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. Michal Novák, Ph.D. **Role** Oponent
Pracoviště	ÚFAL

Text posudku:

Thesis description:

The goal of the presented thesis is to build a neural machine translation (NMT) system between Indonesian and English. Moreover, it aims at exploring possibilities for tackling both domain mismatch and lack of resources for a language at the same time. To this end the author performs experiments with two methods: fine-tuning of models and incorporating synthetic data produced by backtranslation of monolingual corpora.

The thesis is structured as follows. After introducing the topic and the goals of the thesis, Chapter 1 informs a reader about fundamentals of NMT, principles of the Transformer model as well as relevant work related to the domain adaptation in NMT. Chapter 2 presents the experimental setup including the datasets and modeling details. Chapters 3 and 4 are devoted to experiments on the development data, namely to fine-tuning and to backtranslation, respectively. In Chapter 5, the models proven to perform best in the experiments are then evaluated on the test set and compared with the Google Translate system in both translation directions. Finally, the Conclusion chapter summarizes the main findings and suggests the future work. The thesis consists of 55 pages including Bibliography and Lists of Figures, Tables and Abbreviations. A CD containing training scripts is attached.

Comments:

The thesis is written in a good level of scientific English and the text is easy to comprehend. However, the number of grammatical and lexical mistakes could have been smaller (I can supply a list of them if needed). It is also the case of minor citation errors, where the command `\citet{}` should be used instead of `\citep{}` and vice versa. Some of my objections are a matter of style and personal preference. For instance, I would avoid using digits for small numerals (e.g. "3 abstractions", "2 objectives" and "1 epoch"). I would also be cautious about omitting the conjunction "that", especially in the cases where it may lead to a garden-path reading of the sentence (e.g. "... which indicates adding synthetic data to the training set is beneficial ...").
Tables and figures are formatted nicely and together with their captions are self-explanatory. Captions in the lists of tables and figure at the end of the thesis could have been shortened, though. In some of the graphs of learning curves (e.g. Figures 4.3, 4.9), the duration axis is
given in a number of hours whereas in the others a number of training steps is used instead (e.g. Figures 3.1, 4.4). Consistent use of one of the units or both of them (as seen in Figure 1.3 adopted from [Popel, 2018]) would facilitate aligning the information plotted in graphs with other sources of information. For instance, whereas Table 4.3 shows that the non-tuned 4CONCAT system is outperformed by non-tuned CONCAT on the in-domain data, 4CONCAT seems to be the best-performing system according to Figure 4.2. The problem is that the end of the learning curve in Figure 4.2 is trimmed, so three of the four systems reach their maximum outside the graph. However, one could not prove it without comparison with Figure 4.9 later in the thesis, because the best-performing

systems are specified by a training step in Table 4.3 while training duration is specified by time in Figure 4.2.

The thesis is well-structured, clearly separating author's own work from the rest of the thesis. Chapter 1 provides the reader with all the concepts of NMT and the Transformer model and
approaches to domain adaptation that she/he needs to know for understanding the experimental part. As I had not known too much about domain adaptation in NMT before, the summary of works related to this topic was particularly beneficial to me. The only thing I missed here was a section or a paragraph about technical details specific to the Tensor2tensor library. Since I have no experience with this library, I got confused in the experimental part due to a special definition of batch size in Tensor2tensor. Throughout the thesis, batch sizes are specified only by a number with no unit. I personally expected that the batch size is defined as it is usual for NLP tasks - as a number of sentences. Getting evidence from some tables and figures that contradicted my expectation, I soon started having doubts about correct definitions of other terms, e.g. data block and epoch. Everything has finally been cleared
up after I found out that Tensor2tensor uses a batching machinery producing batches of a variable length in terms of sentences. The unit used to specify the batch size is in fact the number of tokens. My misinterpretation could have been avoided if the Tensor2tensor batching mechanism was explained or at least units were given whenever batch size was mentioned.

As for the author's own work, it fully satisfies what has been promised in the specification. All experiments were designed and performed correctly. Results of the experiments are in line with the expectations and findings from related works. The only exceptions are two observations on training in the CONCAT regime. First, behavior of the learning curve in transitions between the data blocks is opposite to what Popel [2018] shows in his experiments.
Second, unlike to Popel's work, the author gains no benefit from checkpoint averaging here. Although the author hypothesizes about reasons behind the discrepancies, they should be explored deeper in some future experiments. Nevertheless, I understand that the author rather concentrated on a great deal of other experiments in her thesis.

I appreciate that throughout the whole thesis the author is sharply focused on her objectives related to domain adaptation. For example, in the final evaluation in Chapter 5, the whole text commenting its results is devoted to evaluation on the TED domain. However, according to Tables 5.1 and 5.2 the best improvement over Google Translate was achieved on movie subtitles (test_B dataset). Although I know it is not in the center of presented research, it would still deserve a sentence or two.

A few of my comments and questions to some particular details of the thesis follow:

- It may be again a feature of Tensor2tensor, but it seems to me that development evaluation and checkpoint storing were run asynchronously with the training loop. Although it may be useful for efficiency, it hinders conducting some experiments more precisely. For example, exploring the performance on in-domain data blocks and transitions between them and the out-of-domain blocks in the 4CONCAT regime requires tripling the data blocks by upsampling and shortening the checkpoint period to 15 minutes. However, I suspect there is still not so high chance for the evaluation checkpoint to hit an in-domain data block. Would not it be better to run evaluation and the training loop synchronously in order to have a better control on timing of evaluation checkpoints?
- A post-processing script to fix word repetitions is introduced on page 20. I missed clearer justification for using the script. Are the word repetitions a well-known issue for Transformer-based translation? Has the author really found a close-to-optimal solution if it makes errors that can be reduced by such simple script? Is the script run only to find an optimal setting of beam size and length penalty or in all experiments? On page 40, the author mentions that Indonesian plural is formed by reduplication of words (nouns?). At the same time, she observes that her system and Google Translate fail to translate such plural forms. Is not it just a consequence of the post-processing script?

- In the regular expression on page 19 (‘`grep -Ev "https+:" | ...`'), should not the '+' sign be replaced with '?' to cover also "http:"?
- On page 32, the author says: "... we hypothesize if the systems (AUTHENTIC and SHUFFLED) are trained longer, the difference between their performance will be more notable". In my view, the claim is in contradiction with the previous sentence: "While the improvement (of SHUFFLED over AUTHENTIC) in early training is around 1.25 BLEU point, it is reduced to as small as about 0.3 point after 60 hours of training." It is also in contradiction with the learning curve in Figure 4.3, where the gap between the two systems is narrowing as the training goes on. Did I misinterpret anything?
- On page 36, the author found no obviously distinct behavior among the performance of the SHUFFLED, CONCAT and 4CONCAT systems fine-tuned by the same scenario. However, according to Figure 4.9, performance of SHUFFLED+FINE-S and SHUFFLED+FINE-AS has a rising tendency even towards the end of the fine-tuning depicted in the graph. Does the author think that performance of the SHUFFLED system would remain the same if she fine-tuned it for a longer time?
- What is the reasoning behind using different batch sizes in different experiments? Are they supported by an experiment that has not been mentioned in the thesis or some well-known recommendations?

Conclusion:

Achieving very promising results, the thesis is a nice contribution to machine translation of low-resourced languages in a specific domain. Despite the shortcomings it clearly shows that the author is capable of conducting scientific research on her own. The work satisfies the requirements set on master thesis. I thus recommend that it is accepted for the defense. I will be also happy to see the experiments extended to some other low-resourced languages and published as a scientific paper.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 2.9.2019 **Podpis**