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
Study branch	Language Technologies and Computational Linguistics

Review author Mgr. Martin Popel, Ph.D. **Role** Supervisor

Department ÚFAL MFF UK

Review text:

The original goal of the thesis was to explore different ways of preprocessing text before subword tokenization for the purpose of neural machine translation (NMT), so that the NMT models can also handle texts that are all-uppercased (or all-lowercased) or without diacritics. The motivation was to improve the Czech-Ukrainian MT system (Charles translator) developed at ÚFAL. The experiments were thus restricted to this language pair, although the algorithms can be applied to many other languages. There are several previous works dealing with the casing robustness, but the idea of using an approach similar to inline casing also for diacritics is novel, as far as I know.

The present thesis includes several other topics beyond the original goals:

- intrinsic metrics and their correlation with extrinsic metrics
- blind spots of Rényi efficiency
- romanization
- stabilization experiments
- data augmentation by noising

I am impressed by the number of experiments conducted within the thesis and by the thoroughness of their analysis. I appreciate Kirill worked on the thesis continuously, actively studied many related works beyond my suggestions, and implemented many novel approaches and experiments (especially those regarding diacritics). He has proven his ability to perform independent scientific work. We plan to integrate the main outcomes of the thesis into the Charles translator.

The main weakness of the thesis is that most of it was written in haste, it is too long, the structure is not ideal and there are many minor stylistic issues, which could have been easily fixed. While in general, the structure of Theoretical Background, Methodology and Experiments

chapters is reasonable, many pieces of information are unnecessarily repeated this way and one has to jump there and back when reading the thesis.

There are also several errors or inaccuracies in the text. For example, confusing *háček* and *čárka* on page 28, or omitting the soft-sign-like latin-script character in *_aziatsk*ym* in Table 6.4. The description of InCa in Section 2.2.1 says that *at the training step, the sentence initial words are lower-cased*, which would mean increasing the count of lower-cased instances (even for proper names that are never lower-cased). Luckily, it is only a wrong description, the implementation of `train_line` in `inca.py` correctly skips sentence-initial words (unless the `--include_sent_initial` option is used).

Despite the above-mentioned weaknesses, I am satisfied with the thesis.

I recommend the thesis to be defended.
I do not nominate the thesis for a special award.

Prague, June 4, 2024

Signature: