Opponent’s review of PhD thesis proposal

PhD student: Jakub Náplava
Title: Natural Language Correction

Content

The thesis proposal is focused on the task of grammatical error correction. The task is well motivated, selected literature is surveyed, novel techniques are outlined along several directions, and mature experimental achievements are already reported. The text is clearly structured and is easy to follow. Yet I outline several points that, in my opinion, could be improved or extended in the future text of the dissertation.

Comments to the literature survey:

- SMT is blamed to be unable to “generalize beyond patterns seen during training” (in contrast to NMT), but I find this claim too general unless “pattern” is specified in more detail; in some sense, NMT also searches for patterns, just that their nature and granularity is different.
- I find the following claim misleading (3.3) “GEC in English is a long studied problem... there has been only a limited research on error correction of other languages”, followed by a (seemingly exhaustive) list of only a few references. “Limited” is a vague word, but still, there are numerous languages for which experiments with grammatical error detection/correction were published, sometimes more than two decades ago. For Czech, this includes e.g. the systems of Karel Oliva or Jan Hric going back to mid 90’s; sequence of publications on an error corpus of Karel Pala and his colleagues is relevant too, starting probably with *Pala, Karel, Pavel Rychlý, and Pavel Smrž. "Text corpus with errors." International Conference on Text, Speech and Dialogue, 2003.*
- More attention should be paid to the quality of bibtex entries:
  - Double braces should be used around titles in bibtex entries in order to avoid undesired lowercasing (“non-native english”, “african languages”, “Correcting esl errors using phrasal smt techniques”, and many others)
  - A wrong bibtex item: Tatranské Matliare is perhaps not the right coauthor of Alexander Rosen.
  - An incomplete bibtex item: Jakub Náplava. 2017. Natural language correction.
  - An incomplete bibtex item: Kevin Knight and Ishwar Chander. Automated postediting of documents.
  - Swapped first name and surname of the first author in Ondřej et al., 2017.
- In the current text, literature review segments are interleaved with text pieces describing novel contributions. In my opinion, in the future dissertation text there should be ideally a single clear cut between the “old” and the “new”. If they remain mixed, then then novel contributions should be clearly distinguished throughout the text by some other explicit means. In addition, the risk of any auto-plagiarms suspicion should be avoided since the topic of 3.1 seems to overlap considerably with the author’s former work in his diploma thesis.

Language quality

- The text reads well, however, it should undergo proofreading if some parts are to be published further. Above all, there are tens of missing determiners.
- Some other language imperfections:
  - “since than” → “since then”
  - “perform satisfactory” → “satisfactorily”

    - “the performance our model” → “the performance of our model”
    - “Brockett et al propose to consider natural language processing to be a machine translation problem” → perhaps you meant “natural language correction”?
    - “multiple of them must be combined” → “more of them …”
    - “bad spelling” → “wrong spelling”
- Several expressions should be written rather with a hyphen, such as “hand coded rules”, “word based approach”, “character level network”, “low resource machine translation”, “black box generation”.

Questions selected for the exam (could you please prepare a brief reaction?):

- Surprisingly, the quality of contemporary off-the-shelf commercial grammar-correction solutions (such as grammarly) is not discussed at all. Why don’t you use some of these systems at least as baseline? Are they known to be well below the state of the art, or is there some other reason.
- Ad “… none of them having comparable performance to systems on English” – are the absolute values actually interpretable across different languages? (let me speculate: you need more words if a same meaning is to be conveyed e.g. from Czech to English, and if the absolute number of errors remains the same, then you get a lower percentage of tokens that must be changed in the English document than in its Czech counterpart).
- Users’ needs are reflected in using $F_{0.5}$ instead of F_1. Would it be useful to present also different levels of severity of detected/corrected errors to the end users? Would it make sense to include it into evaluation measures?

Other comments and questions:

- I am not sure whether it’s relevant, but there’s also a body of literature focused on correcting errors in *automatically* produced texts, especially in the context of machine translation (such as *Rosa, Rudolf, David Mareček, and Ondřej Dušek. "DEPFIX: A system for automatic correction of Czech MT outputs." Proceedings of the Seventh Workshop on Statistical Machine Translation. 2012)* and speech recognition/reconstruction (such as *Fitzgerald, Erin, and Frederick Jelinek. "Linguistic Resources for Reconstructing Spontaneous Speech Text." LREC. 2008.).*
- I don’t understand why UD is used as the data source when approximating diacritization complexity in various languages. Why you don’t use some other multilingual dataset with much more data per language (given that you don’t use any UD-specific information at all)? I mean I like the introduced dictionary baseline, but I think that the results might be biased (it’s hard to guess how much) by the range of UD corpora sizes: the bigger the corpus, the more ambiguity you get, and the sizes of the UD corpora vary in the order of magnitude.
- If there are any annotator manpower reserves at your disposal, I would be really interested in seeing the inter-annotator agreement on grammatical correction, which is unfortunately not discussed at all in the text. Thus we have no idea how far we can get from the current level of performance. In general, I think that it’s always worth to sacrifice some portion of annotators’ time for parallel annotations to estimate the agreement. At the same time, I believe that every presented performance result should be accompanied with a reasonably strong baseline.
- Ad “Although this is only a short survey, we hope that it covers … before the shared task in 2019”. This sentence is irrelevant for the thesis proposal submitted in spring 2020.
- It seems that entities that undergo corrections are only words. What about punctuation marks? For example in Czech there are specific rules e.g. for quotes, for decimal points and commas, or for commas in complex sentences etc.

- Would it make sense only to include detected but unresolved errors in the evaluation too (for example is situations in which edit distance to any other correct solution is high)?

Conlusions:

In my opinion, the author has gained a deep insight into the task, he explores it really systematically, the presented experimental achievements in grammatical error correction are already impressive and the future work outline is realistic.

In Prague, June 10, 2020

doc. Ing. Zdeněk Žabokrtský, Ph.D.