Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Meisyarah Dwiastuti
Název práce	Indonesian-English Neural Machine Translation
Rok odevzdání	2019
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Mgr. Martin Popel, Ph.D. **Role** vedoucí
Pracoviště	ÚFAL MFF UK

Text posudku:

The first goal of the present thesis is to develop a state-of-the-art Indonesian-English (and also English-Indonesian) neural machine translation system. The second goal is to study the interactions between domain adaptation (by fine-tuning to the TED talks "domain") and back-translation (a method for improving the translation quality by leveraging large target-language monolingual data).

The first goal was fully accomplished. As demonstrated especially in Chapter 1 (Background and Related Work), the author has gained a good knowledge of Neural Machine Translation and related areas, including recent advances. She prepared training and test data in a methodologically sound way and trained very strong baseline systems. This involved for example a thorough tuning of beam search hyperparameters in Section 3.2, where I appreciate that Table 3.2 shows also the time demands and the effect of a deduplication post-processing script. Such insights may be useful for other researchers when building NMT systems for other language pairs.

The baseline system produces clearly higher quality translations than any previous work on Indonesian-English MT I am aware of, although this was expected because the previous systems used smaller training data (and thus they are not comparable and no such comparison is reported in the thesis). The final system (with domain-adaptation and backtranslation) achieved a significant improvement over the baseline on the target domain. Surprisingly, it is even insignificantly better than Google Translate, although Google Translate is not constrained to the publicly available training data.

The second goal was accomplished as well. While both the back-translation and domain adaptation using fine-tuning are simple and well-known techniques, their combination as presented in the thesis has not been studied yet, as far as I know. Chapter 4 explores all combinations of four back-translation training regimes, four fine-tuning regimes and two averaging variants, i.e.

32 systems in total. The novel 4-way-concat training regime combines elegantly back-translation (balancing synthetic and authentic training data) and domain adaptation (balancing in-domain and out-of-domain training data). It could remove the need for explicit fine-tuning and possibly result in higher quality due to synergic effects. The present results unfortunately do not confirm this hypothesis, but I agree with the author's conclusion that the selected size of in-domain blocks is still too small (even when three times upsampled). Based on the fine-tuning results in Figure 4.9, it seems probable that upsampling the in-domain blocks e.g. to the size of the out-of-domain blocks would result in significant improvements (at least similar to the fine-tuning).

I appreciate the in-depth exploration and analyses including learning curves with BLEU and training loss, evaluated on both in-domain and out-of-domain test sets. This research could be helful for virtually any language pair and target domain.

The thesis is well organized, written in English with a minimum of typos. There are few grammar/style errors and oversights, e.g. an inaccurate formula for sentence length penalty on page 9 – it should be $(\frac{5+|Y|}{6})^{-\alpha}$.

I acknowledge that part of the thesis was accepted as a paper submitted to the ACL Student Research Workshop [1].

I recommend the thesis to be defended.

Reference

[1] Meisyarah Dwiastuti. English-Indonesian neural machine translation for spoken language domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 309–314, Florence, Italy, July 2019. Association for Computational Linguistics.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Seattlu dne 25. 8. 2019

Podpis: