Thesis proposal review

Ivana Kvapilíková: Towards Machine Translation from Monolingual Texts

The thesis proposal of Ivana Kvapilíková belongs to the field of machine translation and focuses on methods which do not require parallel data for training. This is a very active research area since for many language pairs manually translated parallel texts do not exists at all, or their amounts are very limited and for traditional supervised methods insufficient. Searching for a solution that would provide translation between such languages is indeed very useful, but also challenging.

The proposal is written in English, spanning 16 pages in total (including a rich bibliography). The text is structured into 6 sections. After the introduction, the author reviews related work on unsupervised MT, multilingual MT, unsupervised cross-lingual embeddings, and parallel corpus mining. All these topics are highly relevant to the proposal. Section 3 then presents methodology the author uses (or aims to use) in her experiments which includes various language modelling techniques, methods for cross-lingual embeddings, unsupervised SMT, unsupervised NMT and techniques such as back translation and denoising auto-encoding. This section also presents evaluation methods for tasks the proposal focuses on (machine translation, word translation, parallel corpus mining, and parallel sentence matching). In Section 4, the author presents her experiments that were already conducted and (some of them) published. Section 5 then presents plans for future work following two research directions: the first focusing on inducing multilingual vector representations and the second focusing directly on machine translation for low-resource language pairs. Section 6 contains conclusions.

The text is very well written and readable without disturbing errors and typos. The review of related work and overview of the most relevant concepts and methods is rich and complete (the bibliography includes more than 50 referred papers). The technical parts provide a good view on the problems and allows good understanding of the author's approach and plans. Each idea is well motivated and a proper methodology proposed. The only thing the author does not really discuss is the selection of languages she would like to experiment with. Of course, the proposed methods are expected to be language-agnostic and one can pretend that any language pair is low-resourced by ignoring the data available. However, are there any plans for experiments with real low-resourced languages? Are there some existing test sets for such language pairs?

Conclusion

The thesis proposal of Ivana Kvapilíková is a nice piece of work, some of the experiments were already published at international venues. I consider this research topic and the research plans sufficient and well formulated for the author's dissertation.

RNDr. Pavel Pecina, Ph.D. In Prague June 3, 2020