PhD Thesis Report

Thesis Title: Natural Language Correction With Focus on Czech
Thesis Author: Mgr. Jakub Náplava
Reviewer: Mgr. Ondřej Dušek, Ph.D.

Thesis Contents Summary

Jakub Náplava's PhD thesis addresses the task of grammatical error correction (GEC), i.e., automatically correcting grammatical errors in texts. The main focus of the work is on Czech, but the author also applies his models to several other languages.

The format of the thesis is a "dissertation by publication", i.e. the text consists of two main parts: a preamble and list of published works. The second part amounts to all the author's papers on the topic of GEC published during his doctoral studies, which are attached verbatim to the text. Since the preamble contains a very detailed summary of all the works conducted as part of the PhD thesis (and the corresponding papers), and since all the attached papers have already undergone peer review, this PhD thesis review will mostly focus on the contents of the preamble as far as writing is concerned.

The preamble consists of four numbered chapters:

1. An introduction, motivating the task of GEC, listing all contributions of the thesis and shortly summarizing the structure of the following text.

2. A background description. This includes a listing of GEC subtasks, common metrics and datasets, a brief summary of previous works on GEC, an overview of commercial GEC applications, and an overview of GEC tools available for Czech. The related works overview ends at the point where the author started working on his thesis. The commercial applications overview also involves a performance comparison of commercial systems with state-of-the-art research systems on the English benchmark CoNLL 2014 dataset.

3. The main contributions chapter. This is a comprehensive summary of all of the author's works, in chronological (or even biographical) order. Seven out of eight subsections in this chapter directly correspond to the published papers included in the second part of the thesis (see details below).

4. A short conclusion, summarizing the contributions again and listing possible future work avenues.

Chapter 3 (i.e. the main contributions chapter) includes the author's GEC-related works that could be roughly categorized into the following four areas:

- *Diacritics restoration*: This line of work focuses on adding diacritical marks (accents) to texts where accents have been omitted, as is common in text messaging or email for certain languages using accented characters. The work starts with a new diacritic restoration approach with recurrent neural networks (RNN) and an external language model (LM), accompanied by a new dataset for 12 languages using non-trivial accents in Sect. 3.1. The author then follows up with a new approach using a pretrained language model (PLM) with non-autoregressive token-level tagging in Sect. 3.6, which also includes a detailed error analysis.

- *Spelling error correction*: This is just a short account of the only unpublished experiment in the thesis (in Sect. 3.2). The author did not reach any tangible results due to difficulties defining the task and quickly reoriented to full GEC.

- *Full GEC*: This strain of work starts with the application of a transformer-based model and the additional use of Wikipedia edits (Sect. 3.3) and synthetic data (Sect. 3.4 and 3.5) for training. This is followed up with a non-autoregressive PLM-based model similar to the diacritic restoration one, in Sect. 3.7. The author also created two new GEC evaluation datasets for Czech (the first one presented in Sect. 3.4 is subsequently reworked and expanded in Sect 3.5). He further carried out a manual evaluation of his systems' outputs, which serves for GEC metric evaluation and detailed error annotation (in Sect. 3.5). Finally, the full GEC work includes an evaluation of commercial GEC tools and comparison to research tools (in Sect 3.5 For Czech, complementing Sect. 2.4 where this is done for English).

- *Downstream task evaluation*: This work, in Sect. 3.8, evaluates various natural language processing (NLP) tools (parsing, named entity recognition, machine translation) under a noisy input condition, and compares retraining the NLP tools on noisy data with applying GEC as preprocessing.

Overall Evaluation

The extent of research work carried out here clearly exceeds the usual requirements for a PhD thesis. The author demonstrated without doubt his ability to conduct novel and independent research. This work pushes the state of the art in GEC and its subtasks (diacritics restoration in particular).

The datasets collected as part of this work (a diacritics restoration benchmark for 12 languages and two Czech GEC sets) will probably have the greatest, longest-lasting impact. This work essentially leads the latest trend in a lot of areas of NLP, where research is now being expanded to languages other than English. The author effectively ensured that Czech is a high-resource language for the task of GEC, with the second-largest amount of annotated data available, trailing only English.

The GEC systems developed here also show some novel approaches and especially a lot of inventive applications and adaptations to new languages. The RNN+LM combination approach to diacritics restoration is novel and was state-of-the-art at the time of publication. The transformer-based models used in Sect. 3.3 through 3.7 are mainly novel applications of existing approaches to new languages. The most valuable part here are the detailed evaluation in Sect. 3.5, including metric evaluation, and further detailed analyses in Sect. 3.6 and 3.8. The downstream evaluation in Sect. 3.8 is novel with respect to the number of tasks compared, and it comes with a really interesting finding (applying GEC as preprocessing is better than training noise-aware models for low-resource NLP tasks). I especially appreciate the author's focus on practical aspects such as implementation speed and the inclusion of commonly used commercial applications in the evaluation.

