dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Dr. Cristina España i Bonet
Senior Researcher, Team Lead
DFKI GmbH
- Saarland Informatics Campus -
Campus D3 2, Stuhlsatzenhausweg 3
66123 Saarbrücken
Telefon: +49 (0)681 85775-5354
E-Mail: cristinae@dfki.de

22nd January, 2024

Review for Ivana Kvapilíková's doctoral thesis, with title “Towards Machine Translation Based on Monolingual Texts.”

Summary and Relevance

This thesis deals with a hot topic in the field of natural language processing: unsupervised machine translation (UMT). In an era where machine translation has achieved (quasi-)human performance when huge amounts of parallel data are available, language pairs with low resources lag behind. A way to overcome the problem, and make machine translation accessible to more languages, is to get rid of the need for parallel data and be able to use only monolingual texts for the task. This is the goal of UMT.

Ivana’s work focuses on the improvement of existing UMT methods mainly, but not exclusively, with new approaches to data collection and processing. The main contribution of the thesis is a method to extract pseudo-parallel data from monolingual corpora using multilingual embeddings (Chapter 5). A second contribution affects UMT architecture with modifications to the standard pre-training strategies (Chapter 6). Finally, after the observation of some of the translation errors made by UPBMT systems, several heuristics are devised to process the data in a manner that can help minimise the errors present in the subsequent steps (Chapter 7).

Document and Content

The thesis is well written (with few typos) and easy to follow with a structure, length and content that are adequate for a PhD dissertation. The main content of the report is distributed as follows:

Chapter 1: Introduction. This a short chapter that motivates the research question of the thesis and introduces the following chapters.

Chapter 2: Background. The text defines first the kinds of data used in machine translation (monolingual, parallel, comparable, pseudo-parallel, synthetic parallel, lexicons and pre-trained models). Especially nice to read is the definition of low-resource language and the description of the settings considered in thesis. The work explores several language families, amount of data, and domains —the latter with minimal experiments.

Chapter 3: NLP Fundamentals. Similarly to the previous chapter, this is a descriptive chapter that introduces basic models and architectures relevant for the work. This includes cross-lingual embeddings, pre-training of (multilingual) language models and supervised statistical and neural machine translation.

Chapter 4: Related Work. The final introductory block defines a taxonomy of UMT divided in model-centric and data-centric approaches. The chapter describes related work for both of them.

Chapter 5: Parallel Corpus Mining. This is the first experimental contribution: the creation of pseudo-parallel corpora by extracting comparable sentences in monolingual data. The novelty implies using UMT to create a first synthetic parallel dataset to better align cross-lingually a multilingual LLM. With the alignment-improved LLM, one can generate sentence representations for the monolingual texts and extract the (pseudo-parallel) pairs as the most similar sentences. The method is evaluated in the "parallel corpus mining" and "corpus deshuffling" tasks. Results do not lie far from the supervised comparison, are more robust to a domain change than other models, and the analysis layer-wise is interesting. Also relevant is the observation that it is enough to align the LM with respect to one language pair to obtain improvements in general.

Chapter 6: Unsupervised Machine Translation Methodology. Extension of Chapter 3 where unsupervised statistical (UPBMT) and neural MT (UNMT) are introduced. The architectures used in the second part of the thesis are presented, including some modifications to the standard approaches.

Chapter 7: Experiments and Results. This is the second experimental contribution: the analysis and improvement of standard UMT approaches. It is the most extensive part of the thesis. First, experiments with UPBMT are shown. These are used to detect systematic errors in the translations and postprocess them in a way that these errors can be avoided in the next steps. Second, the data generated by the UPBMT system is used to initialise a UNMT possibly followed by a fine-tuning step. Different pre-training strategies are also presented. Finally, the data created in Chapter 5 is used for fine-tuning different modalities of the systems. These techniques are applied in medium and low-resource settings. As expected, the more the languages differ among them and the further the domain in both languages is, the lower the final translation quality. This observation is common to all UMT systems but, in most of the cases, the contributions of the thesis beat the standard models.

Chapter 8: Discussion. This chapter summarises the main findings and challenges of the work. The thesis ends with the conclusions and directions for future work.

Comments

Ivana has conducted thorough research on unsupervised machine translation. Her contributions are sound and improve over the seminal works by Lample et al. and Artetxe et al., providing interesting insights to the machine translation community. The work has been published in

two refereed conferences and several shared tasks. Looking at the bibliography, I see that a couple of references are simultaneous or previous to the date of Ivana's bachelor thesis, so it is not completely clear to me which contributions belong where.

The document is well written and I enjoyed reading it. The state of the art and background are comprehensive and definitely useful for newcomers to the field. Experiments are well designed and allow to see the relevance of each component in the final unsupervised systems. Something that is not explored in depth is the effect of the domain, not only the mismatch between languages in training but also in test. Given the importance of data in the thesis I think this would have been a nice addition. This does not diminish the quality of the work which, in my opinion, fulfills the requirements for a PhD thesis.

Dr. Cristina España i Bonet
Senior Researcher, MT Team Lead
Multilinguality and Language Technology
DFKI GmbH

Saarbrücken, 22nd January 2024