PhD Thesis Proposal Review

Thesis Title: Are We Losing Textual Diversity to Natural Language Processing?
Thesis author: Josef Jon
Review author: Ondřej Dušek
Date: 30 May 2024

Contents summary: Josef Jon's PhD thesis proposal is motivated by the currently very topical issue of language change prompted by the proliferation of machine-written texts, which is most apparent on the use of large language models (LLMs) and their propensity for using specific, previously rare expressions, which results in an overall increase of use of these expressions in, among others, scientific literature. The author, however, focuses his research mostly on the area of machine translation (MT), which receives less attention, but has had a lot of influence on texts published on the internet for a considerably longer time than LLMs. MT outputs are supposedly less diverse than human-produced texts, and MT seems to struggle most with very original and creative types of texts, such as fiction works or even poetry. Josef Jon's thesis aims to measure this difference in diversity (supposed diversity loss), e.g., in terms of measure of word surprisal and its distribution over the course of a text. The thesis further aims to identify potential causes of the diversity loss, then address them by adapting the MT models' decoding approaches and training objectives. The final aim is to apply the newly developed approaches in a real-world scenario.

The author has already published some work on the topic at multiple conferences (ACL 2023, WMT 2022-2023, LREC 2024), where he focuses on an alternative decoding approach making use of minimum Bayes risk decoding and genetic algorithms. He applies the approach both to a standard MT scenario (within the WMT shared task) and to a specific scenario, where he produces adversarial examples for automatic MT evaluation metrics.

The proposal text is structured into five sections. Section 1 represents the introduction, which is mostly dedicated to the main motivation for the research. Section 2 then describes problems considered in the proposal: measuring textual diversity, the apparent loss of diversity in MT outputs and its potential causes, as well as potential remedies. In the end, this leads to the formulation of the thesis' main objectives. Section 3 focuses on related works and discusses specifically some of the approaches considered for the thesis or even already applied in published work. It also provides additional motivation to the research in showing previous works addressing the loss of diversity in MT and problems of decoding algorithms and training objectives. Section 4 then describes the author's work so far, addressing each of the objectives one by one. Section 5 provides a brief overview of the future research plan.

Overall evaluation: I consider the topic and framing of the thesis really novel. It appears non-mainstream, yet very current, and it addresses a very important problem. The objectives of the thesis are formulated very clearly and even if only partially fulfilled, they will represent an extremely valuable contribution to our field. In fact, I believe that a full thesis could be written on every single one of them. Even though the first two objectives do not seem to successfully confirm their underlying hypotheses in the research conducted so far, the remaining objectives of alternative approaches to MT are valuable on their own.

The proposal shows very solid past work of the author, published at reputable venues. The future works plan is laid out reasonably, and I have no doubt the future experiments will produce publications of at least the same quality.

The focus on MT in the thesis seemed a little forced at first, but the proposal convinced me that MT is a good testbed for this type of research. The focus on literary or creative translations is definitely well chosen, as these are the area of translation where MT still does not match humans. Nevertheless, I would

still really like to see an expansion of this research into the general area of LLM output – be it as part of Josef Jon's thesis, or as a follow-up to it.

The proposal text is written in excellent English with little to no issues. It is slightly repetitive, but not so much to pose a problem; in fact, the repetition increases ease of reading. Sections 2 and 3 experience some skips that could potentially have been avoided by merging both sections. On the other hand, I really liked the structure of Sections 4 and 5, where the text is clearly aligned with the thesis' objectives.

Overall, I really enjoyed reading the proposal. I believe its quality is excellent, and it has all the ingredients to become a successful PhD thesis in due time.

Questions: I have several questions, which could be discussed during the exam, or could be potentially addressed in the author's future work:

- Regarding the linguistic levels to be addressed (Figure 2), have you considered going below the word level, into e.g. rhymes and rhythmic patterns of poetry? I believe that is one of the toughest problems for current MT.
- Do you have any alternative hypotheses on the surprisal uniformity and/or loss of diversity – how to measure them and what, besides the model decoding, could be the cause?
- Are you considering direct preference optimization (or its variants) as an approach to segment-level training?
- Do you think that changing the architecture of the models itself could also help, in addition to changing decoding and training?
- Regarding your example in Section 4.1 with more and less uniform surprisal: It feels like the less uniform text is just overall more complex. Could you somehow compensate for that and show an example with a same surprisal on average, yet more evenly distributed?
- Could you provide more explanation on how the SLOR measure works and what Figure 5 signifies?
- Regarding Figure 6 and the corresponding describing text: It seems to me that the text does not entirely match – it looks like the *books* domain behaves the same as *wilde*, and the trends you describe are only there for *wmt* and *global* – could you please explain that?
- Would it make sense to evaluate the MT quality and/or surprisal with humans instead of automatic metrics? Do you think it would change anything? Do you think that using COMET specifically might bias your results in any way?
- Do you think there would be any difference in the results with respect to different language pairs and their properties (such as morphological richness)?
- Your genetic algorithm presented in Section 4.3 seems a little crude – is the crossover point fixed in terms of position for both sentences? If so, does this not break the translation in a vast majority of cases? Also, how do you ensure that "the final translation always scores better or at least as well as the best translation from the initial candidates"?
- In your reference to BLEURT in Section 4.3, do you do you mean words or tokens?
- When you combine multiple metrics for the adversarial example generation approach, would it make sense to use dynamic combination (i.e. change the weights randomly)?

Small remarks: These are mainly meant as additional writing feedback for the author:

- It would look great if Figure 4 could show the underlying words.
- The equations in Section 4.1 are a bit underused in the text and hard to parse for the reader. I believe they would be better left out of the text on this level.
- The ORT dataset could use a citation or more explanation.
- I'm not sure what you mean in Section 4.4 by the paragraph stating: "Traditional loss functions [..] struggle [..]"