Reproducibility Proposal

Chahyon Ku, Daniel Cheng, Sherry Zhao, Shubhkarman Singh

1. Minimum Viable Plan

Our initial goal is to train the model for the English-German data and reproduce the performance metrics on WMT 2019 sentence-level test set. To do so, we would have to adjust each components of their models to run on our computers. Some examples of these components are: the multilingual BERT model, the baseline model from previous works (SyntehticQE), the transformers library, and the custom tokenizer. Algorithmically speaking, Monte Carlo Dropout and Whole Word Masking (WWM) seem challenging.

By the end of week 3, we would like to get the pre-trained multilingual BERT model on our computers and evaluate data. Then, on week 4, we would proceed to work on training the paper's method (basically fine-tuning the mBERT model) on the word-level en-de data from WMT 2016. Then, on week 5, we would train the model on other datasets (WMT 2017, 2018, etc.) for word-level en-de data. Then, on week 6, we will train and evaluate on sentence-level data as well as English-Russian data. Over this process, we would evaluate the model on a single dataset, WMT 2019.

Once we've reproduced the results for the single-model setting, we will move on to the training of ensemble models. Since ensemble model shouldn't need more training on top of the previous model, we will spend one week to reproduce the results.

2. Stretch Goals

Validate the performance of this model on other Quality Estimation datasets (tentative: En-De dataset from WMT20). `https://www.statmt.org/wmt20/quality-estimation-task.html`.

Train a custom model to post-edit and revise on the word that we tag as BAD from the QE model. Validate the performance of this model on ground truth En-De translation datasets.

3. Bibtex Citation of Paper We're Reproducing

Zheng et al [1]
`https://aclanthology.org/2021.emnlp-main.267/`

4. Original Hypotheses We're Reproducing

The paper's self-supervised method (masked language model and machine translation model with transformer encoder) outperforms previous unsupervised methods (SyntheticQE, which trains on synthetic data and suffers from biased errors).

5. Accessing Data in the Paper

The paper uses data from Conference on Machine Translation (WMT) and the open parallel corpus (OPUS). While they provide a python script for downloading all relevant datasets, these can also be found on their corresponding websites:

- https://www.statmt.org/wmt22/
- https://opus.nlpl.eu/

6. Using the Original Code

We will be using their existing code at:
`https://github.com/THUNLP-MT/SelfSupervisedQE`

We'll modify this code to run custom experiments as necessary.

7. Feasability of Computation

The main computation cost will come from training the model in this paper, which is just fine-tuning mBERT, which is relatively cheap and should be very computationally feasible.

8. References

[1] Zheng, Y., Tan, Z., Zhang, M., Maimaiti, M., Luan, H., Sun, M., Liu, Q., and Liu, Y. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic, Nov. 2021), Association for Computational Linguistics, pp. 3322–3334.