Machine Translation Course Shared Task: Team Toneko [1]

Tõnis Ojandu and Eneko Ruiz de Loizaga

UNIVERSITY OF TARTU 1632

GOAL

Goal of this project was to improve neural machine translation system. Target translation was from Estonian to English. Intent was to use previously established methodologies to train a baseline system, to observe the shortcomings of this system and thereafter propose and implement improvements to this system.

PROJECT STEPS

- Pre-processing training corpus of Estonian - English language pairs
- Splitting training corpus into training, dev and testing sets
- Training the baseline neural machine translation system using the established corpus
- Machine translating an sentence set using the system trained including the postprocessing of the result
- Evaluating the translation against accurate human translation
- Proposing improvements
- Implementing improvements to the baseline system that involves modifying pre-processing and postprocessing
- Retraining the neural translation system with said improvements
- Retranslating an sentence set with the improved translation system
- Evaluating the resulting translation against accurate human translation

IMPROVEMENT IDEA

Main idea of the improvements was to improve grammar of the translations generated by the established neural machine translation system.

To this extent the goal was to use the method used by R. Sennrich and B. Haddow for WMT16 [2].

The idea was to annotate the tokens in pre-processed sentences with Part-Of-Sentence (POS) tags for both source and target languages and later remove them in postprocessing step.

Since the baseline system also included splitting the tokens using the Byte Pair Encoding (BPE) model that was also trained on the initial corpus, then this tagging would also involve duplicating the tags to BPE splits.

The aim of this tagging was to create a stronger association between the words (and also in some cases between BPE splits that had not joined in the in the trained model) with their counterparts in the opposing language. Hope was that this would help to differentiate token-wise equivalent n-grams that had different grammatical meaning in the source text.

TOOLS

OpenNMT python implementation [3] was used as the underlying framework for the whole project.

For POS tagging the English sentences Standford CoreNLP toolkit [4] was to be used.

For POS tagging the Estonian sentences Estonian natural language toolkit [5] was to be used

RESULTS

Baseline neural translation system was successfully trained by splitting 19051439 language pairs into 25000 large dev set, 50000 large test set and with remaining as the training set.

Baseline system trained 13 OpenNMT epochs and resulted in BLEU value of 35.97 points on the test set.

Due to technical difficulties the annotating was not implemented in time for the course end.

REFERENCE

[1] Prject repository: https://github.com/mt2017-tartu-shared-task/nmt-system-F

[2] Linguistic Input Features Improve Neural Machine Translation. / Sennrich, Rico; Haddow, Barry. Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers. Association for Computational Linguistics (ACL), 2016. p. 83-91.

[3] OpenNMT python implementation: https://github.com/OpenNMT/OpenNMT-py

[4] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

[5] Estonian Natural Language Toolkit: https://estnltk.github.io/estnltk/1.2/index.html