With regards to the experiments, I see only three minor issues:

- The experiments in Sect. 3.1, 3.3 and 3.4 do not account for randomness in training. They do not involve statistical significance tests or multiple runs with different random initialization. These steps are actually present in the more recent experiments. The thesis thus essentially follows the usual standards in the field at the time when the respective papers were published, but it would have been better to extend these steps to all experiments.

- The comparison with commercial GEC systems in Sect. 2.4 and 3.5 is done on a smaller sample of the data than the evaluation of research systems. The use of a smaller sample is understandable as human input is needed for the commercial tools to apply their proposed corrections. However, this makes the scores not directly comparable to scores measured on the full set. While I do believe that the scores of research systems on the smaller sample will be very similar to the full test set scores, it would have been better to verify this.

- The non-autoregressive approach to diacritics restoration and full GEC in Sect. 3.6 and 3.7 is using an approach of classifying into rules, which is relatively common in morphological inflection generation (see e.g. Bohnet et al., COLING 2010, Dušek & Jurčíček, ACL-SRW 2013 or the SIGMORPHON 2018 Shared Task baseline). This connection is never made in the text. The author might not be aware of this work, but the tasks are very closely related.

While the research described in Jakub Náplava's thesis is of a very high standard, I am not so happy about the presentation form. I am not opposed to the concept of dissertation by publication, on the contrary. The format actually allows for a very efficient review as it includes comprehensive summaries of all the works, and I could refer to the published papers if more detail or clarification was required. I also believe that the author did not spare energy using this format, and that writing a conventional thesis would have taken a similar amount of work.

What I do oppose, however, is the "biographical" narration style chosen by the author. I believe that the main aim of a thesis (or any scientific publication) is to present the results obtained; the process of obtaining the results should be secondary and only mentioned if required to explain experimental decisions. Since the thesis presents results in a chronological order, it makes several otherwise unnecessary topic shifts there and back. The author could have structured the thesis around the tasks, models and datasets – one option would be the order I used to list the contents of Chapter 3 above, for instance. The same problem is apparent in the related works overview:

Related works from the time period after the author started his PhD research are explained along the way (e.g. Sect. 3.3.4, 3.4.4 and 3.5.1) but would be better located in one place.

I found three more general issues with the writing:

- I believe that a thesis should be roughly self-contained, which means that the background chapter would need some expansion. The present text assumes quite a lot of previous knowledge, e.g. about neural network architectures, training and decoding techniques.
- Some sections of the preamble lack citations, even for claims that are not obvious to a reader outside of the GEC field (e.g. all of Chapter 1, GEC subtask definitions in Chapter 2 introduction, Sect. 3.8 introduction mentioning "several studies [..] on assessing performance drop"). While these claims are plausible and partially covered in the published works part, the current state is not ideal.
- The thesis is written in understandable English, but, ironically, it contains a non-trivial number of grammatical errors. Most of these do not hinder understanding, but they do make the text harder to read (see detailed remarks below).

Recommendation

Based on the evaluation above, my recommendation is that the thesis be **approved** for a PhD. My reservations are relatively minor and do not challenge the thesis as a whole.

Detailed Comments

Regarding Contents:

Chapter 1

- A more explicit listing of research questions or sub-goals would have been nice here.

Chapter 2

- The split of various listings in Chapter 2 into the period before and after the author started his PhD work may be informative in some regards, but should not be so prominent.
- Sect. 2.3 does not explain all the systems included in the comparison in Sect. 2.4.
- Regarding the commercial system evaluation in Sect. 2.4: The fact that the same type of evaluation for English and Czech is located in very different places in the thesis is a bit strange. In addition, this evaluation is a novel contribution of the thesis and should be announced as such.
- The differences between English and Czech with respect to application of statistical machine translation approaches could be discussed in Sect. 2.5.

Chapter 3

- I do not see any point in discussing seq2seq models in Sect. 3.1.1.
- I am not sure if Sect. 3.2 is even needed as a standalone section since there is no tangible output. It could have been just a short note somewhere else.
- It would be good to report the ablation of using authentic data only for finetuning in Sect. 3.4.2.
- Table 3.14 should clearly mark that it only depicts development and test sets.
- It would have been interesting to see intra-annotator agreement as well in Table 3.16 (i.e. the agreement of the same annotator with themselves when doing the same annotation a second time).
- The sudden introduction of the KaziText tool in Sect 3.5.3 is confusing and would need more detail. This is just another problem of the chronology approach.

