Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis Author	Kirill Semenov
Thesis Title	Pre-processing of the Subword Encoding for the Neural Machine Translation
Submission Year	2024
Study Program	Computer Science
Branch of Study	Language Technologies and Computational Linguistics

Review Author	Abishek Stephen
Role	Opponent
Department	Institute of Formal and Applied Linguistics

Review Text:

Summary

The main goal of this thesis is to address the pre-processing techniques to improve subword tokenization and implement it for Czech-Ukrainian machine translation. The execution of the experiment is based on inline approaches to casing and diacritics in the texts. The motivation for this research is that subword tokenizers trained on general-purpose data show poor performance in tokenizing words with similar semantics but different in orthography due to casing or diacritization. The author systematically compares the extrinsic and intrinsic metrics for subword tokenization. The data used for the experiments comes from Czech and Ukrainian; no other languages were involved.

The text is divided into an introduction, 7 numbered chapters, a conclusion, a bibliography, and an appendix. The total length of the thesis text is 135 pages.

Comments

The first thing that struck me was the length of the thesis. 135 pages in total at first glance invokes excitement as to what this thesis might bring to the table. The excitement comes from the notoriously challenging subword tokenization task and how this thesis aims to address this challenge through specific preprocessing techniques handling the noisy data. The title creates a sense of a diverse multilingual experiment but the author goes on to say "The main experimental contribution of this work is related to developing the Czech-Ukrainian MT system of Charles University" hence this work focuses only on Czech and Ukrainian. The thesis is highly experimental and I appreciate the author's efforts in running such large-scale experiments. The positive and negative results are presented and inferred accordingly. Evaluation measures or metrics have been

very clearly defined. Full points for the experimental setup. Given the experimental nature of this thesis, I would not criticize the decisions for selecting the algorithms. Still, I would like to point out the heavy reliance on Popel et al. (2022) because this thesis tries to do what the authors of Popel et al. (2022) could not investigate due to time constraints as mentioned on pg.20. As I continued reading, the presentation or the stylistics of this thesis made the 135 pages a nightmare to read and comprehend. Most of the sections have details that look very similar to other parts in the text and as a result, I had to go back and forth multiple times. In the end, I was very confused and had to read some sections 2-3 times. I strongly believe that the author could have done a better job in weaving the blocks of this text in a comprehensive manner. Much of the information could have been shown as a diagram which would have aided in better understanding the text.

The theoretical background is covered extensively and cites all the issues in machine translation concerning casing strategies, diacritization, and writing systems across multiple languages. Many solutions adopted for different language pairs are also mentioned but there are no references here to previous related works for Czech or Ukrainian. This gave me the impression that not much has been done for Czech-Ukrainian machine translation tasks but WMT 2022 submissions say otherwise. The author also says on pg. 12 that "Limisiewicz et al. (2023), will be analyzed more thoroughly in terms of the metrics suggested for tokenization evaluation" but I am not sure if that was done because the paper was not cited anywhere else in the thesis.

In the methodology section, on pg.22 the author mentions "The canonical normalizations (NFD, NFC) are mostly lossless", the abbreviations are not clear as to what they stand for. The data normalization part describes two oppositions which to me seems circular and unnecessary. This part could been written in a simpler way only citing which normalization was preferred and why. The preprocessing algorithms have been explained well at least in terms of the core ideas. It would have been better if there was a table or a diagram that shows how the different preprocessing algorithms work here itself but the tables are provided in section 4, pg.62 onwards which is roughly 30 pages ahead of the algorithm descriptions. There also some spelling or syntax errors. On pg.28, the author says "For instance, the base "radi" can be diacritized as "rádi" (meaning "happy [Plural, animate]") and "radí" ("[she/he] gives advice"), which both use the same "háček" diacritic but on different vowels". I think the author meant to say "čárka" instead of "háček". The following line also has the same issue. In Table 5.1, the input sentence should have the diacritized word "komisí" instead of "komisi".

The preprocessing steps or algorithms in general compare the scores on extrinsic or intrinsic metrics based on the different flags used either to mark the type of casing i.e. upper or lower casing or the diacritics. I like these 'flagging' decisions. Very smart indeed. But I would have liked to know the motivations behind some decisions, for example, why does the Char-InDia tag have the morphology KV–idx1–ID–idx2–KV–d1–d2? However, the results did not show tremendous

improvements or scores beating the baselines by huge margins but were slightly better in some cases or on par in many. One other thing that I liked is the use of training data augmentation. But here too no surprising gains were achieved which is completely fine. Overall, it seems like the incasing algorithm is strong in itself and doesn't need too much tweaking.

Out of the three experiments conducted i.e. experiments with casing, diacritization, and romanization strategies, the romanization one only received 4 pages which was one of the goals of this thesis. The author says on pg13. that "The romanization improves performance on the genealogically close languages that use different writing systems, while it does not have a positive effect on the non-related languages." but fails to test this extensively for the selected pair of Slavic languages. I think that this track should have been analyzed in more detail because of its potential to be extended and applied to multiple languages comparatively more than the diacritization track.

The conclusion section addresses a lot of the limitations of this research. I feel some of the limitations could have been investigated. For example, the experiments have shown that the inline casing techniques help considerably with noised texts, it would have been really interesting to see how these preprocessing techniques influence the performance of the Charles translator which was presented as the motivation of this research.

Questions

I would like to ask one question to the author, which if time permits could be answered during the defense. The thesis contains three tracks of experiments. All of these three tracks could possibly be combined in a single setting. For example, the sentence "Tokio bude jediným asijským městem" could be encoded like T_tokio _bude N_jediným N_asijský m N_městem following Incasing (InCa), Word-InDia, and the Roman methods. Why was such a setup not tried at the first place?

The presented thesis is a testimony that the author is capable of devising and running large-scale experiments and hence:

I recommend that the thesis be defended.

I do not nominate the thesis for a special award.

Prague, 31st May 2024

Signature: