Master thesis review

Faculty of Mathematics and Physics, Charles University

Thesis author	Iryna Tryhubyshyn
Thesis title	Mutual Relation of Machine Translation and Quality Estimation
Submission year	2022
Study program	Computer Science **Study branch** Artificial Intelligence

Review author	Mgr. Martin Popel, Ph.D. **Role** Reviewer
Department	ÚFAL MFF UK

Review text:

The topic of the thesis is to study the relationship between machine translation (MT) systems of different quality and sentence-level quality estimation (QE) systems on the English-to-Czech translation direction. The main research questions are how the performance of QE (measured as Pearson correlation with "gold targets") depends on:

1. quality of the MT systems used for training QE,
2. quality of the MT systems whose quality is being estimated,
3. capacity (number of parameters) and architecture (BiRNN vs PredEst) of the QE model,
4. QE training data size,
5. domain of the test set (in relation to the domain of the QE and MT training data).

The author explored this 5-dimensional space, which resulted in an impressive number of experiments. This is related with both the main strength and the main weakness of the thesis: Exploring all the interactions between individual dimensions is a very ambitious goal with great practical impact. Presenting such results in an easy to understand and well-arranged way is very difficult and there is a large space for improvement in this aspect.

The text is clearly structured into 4 chapters, written in English. The are some errors in English grammar and very rare typos (e.g. "1m" instead of "10m" on the last line of page 31), but in general the text is easy to read.

Chapters 1 and 2 give an overview of MT and QE, respectively, and the related work. Despite minor inaccuracies, the text shows the author is familiar with these topics. I appreciate the review of related work includes also author's critical opinions supported by facts (e.g. *We think this*

evidence is rather weak since... in Section 2.3.5). Chapter 2 mentions three types of sentence-level QE: detection of critical errors, regression estimating direct assessment (DA) and regression estimating post-editing scores measured by HTER. It does not discuss estimation of post-editing time nor different purposes of QE: to select the optimal MT system for a given text vs. to decide whether to use a given pre-selected MT system for a given text (and post-edit it or discard the machine translation and hire translators to translate it from scratch).

Chapter 3 describes data sets and tools. The author trained two MT systems (called 1m and 10m according to the number of training sentence pairs) in the Marian toolkit. She also prepared several QE training data sets. While Chapter 2 mentions only HTER (and DA) as the way of assigning "gold target" labels to the sentences in QE training and test data, Chapter 3 uses TER for this purpose. This is justified by the expensiveness of (human) post-editing of selected MT outputs, which is needed in HTER. However, the implications of the change from HTER (or DA) to TER are not discussed.

Chapter 4 describes an impressive number of experiments. Figures 4.1–4.6 summarize the most important results of the thesis. On one hand, I appreciate the compact visualization summarizing multiple dimensions of the experiments. On the other hand, it took me a lot of time to understand the visualization and it was very difficult to follow the whole discussion.

I admit the author tried to describe it clearly (and some of the description in the main text were indeed helpful), and it would be very difficult to make the visualizations and discussion substantially easier to understand without sacrificing some of the details (complexity of the 5-dimensional results). I have just a few minor suggestions for possible improvements:

- *Data quality* should be renamed to *MT Quality*. *QE traininig data quality* should rather mean the amount of noise in the QE training data, i.e. whether the target labels reflect reliably the real translation quality (or the post-editing effort). For example, TER scores (used in this thesis) seem to be of lower quality than HTER or post-edit time or DA because a given sentence may have multiple fully correct translations which are very different according to TER. However, this aspect (quality of the QE training data) has not been explored in the thesis.

- *Data amount* should be renamed to *QE data amount* or *QE data size*.

- *Model capacity* should be renamed to *QE model capacity* (to prevent confusion with *MT model capacity*, which is a dimension that was not explored in this thesis).

- It is very confusing that names "*1m*" and "*10m*" are being used for three different dimensions: MTs used for test sets, MTs used for training and the QE training data size. The MTs used

for training could be renamed e.g. to *weak* and *strong*. The names *1m* and *10m* could be kept only as names of the QE training data size.

- Each row in Figures 4.1–4.6 corresponds to one MT system used for translating the test set. While this dimension is interesting and shows some interesting insights, in most figures each row shows similar phenomena (trends). I would suggest to primarily report results on a test set translated with *all* the MT systems and only secondarily report exceptions from these general trends. This is partially done with the *Common test set*, which includes translations by the 1m and 10m MTs (but not LINDAT and Google). The main motivation is to make the figures easier to interpret by having a single row for each figure. This way, all the figures could be places on one page, so that comparison across test sets (CzEng, WMT18, IWSLT) are possible.

The last item is related to the expected purpose of QE (cf. my review of Chapter 2 above). If the purpose is to select the optimal MT system from a given set for a particular sentence or document, it makes sense to measure the QE performance on all these MT systems.

Questions for the defense:

- Do you expect any of the reported results would change when using HTER-based on DA-based QE (instead of the TER-based)? Of course, there is no time to do such experiments (using some of the existing HTER-based QE data sets) before the thesis defense; so I am mostly interested in a discussion of possible differences (and thus limitations of the selected approach).
- Do the experiments in Section 4.1 use Bi-RNN or PredEst?

Detailed comments:

- According to Section 3.1.1, the main *CzEng test set* is a subset of the CzEng 2.0 training data (czeng20-train.gz): "*we use around 20 million sentence pairs from the whole dataset ...we further divide into train, validation, and test.*" According to Section 3.3, "*We translate the CzEng test set by Google Translate and LINDAT Translation*". However, LINDAT Translation was trained on the whole CzEng 2.0 training data, i.e. including the test set used in this thesis. Evaluating a system on a subset of its training data without explicitly acknowledging this is a methodological flaw. A possible solution would be to extract the 5k validation and 10k test sentences from the 492k sentences in czeng20-test.gz instead from czeng20-train.gz.

- "*the probability that the model assigns to some words does not correspond to the actual frequency at which this word will appear in the valid translations.*"
  A reference is missing for this claim, which seems to be misleading. NMT models assign *conditional* probability to each generated word (conditioned by the whole source sentence and the previous target words), which is not comparable to a frequency of that word in the reference translation of the whole test set.

- "*According to Table 3.4, the SAO dataset is the most diverse in the group.*" How is the diversity of a data set defined in this thesis? The claim that "The WMT18 dataset is more diverse than CzEng." was based on a comparison of distinct words and total words. Type-token ratio (TTR) is indeed one of many possible ways to evaluate (lexical) diversity. However, according to Table 3.4, Antrecorp has higher TTR than SAO for both English and Czech:

  ```
  English SAO       TTR = 1897 / 13158 = 0.14
  English Antrecorp TTR = 1532 /  7893 = 0.19
  Czech SAO       TTR = 3011 / 10818 = 0.28
  Czech Antrecorp TTR = 2104 /  6505 = 0.32
  ```

- "*We should note that the WMT18 MTs are different but do not have lower power than the rest.*" I would say it is the other way round. The MT models evaluated on WMT18, CzEng and IWSLT test sets are exactly the same (1m, 10m, LINDAT and Google). The MT power was defined as MT quality as evaluated by DA, BLEU etc., so according to this definition, the models have lower power on WMT18 because it is a more challenging test set.

Overall, I am satisfied with the thesis: the goal was ambitious and it was partially achieved. The author has proven her ability to perform independent scientific work.

I recommend the thesis to be defended.

I do not nominate the thesis for a special award.

Prague, 31 August 2022

Signature: