Doctoral Thesis Review

Roman Grundkiewicz

June 17, 2022

Thesis: *Natural Language Correction With Focus on Czech*
Author: Jakub Náplava, Mgr.
University: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Supervisor: RNDr. Milan Straka, Ph.D.

Reviewer: Roman Grundkiewicz, Ph.D., Microsoft Research

The reviewed doctoral thesis concerns natural language correction, which is a subfield of natural language processing (NLP), and presents Jakub Náplava's research work and contributions to the tasks of diacritics restoration and grammatical error correction (GEC) with the focus on, but not limited to, the Czech language.

The thesis consists of two parts. The first part begins with a short introduction to the field, where evaluation metrics, data sets, and a brief history of automatic text correction before mid-2017 are presented. It is followed by a description of the author's research conducted between 2017 and 2022. The first presented work describes the development of new state-of-the-art models and data sets for the diacritics restoration task for 12 languages. The next sections are dedicated to work done in the area of grammatical error correction: the development of English GEC systems for the BEA 2019 shared task, adaptations of GEC models to low resource scenarios, and the creation of new error corpora for Czech, including meta-evaluation of metrics and systems utilizing human evaluation. After that, in two separate sections, new BERT-based models are proposed for each of the task. The whole research work is finalized with analysis of robustness to different types of noise in the input data across multiple NLP tasks.

The second part of the thesis includes 7 author's publications in their original forms. The papers complement the previous part providing necessary details and insights to the previously described experiments.

Majority of the included publications were published at the established international workshops, journals and conferences. Jakub Náplava is the main author of 6 papers that formed the basis of this thesis and provided significant contributions to the last one. They have collected more than 65 citations altogether, which is a promising number at an early stage of research career considering recent publication dates of some of them.

The structure of the thesis is appropriate and clear. The chronological order of presentation is logical and easy to follow. The last section on model robustness to the noisy data for multiple NLP tasks summarizes the entire research work and demonstrates its importance and usefulness. The author cites appropriate number of bibliography sources and examines related work where relevant, which both confirm the deep knowledge and orientation in the areas of natural language processing discussed in the thesis. The language is coherent while language inadequacies are rare. A minor remark is that not all examples in Czech provides translation or word-to-word translation into English. Overall, the thesis meets the standards expected from a doctoral dissertation.

The topic of the doctoral thesis is current and relevant in the context of the recent research in natural language correction. Motivations, objectives, methods are clearly described for every stage of the research and satisfy requirements of creative scientific work.

The research outcome provides novel results establishing new state of the art for diacritics restoration and low-resource grammatical error correction across several languages. For both tasks, the author has created especially valuable and publicly available data sets, which established new benchmarks and will facilitate further research. In particular, the GECCC error corpus, considering its size, quality of annotations and domain coverage, is a distinctive resource among available GEC corpora for non-English languages and is a remarkable contribution to the research community.

In most cases the developed methods are compared with appropriate baselines and relevant alternative methods reported in the literature. However, the work on the authentic noise generation and the proposed *KaziText* framework would benefit from more thorough comparison with existing approaches to artificial error generation which aim at resembling human-made error distributions, for instance, with methods utilizing the back-translation technique. Similarly, if one of the main motivations for developing BERT-based GEC models was improving the inference time, it would be interesting to compare them with the teacher-student training utilizing the popular sequence-level knowledge distillation technique applied to the encoder-decoder models, which has been proven to be very effective in boosting the decoding speed, for example, in neural machine translation. The absence of the aforementioned comparisons, however, does not understate the main results of the relevant parts of the thesis.

It can be concluded that the conducted research work of Jakub Náplava brings significant contributions to the field of natural language correction. The thesis clearly demonstrates his ability to understand existing scientific work, design and conduct own research experiments, and critically analyze the obtained results.

In my opinion, the doctoral thesis of Jakub Náplava proves his ability to perform creative scientific work and fulfills requirements to seek a Ph.D. degree in computer science.