Master thesis review

Faculty of Mathematics and Physics, Charles University

Thesis author	Jiří Balhar
Thesis title	Improving Subword Tokenization Methods for Multilingual Models
Submission year	2023
Study program	Computer Science **Study branch** Artificial Intelligence

Review author	Mgr. Martin Popel, Ph.D. **Role** Reviewer
Department	ÚFAL MFF UK

Review text:

The thesis tries to find optimal (subword) tokenizers for large multilingual neural language models (LLMs). To this end, it defines several intrinsic metrics for assesing tokenizers without training LLMs, i.e. quite cheaply. It is shown that these intrinsic metrics correlate well with extrinsic evaluation on several downstream NLP tasks. The thesis confirms three previous works in observing improved performance on low-resource languages when better balancing the vocabulary, but shows that this can be achieved by a much simpler way – sampling the training data for the tokenizer uniformly across languages.

I appreciate the overview of related work, which very nicely summarizes relevant papers including very recent ones. It is obvious that the author understands deeply the papers and approriately comments on unclear parts or even criticizes some of the decisions. The only exception I found is in Algorithm 1 (BPE), where the line "p ← most frequent pair in V" is wrong (probably because not understanding the Python code in the original paper). It should be the most frequent pair of subwords in the training data, not in the vocabulary.

Section 1.1 states that *The work on this thesis began as a collaboration with Ing. Tomasz Limisiewicz on the paper "Tokenization Impacts Multilingual Language Modeling"*. I acknowledge that this paper has been accepted to ACL Findings 2023.[1] I would appreciate a more explicit statement explaining which parts of the thesis are based on this paper and what is the unique contribution of the thesis author vs. other authors of the paper (Tomasz Limisiewicz and David Mareček).[2]

I like the theoretical analysis of relations between Average Rank and other measures in Chapter 3. It is sound and I found some parts novel for me (maybe even "surprising" as stated in Section 3.4.2).

[1] `https://aclanthology.org/2023.findings-acl.350.pdf`

[2] The thesis could also provide a link to `https://github.com/tomlimi/entangled_in_scripts/graphs/contributors` showing that Jiří Balhar is the author of about 40% of the commits in this repo (in addition to all commits in the main repo `https://github.com/kukas/multilingual-tokenizers`).

The attached source codes are of high quality and well documented. They include a re-implementation of two vocabulary balancing methods whose source codes were not publicly available yet. A notable number of experiments has been done, while taking into account their computational requirements (and carbon footprint) and cleverly downscaling the size.

The correlation between intrinsic and extrinsic metrics is measured only for three implementations (not comparing Huggingface and Sentencepiece implementation of Unigram) and by varying only the language of the test set in Figure 4.1 (and the "source" language, i.e. langauge of the training set in Figure 4.2). The effect of the parameters studied in Chapter 5, most notably the data balancing parameter α, is not explored. This is understandable because training LLMs is an resource-expensive experiment. However, Chapter 5 interprets higher values of CPT and AR as better even when varying parameters, where the correlation with extrinsic metrics has not been proven, e.g. "*the marginal benefit of adding more data to the high-resource languages is lower than the incurred cost on the quality of tokenization for the low-resource languages*". We know that high-resource languages have more data for training the tokenizer, but also for (pre-)training the LLM and usually also for fine-tuning the downstream tasks, so while $\alpha = 0$ seems to be the optimal for the intrinsic metrics in Table 5.4, extrinsic metrics may give a different optimal value.

Questions:

- When measuring the overlap in tokenization between two corpora (of different languages), Jensen-Shannon divergence is suggested as a better alternative to both the Wasserstein distance and the "absolute number of overlapping tokens". We could also treat the two tokenized corpora as multisets (bags of words) and compute their Jaccard similarity. Is there any advantage of Jensen-Shannon divergence over the multiset Jaccard similarity?
- The Huggingface implementation of Unigram tokenizer is reported to be worse than the original Sentencepiece implementation according to intrinsic metrics in Table 5.1. Does it affect all the languages? Does it affect also tokenizers trained on a single language? Do you have any insights what is the reason? Can you show a sample text tokenized by the two implementations?

Overall, I am very satisfied with the thesis. The author has proven his ability to perform independent scientific work.

I recommend the thesis to be defended.

I nominate the thesis for a special award.

Prague, August 28, 2023

Signature: