Thesis Review by the Supervisor

Reviewer & supervisor: doc. RNDr. Ondřej Bojar, Ph.D.; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, 181 00

Date: 5. 2. 2024

Thesis Title: Towards Machine Translation Based on Monolingual Texts

Candidate: Mgr. Ivana Kvapilíková

The doctoral thesis submitted by Ivana Kvapilíková focuses on the challenging topic of unsupervised machine translation, i.e. methods of training translation models without providing them with texts previously translated by humans. Such a goal seemed elusive and non-reachable until first approaches by the teams of Guillaume Lample and Mikel Artetxe were published in 2018. Ivana enrolled into her PhD study that year and we were lucky to secure the opportunity for her to spend two months of a research visit with Mikel Artetxe, Gorka Labaka and Eneko Agirre, as a kick-start of her studies.

After a joint paper with the colleagues, which was published in 2020, Ivana continued the research on her own and I am very delighted to see the very good results of her work today.

In her doctoral study, Ivana has covered a broad range of approaches to unsupervised learning of translation which I could even call "full" or "encompassing", if I was not aware of the fact that research in our discipline expectably brings unexpected future developments.

After a brief introduction in Chapter 1, Chapters 2 (Background) and 3 (NLP Fundamentals) provide an excellent motivation from the real world and compactly summarize the basics of the techniques Ivana builds upon. Chapter 4 (Related Work) provides a nice and well-organized survey and a taxonomy of unsupervised MT, and I really wish Ivana will have an opportunity to publish this type survey as an independent paper in the near future.
The scientific contributions of Ivana are described in Chapters 5 to 7. Chapter 5 focuses on finding parallel sentences in a pair of collections of monolingual texts. With such a pseudo-parallel corpus as an additional resource, Chapter 6 provides a nice and clear overview of possible methodologies that Ivana explored: creating cross-lingual word embeddings, creating unsupervised phrase-based and finally neural machine translation systems. Chapter 7 is a detailed account of six experimental protocols, published in Ivana's papers throughout her studies. The discussion in Chapter 8 highlights the lessons learnt: observations on successes and failures across the experiments and the forward-looking remaining challenges.

The thesis itself is excellently written, starting from the overall structure to the internal structuring of individual sections, from the very high standard of English to excellent presentation in pictures and tables.

Despite the fact that Ivana started her PhD study about five years ago, I would like to point out that the relevance of the topic has not faded. As Ivana documented, when we want to

apply unsupervised MT in the true real-world conditions of low-resource languages with less common scripts, rather particular domain or genre data and difficult access to native speakers or language informants, the work is far from finished. Importantly, this insufficient performance in such conditions applies also to the more recent multi-lingual large language models (LLMs) which might bring the expectation of fully solving unsupervised machine translation; consider the result presented in Table 7.10 for ChatGPT. Furthermore, the broad experimenting and careful analysis of Ivana has provided us with a good understanding of what and why can be achieved in unsupervised learning, and as such this understanding will be very helpful when exploring and advancing the more opaque LLMs.

A few remarks are also due describing my collaboration with Ivana throughout her studies. Here I can only express my full content. Ivana was very independent, self-managed, focussed and technically proficient. I enjoyed all our meetings, regardless if we were discussing the plans, checking the progress, discussing the good results or the bad ones where we had to search for justifications and reasons for method failures. In short, it was only a pleasurable experience for me to supervise Ivana's work and -- as it actually should happen with PhD students -- Ivana has definitely been more of a research colleague to me than a student.

In sum, I consider the doctoral thesis by Ivana Kvapilíková as excellent work, documenting Ivana's research proficiency and expertise. The thesis represents well the scientific contributions Ivana has made in the area of unsupervised machine translation. I fully recommend accepting the thesis.

In Prague, February 5, 2024.

Ondřej Bojar