceptions to almost every regularity in languages. No language is entirely consistent in following a morphosyntactical paradigm, meaning no language is entirely fusional or agglutinative (same applies to the synthetic and isoling spectrum). Judging from the word shapes in the data and literature, the way languages modify their stems or lexemic morphemes is largely based on affixation. In a tokenizer acknowledging lexemic parts of words, the knowledge of word formation in the target language should be conveyed.

3 Methodology

4 Results

5 Discussion

6 Conclusion

7 Testchapter

7.1 Citing

Abrami et al. 2022

7.2 Quoting

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

7.3 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.

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.

Iacobini, Claudio and Francesca Masini (Sept. 2005). "Verb-particle Constructions and Prefixed Verbs in Italian: Typology,Diachrony and Semantics". In: *Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*. Università degli Studi di Bologna. URL: https://www.academia.edu/1183339/Verb_particle_constructions_and_prefixed_verbs_in_Italian_typology_diachrony_and_semantics.

Kiss, Katalin É. and Veronika Hegedus, eds. (2021). *Postpositions and Postpositional Phrases*. Amsterdam: Amsterdam University Press, p. 262. ISBN: 9789048544608. DOI: doi:10.1515/9789048544608. URL: https://doi.org/10.1515/9789048544608.

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

Nikanne, Urpo (Dec. 2017). *Finite sentences in Finnish: Word order, morphology, and information structure*. DOI: 10.5281/zenodo.1117710. URL: https://doi.org/10.5281/zenodo.1117710.

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.

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.

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: Asso-

ciation 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.

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.