- The information about GECCC human judgment annotation in Sect. 3.5.4 is shortened in a misleading way.
- The notes on named entities and Korektor behavior in Sect. 3.5.5 are confusing.
- Remarks on BERT in Sect. 3.6 are somewhat inaccurate. BERT's training is self-supervised, not truly unsupervised (it is a standard supervised setup, just using naturally-occurring data instead of specific annotation). The label "BERT" typically only refers to English models, so stating "multilingual BERT" outright would have been better. The citation of Vaswani et al. in Sect. 3.6.1 cannot be connected to BERT – this work predates BERT by ca. 1.5 years.
- The analysis at the end of Sect. 3.6.4 is vague and confusing. Based on reading it, I do not know if there are any clear trends. In addition, "As can be seen" implicitly refers to a table that is not included in the preamble.
- The training tweaks of Omelianchuk et al. mentioned in Sect. 3.7.2 are unclear and would need more explanation.
- Regarding a remark in Sect. 3.8.2: I believe that BERT is not state-of-the-art for the GLUE benchmark anymore, and has not been for some time.

Chapter 4

- The reference to "a neural network called BERT" is inaccurate.

Regarding Writing:

General comments:

- Some Czech examples in the text are left untranslated (e.g. on page 65).
- It would have been better to add hyperlinks to each paper individually instead of just stating "The full paper is in Chapter 5". This makes it harder to refer to the particular paper for details.
- While the reference list is formatted to a very good standard, it misses capitalization in names that BibTeX lowercases by default (e.g. in Alfaifi & Atwell, Brocket et al., Dale & Kilgariff). Some references are duplicated (e.g. Chollampatt et al., Cotet et al., Napoles et al. 2017).
- I am not sure if this is a commonly used term in GEC, but I do find the expression "second learner" strange. I would prefer "second-language learner", "L2 learner" or potentially "ESL learner" (for English as a second language).

English Grammar: All problems listed below appear repeatedly in the text; I only include a few examples for each type. The text is still fully understandable, but it is harder to read due to these errors.

- Word order is probably the most frequent problem. The writing often splits verbal groups, sometimes with very long phrases: "to already be outperformed", "was on the model side attributed" (page 33, 34). Focused clauses at the start of a sentence are much more common than usual in English.
- Missing articles or incorrect article use is also very common: "from W&I+L development set" (page 33), "loss by term that assigns [..]" (page 34), "we follow an approach of Lichtarge et al." (page 34).
- In some cases, punctuation follows Czech instead of English rules: "decide, whether" (page 34), "recall, that" (page 36).
- Vague coreference: English is more tolerant to repetition than Czech, so many pronouns would better be expressed in full to make the reference clearer: "Shared Task [..]. One of its main objectives" (page 33).
- Coordination is used very leniently and does not always follow usual syntactic conventions: "to reduce variance during training and hopefully also achieve better results" (page 34), "model can be quickly and using a relatively small amount of supervised data finetuned" (page 60).

- Modal verbs are sometimes used incorrectly: "These [..] should have contained" (page 41), meaning "These likely contained".
- The word "decent" is used frequently, but its meaning is vague and sometimes the use does not seem appropriate: "more decent base configuration", "learner corpora are rather decent and local" (both page 45), "despite being only a test set of decent size" (page 46).
- The phrase "apart for" or "apart to" is used multiple times (e.g., page 44-45), meaning "in addition to".
- Overuse of participles to start sentences, some of them are dangling participles: "Having the model pretrained, they [..]" (page 45), "Having the need for better [..], it was obvious" (page 47).
- Incorrect use of "although that" or "despite that" – this combination is not correct English (page 47).
- Use of "however" mid-sentence, with no separators (page 70).

Questions

I have a few questions for the author, some of which could be discussed during the defense:

- Is using just one metric per dataset a common practice in GEC (as you imply in Sect. 2.2 and 3.5)? Why not use multiple metrics?
- For the dataset in Sect. 3.1, did you check if people commonly write in other languages without diacritics, as they do in Czech (in text messages etc.)?
- Do you think the metric combination you mention in Sect. 3.5.1 would generalize to other datasets, languages or different GEC systems?
- Regarding error classes in Sect. 3.5.2, how do you handle combinations of multiple errors in one spot?
- What exactly do you average in GECCC human annotation when IAA is measured in Sect. 3.5.4? Why do you not measure metrics' correlation on the sentence level as well?
- Would it be possible to predict each diacritization operation individually in your system from Sect. 3.6 (i.e. use multiple binary classifiers instead of a single softmax)?
- Are any other works apart from your own evaluated on your diacritization test sets?
- Was there a reason to omit English results in Sect. 3.7?
- What is your view on the use of BART or T5 PLMs for GEC? These models are pretrained on a denoising task, which is related to GEC (essentially very simple GEC synthetic data). Are you aware of any works using these models, or did you even try any of them out yourself?

Prague, 16 June 2022

Mgr. Ondřej Dušek, Ph.D.
Institute of Formal and Applied Linguistics
Charles University, Faculty of Mathematics and Physics
V Holešovičkách 747/2, 18000 Praha 8