Review of the Diploma Thesis

The Faculty of Mathematics and Physics, Charles University

Thesis author	Jiří Balhar		
Title	Improving Subword Tokenization Methods for Multilingual Models		
The year of completion	2023		
Study programe	Informatics	**Study area**	Artificaial Intelligence

Review author	Tomasz Limisiewicz	**Role**	supervisor
Workplace	Institue of Formal and Applied Linguistics		

Text of the review:

Content. The thesis explores the caveats of subword tokenization, the method widely used for input representation of the NLP models. The author aims to identify and evaluate the aspects of subword tokenization that benefit the quality of representation in a multilingual language model.

The main text consists of a background overview and three experimental sections. The former presents a clear and detailed overview of the past works spanning the research on subword tokenization and methods for obtaining balanced multilingual vocabularies.

The first experimental part of the thesis is dedicated to proposing metrics for multilingual vocabulary (namely, allocation and overlap) and empirically evaluating their impact on the model performance. For that purpose, the author trains multiple models based on the same architecture with different parameters of the tokenizer. The models are compared in multilingual language modeling and a set of downstream tasks (NER, POS, NLI, Sentence Retrieval). The chapter concludes that allocation and overlap significantly impact the results.

Secondly, the thesis analyzes the hyperparameter (character coverage, data size, subsampling factor) and implementation choices benefiting tokenization metrics.

The last experimental part focuses on evaluating the methods for obtaining balanced multilingual vocabulary. The author evaluates a range of previously proposed methods, showing that they are in most cases on par with simple subsampling of the data for each language for tokenizer training Therefore, the thesis ends with the strong conclusion that a simple proposed method (strong subsampling) is competitive with the more complex ones. For all the experiments, results are presented for each of the analyzed languages and the average across all languages. The experiments test both in-language performance and cross-lingual transfer.

Limitations. The experiments are performed on relatively small models of the same architecture (as acknowledged in Limitations Section). This choice can be justified by the limited computational resources available for the experiments. Nevertheless, the question remains whether the result would hold for the larger models.

The new method proposed in the thesis is simple and the past research already utilised subsampling in model training. However, best to my knowledge it has not been evaluated in such a systematic and detailed way.

Main strengths. The thesis is generally well written and its structure is easy to follow. The experiments in the thesis are well-designed and comprehensively analyze the impact of tokenization aspects that have not been sufficiently researched before. Each chapter of the work clearly refers to and answers specific research questions.

Furthermore, the author is not afraid to challenge the past established method and shows that comparative performance can be obtained with a simpler approach. In my opinion, these findings would be interesting to the broader research community and should be submitted to a peer-reviewed venue. The part of the findings of the thesis (first experimental chapter) has been published in the paper "Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages" accepted to the Findings of ACL.

Formal and technical aspects. The thesis conforms to all the formal requirements. The graphics and writing are of high quality, with rare stylistic shortcomings not affecting the overall clarity. The thesis is accompanied by a source code with README allowing the reproduction of the experiments.

I recommend the thesis to be accepted.

I recommend the thesis for the best thesis competition.

The thesis presents an in-depth analysis of the research problem, examining various aspects of multilingual subword tokenization. Furthermore, the findings of the thesis are novel and interesting to the broader research community.

Prague on 22. 08. 2023

Signature: