Master Thesis Review

Bc. Petr Pechman
Czech Grammar Error Correction

Faculty of Mathematics and Physics, Charles University
Institute of Formal and Applied Linguistics

Supervisor: RNDr. Milan Straka, Ph.D.
Study programme: Computer Science
Study branch: Artificial Intelligence

Reviewer: Ing. Alexandr Rosen, Ph.D.
Affiliation: Faculty of Arts, Charles University, Institute of the Czech National Corpus
Role: Opponent

Following up on earlier reserch, the thesis describes a Grammatical Error Correction (GEC) system for Czech using a Neural Machine Translation approach based on the Transformer architecture. Like some of his predecessors (Náplava & Straka 2019, Náplava et al. 2022), the author addresses the challenge of limited error-annotated data by generating noisy sentences from error-free input, based on probability distributions of formal error types. The resulting synthetic data are modified by incorporating several error types "typical" for Czech. This step represents a novel contribution by the author with significant improvements over the previous setup, without the language-specific additions.

The noisy data are generated in real time, enabling easy adjustments to data generation parameters, which turns up to be useful in the experiments, with pre-training and fine tuning as their main components. Here the reviewer would prefer a clearer distinction, especially about *grammatical corrections* and the addition of language-specific errors: during the pre-training stage *the model learns general language patterns, syntactic structures, semantics, and mainly grammatical corrections. This stage corresponds to the situation where we have no annotated data but want to build a model for grammar correction.* The fine-tuning stage *aims to improve the model in grammar corrections further. We perform it on annotated data in the Czech language.*

The thesis found that better results are achieved when (i) the addition of language-specific errors in pre-training is combined with simple language-independent rules using Aspell and MorphoDiTa, (ii) fine-tuning is performed only on manually annotated data rather than on a mix including also synthetic data, and (iii) in fine-tuning, the share of the smaller GECCC dataset domains is increased by 25%, (iv) the pre-training stage is longer, (v) models are trained on higher-quality corpora, (v) larger models are used. To produce such results, the author designed and implemented a sophisticated model-training pipeline, compiled a list of some common errors in Czech, and checked how clean (error-free) the corpora used in training actually are.

Formally, the thesis is well-structured and clearly written, without any disturbing faults, with a proper introductory and background sections. The following comments and questions do not compromise its overall positive evaluation, based on the extensive and appropriately designed experiments, driven by the substantial and confidently answered research questions.

General comments:

At least according to the reviewer's non-native intuition, the author's command of academic English seems well up to the task, except for occasional stylistic mishaps and handling of articles (or lack thereof).

Czech examples should be typed and glossed according to a standard common in linguistic literature as in *Přišel ke mně*. 'He came to me.'

Morphological cases should be named rather than numbered (*dative* for *third case*).

Specific comments and questions:

p. 6, example of the M2 format: deserves a gloss

p. 13: *We apply one operation (deletion, substitution, swapping with a neighbor, and inserting a random word) for every chosen word with a given probability.* > *They apply...*?

p. 16: about the size of error-annotated data needed to train the sentence2sentence Transformer model: *Unfortunately, there is a lack of such data for Czech, or rather it is a problem for almost every language.* – **How much data would be enough?**

p. 17, fig. 2.1: Output sentence seems to be missing.

p. 18: About Aspell: *For example, it can generate errors in subject-verb agreement, verb tense, or morphology (see example on Figure 2.2).* – **Isn't it more like generating errors resulting in another existing form within the morphological paradigm of the same lexeme?** This might be called morphological or even morphemic errors in inflectional.

p. 18, fig. 2.3: The generated forms seem to include all inflectional variants of all lexemes derived from the same lexical base, however unlikely. **There could be many, some very unlikely, isn't that a problem?**

p. 19: Czech Typical Errors > Typical Czech Errors

p. 21: *Here, we also have errors…* – reference unclear: where?

p. 21: Errors in conditionals: **Did you also include the error type *by jste* > *byste*, *by jsme* > *bychom*?**

p. 21: **Numerals: also *dvěmi*, *oběmi*, *dlouhými* (*rukami*)?**

p. 22: *We are coming up with a pipeline implementation…* >*We propose …*

p. 36: *The third is fine-tuning Second Learners (on the development data), which is quite surprising given the size of the training data (30 812), the error rate, and the representation size in the development and test data.* – This may (also) be due to the heterogenious content in terms of proficiency level and L1.

p. 51: *Moje oblíbená kavárna je Louvre Café, **který** je blízko stanice Národní třída* – the "incorrect" correction (missed antecedent–pronoun agreement *kavárna – který*) should be glossed and commented.

I recommend the thesis for defense.

I am not nominating the work for a special award.

31 May 2024

Alexandr Rosen