Review of Thesis Proposal

Faculty of Mathematics and Physics, Charles University

Review author: Martin Popel, ÚFAL MFF CUNI

Date: September 4, 2023

Candidate: Dávid Javorský

Thesis title: Global and Local Constraints for Neural Models of NLP

Supervisor: Ondřej Bojar, ÚFAL MFF CUNI

The goal of the thesis according to the proposal is to *"design a new method which allows to transfer the knowledge between models trained on tasks of different nature"* by using (user-defined) *constraints on the output*. The method is introduced and described in very general terms, perhaps aiming at *"providing as unified picture as possible"* as specified in the Guidelines for thesis preparation in SIS. Luckily, several concrete examples of applications are given. All of them are types of machine translation (MT).[1]

The proposal focuses on two applications in Section 4: shortening MT and gender bias in MT. Unfortunately, no experiments in the latter task have been conducted and I have doubts how the approach specified in Section 4.2 could work in practice. It mentions an example sentence where the gender of word *doctor* can be inferred from the sentence itself and there are no reports showing how frequently modern MT systems make errors (caused by gender bias) in this type of sentences. Similarly, it is unclear what would be the approach (the constraints and the constraining task) for other applications mentioned in the proposal, e.g. MT for low-resource languages or multi-modal MT. Two types of constraints are mentioned – global and local, but only the global ones are planned to be explored in the experiments and it is not clear how would the selected approach with importance scores work with local constraints.

Thus, the proposal actually focuses only on a single application – shortening MT, which is a well motivated task (needed in subtitling and dubbing) and could be complex enough for a whole PhD thesis topic, especially if considering all the trickiness of real-world needs. We should distinguish between applications where only the total length of the whole translated document needs to be decreased and possibly the language simplified (so a standard summarization tools can be used) and applications where the units of text to be shortened are much smaller (one subtitle) and the goal is to keep the translation length similar to the source length, which sometimes means making

[1] The proposal mentions speech translation, but non-text input is not mentioned, so I expect a standard text-to-text MT using 1-best ASR predictions as inputs is planned to be used. The specification in SIS mentions also "summarization, dialogue systems or their components" and "co-reference of pronouns or term translation choice which needs to be consistent across the whole document", but these applications are not mentioned in the proposal.

the translations longer.[2] Even if the subtitling and dubbing applications operate on smaller units of text, a global consistency (within the whole movie) is needed, so e.g. the translation is not referring to a fact that was omitted in previous translations (which naturally leads to a local-constraint task).

The proposal is well structured and appropriate in formal aspects. It shows notable effort in related work overview. Section 2.2 tries to classify transfer learning into meta learning, continual learning, multi-task learning and ensemble learning. Admittedly, the "descriptions of these concepts in the literature are quite confusing as they are closely related and overlap"[3] but the provided classification is even blurrier, I think. For example, it does not mention/reflect that in transfer learning, only the final performance on the target task is important, while the performance on the source task(s) is irrelevant, unlike in multitask learning, where the performance in all tasks is relevant.

A notable effort has been invested also into the experiments in Section 5 and their analysis. I appreciate these experiments were published at ACL 2023.[4] Sections 5.1.1–5.1.3 show that the word importance scores have intuitive properties and all these experiments are well analyzed and described. However, I don't see any hints that these scores may be useful for the final task of shortening MT.

The Future Work section of the proposal starts with *"First, we anticipate finishing the task of shortening in NMT."* Unfortunately, the experiments done do not involve any actual shortening of translations yet. I don't even see any of the promised baselines replicated and evaluated. I think it would be wise to evaluate the whole task as soon as possible.

Section 3.3 mentions several planned approaches how to use the word importance scores for the final task. Interestingly, it is missing the possibility to first translate and then shorten the translations.

Section 3 starts with an example sentence *'Two elderly women having a conversation with their children'*. This is actually a very nice example for showing problems in the approach of cascade shortening MT where words with scores below a given threshold are deleted from the source-language text. If we delete words *elderly* and *their*, there is no way to reconstruct these two pieces of information in the translation. However, there are target languages where the information can be preserved without making the translation longer.[5]

[2] `https://iwslt.org/2022/isometric` evaluates *the percentage of translations in a given test set falling in a predefined length threshold of ±10% of the number of characters in the source sentence.*

[3] Citing `https://www.cs.uic.edu/~liub/IJCAI15-tutorial.html`

[4] `https://aclanthology.org/2023.findings-acl.563.pdf`

[5] There are one-word translations of *an elderly/old woman* in many languages. There are languages with different words for *a child* as a young person (not an adult yet) and *a child* as an offspring (son or daughter) of someone. Moreover if the children of the elderly women are already adults, only the latter translation is correct.

Sections 5.2 and 5.3 report on an experiment with downscaling word embeddings *"of tokens that we aim to alter"*, but it is not mentioned that the experiments focus on scaling (or zeroing) all[6] words with a given POS tag. Section 5 is introduced with *"Second, having the importance scores, our goal is to use these scores to make MT models aware of the importance of the tokens on their input. For this, we examine scaling word embeddings because we suppose that they encode meaning. We show results of such experiments in Section 5.2."* However, it seems the importance scores (as defined in Section 3.2) were not used at all in Sections 5.2 and 5.3. More importantly, it is not clear what is the relation to the original goal of Shortening MT. In this experiment, all model parameters are frozen and some are modified (by scaling or zeroing some word embeddings). This leads to the decoder working with inputs (encoder outputs) it has never seen during training, which usually leads to catastrophic errors such as generating repeated tokens, in my experience. So I was not surprised to read this was the case also in this experiment.

Questions:

- What was the goal of the experiment with downscaling word embeddings in Sections 5.2 and 5.3?
- *Regular machine translation systems* are promised for baselines in the gender bias MT task. Do you plan to use datasets where the gender cannot be inferred from the single sentence? Do you plan to use document-level MT (e.g. trained on whole paragraphs) as baselines as well? Do you plan to consider also 2nd person gender (in dialogues)? Do you plan to consider the usecase of non-English translation pairs pivoted via English, where both the source and target language express gender in similar ways, but English does not?
- For the shortening MT task, *pretrained models for sentence summarization (Zhang et al., 2019; Rothe et al., 2020)* are promised as baselines and *SARI (Xu et al., 2016), Rouge-N or Rouge-L (Lin, 2004)* as the evaluation metrics. Does this mean that the task will constrain only the total length of the translation and the Length Compliance metric defined in the Isometric Spoken Language Translation task[7] would not be usable? Do you plan to include the self-learning based approach of Lakew et al. (2022)[8] into the baselines?

Prague, September 4, 2023

Martin Popel

[6] It is not clear if all words of a given POS are scaled/zeroed or just a single random word in each sentence. Section 5.3 says *"Specifically, zeroing out one word in the source leads to an edit distance of 3–4 on average."* so the latter is possible. What has been done with sentences without that POS, in that case?

[7] `https://iwslt.org/2022/isometric`

[8] `https://arxiv.org/pdf/2112.08682.pdf`