Proposal: Almost Black-Box MT Model Stealing with Reference Translations Supersampling

Vilém Zouhar, Philipp Zimmermann

1. Problem Statement

We plan to work on stealing a neural-based machine translation model via translation querying and comparison to reference translations and teacher score.

2. Motivation

Model stealing is a practical concern for production MT systems [4]. Therefore we wish to document another approach which strengthens this claim. The method could also be used for distillation in general.

3. Proposed Strategy

After querying then comparing these translations to references and supersampling the train dataset based on (1) distance to the reference translation and (2) decoder confidence. The closer it is to the reference or the higher confidence it has, the more supersampled it is going to be.

4. Related work

The authors of [1] presented student models that distil knowledge from a larger teacher model without loss in BLEU performance. These models are significantly faster and do not require a GPU for inference. Their way of manipulating the queried data was based on target sentence quality *without reference and teacher score*.

5. Existing code/software

We wish to use MarianNMT [2] because we have access to pre-trained NMT models which can be used as teachers. One of us also has practical experience with this framework and knows how to, e.g. extract sentence confidence.

6. Implementation

Reference and inferred data loading and processing, sentence similarity metrics, sentence supersampling based on heuristics from sentence similarity and confidence. Training of multiple models and their evaluation.

7. Evaluation - metrics

The standard metrics of MT quality is the BLEU score. Even though it is widely known to not be the best metric [5], it is widely used and sufficient.

8. Evaluation - datasets

Given the pretrained model availability, we will be working with either English-Czech or English-German language pairs. Both are part of the Europarl corpus [3]. It is a relatively small corpus, but given our computational and time limitations, it will have to suffice.

9. Evaluation - baselines

We plan to compare the results against (1) training only on references (2) training only on translated data (3) training on the mixture of the two.

10. Success criteria

We hope that one of the metrics will result in better BLEU scores. This is, however, hard to predict. If that does not happen (negative results are also a good finding), then we will at least have a comparison between the three baselines.

11. Team

Team Mittwoch with the two members Vilém Zouhar (vilem.zouhar@gmail.com) and Philipp Zimmermann (s8phzimm@stud.uni-saarland.de).

References

[1] Ulrich Germann, Roman Grundkiewicz, Martin Popel, Radina Dobreva, Nikolay Bogoychev, and Kenneth Heafield. Speed-optimized, compact student models that distill knowledge from a larger teacher model: the uedin-cuni submission to the wmt 2020 news translation task. pages 190–195. University of Edinburgh and Charles University, 2020. 1

[2] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann,

Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. 1

[3] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005. 1

[4] Eric Wallace, Mitchell Stern, and Dawn Song. Imitation attacks and defenses for black-box machine translation systems. *arXiv preprint arXiv:2004.15015*, 2020. 1

[5] Vilém Zouhar, Tereza Vojtechová, and Ondrej Bojar. Wmt20 document-level markable error exploration. *Submitted to WMT2020*, 2020. 1