Supervisor’s review of doctoral thesis

Student: Mgr. Jakub Náplava
Department: Institute of Formal and Applied Linguistics
Title: Natural Language Correction With Focus on Czech

Supervisor: RNDr. Milan Straka, Ph.D.

The topic of the doctoral thesis of Jakub Náplava is natural language correction – correction of user-generated texts. Jakub has started working on this topic already in his master thesis, and focused entirely on this area during his whole doctoral studies, authoring 7 peer-reviewed publications about natural language correction. Based on these publications, Jakub submitted a “dissertation by publication”.

The thesis consists of two parts. The first part is a 80-page Preamble that has been newly written, and the second part is composed of the 7 published papers by Jakub. The Preamble contains an introduction, a description of the individual tasks, existing datasets and models, a 55-page summary of the author’s papers, and finally a conclusion. The Preamble can be read without also reading the papers in the second part of the thesis, it is written chronologically, and describes in detail the reasons for working on the individual papers. Personally I find this format very interesting, because it offers a comprehensive yet concise overview of the performed work (with the papers acting as an appendix containing technical details like hyperparameters, detailed architectures and ablation experiments), including detailed motivations and hindsight interpretation not available in the publications themselves. Therefore, I believe there is merit in reading the Preamble in addition to the papers themselves.

The scientific contributions can be divided in three areas.

- **Diacritic restoration** Jakub collected and published a 12-language dataset, including a contemporary state-of-the-art model based on RNN with an external language model. Later, the performance was improved by using a multilingual BERT model, without the need of external language models. Last but not least, Jakub lead an annotation effort to classify the remaining errors of the Czech model, showing that roughly half of them were plausible variants or gold data errors.

- **General grammar error correction** Jakub first implemented a Transformer-based system for a BEA-2019 shared task in English grammar error correction. Based on the shared task results, Jakub incorporated an approach for synthetic data generation and published a system reaching state-of-the-art results for low-resource datasets of German, Russian and Czech. Furthermore, he released two Czech grammar error correction datasets. The second one contains annotations in four different user domains, it is to my best knowledge the largest non-English dataset, and on top of that, Jakub organized manual evaluation of metric correlation, producing the metric most appropriate for this dataset. Finally, Jakub also proposed a non-autoregressive system for morphologically rich languages.
- **Handling user-generated text in downstream applications** In addition to the grammar correction itself, Jakub also devoted effort to the evaluation of several downstream applications on erroneous user-generated texts. He created a system capable of injecting synthetic errors corresponding to given data, and he proposed and compared two mitigation strategies (correcting the input texts with a stand-alone system versus training the downstream application on the texts with errors).

Overall, I consider the thesis to be of very high quality, demonstrating independent and high-quality scientific work on international level. Since the beginning of his doctoral studies, Jakub has been an author of 9 peer-reviewed publications, which according to Google Scholar collected more than 80 citations in total.

I fully recommend the doctoral thesis of Jakub Náplava to be defended.

Prague, 17th June 2022
RNDr. Milan Straka, Ph.D.