Capstone Proposal
Transfer learning for machine translation quality estimation

Lucia Specia and Paco Guzmán

Motivation: Machine translation quality estimation (QE) is the task of predicting the quality of a machine translated segment (generally a sentence) without access to a gold-standard (reference) translation. This is an important field of research especially given recent advances in machine translation: for medium to high resource languages, the level of fluency of translations tends to be very high, making translation seemingly high quality, while they may still contain meaning preservation errors. This makes it hard for models and even humans to identify mistakes in translations. Quality prediction models are built from a set of labelled instances for a given language pair and information from both the source and translated segments. This information varies from features such as language model scores for these sentences (Specia et al., 2015), to more complex word representations learnt with neural models (Kim et al., 2017; Ive et al., 2018; Kepler et al., 2019a).

Limitation in current approaches: As with most supervised machine learning problems, an important bottleneck in QE is the need for labelled data. Most existing datasets predict a continuous score, e.g. in [0,100] and suffer from skewed distributions towards the high end of the quality spectrum (i.e. most translations are good). To alleviate the need for labelled data, state-of-the-art models for QE rely on pre-trained representations, such as those provided by BERT, XLM, or Laser (Kepler et al., 2019b). However, these models still require at least a few thousand instances for fine-tuning. This means that for each language pair (and possibly text domain), a new labelled dataset has to be collected to build language & domain-specific models. In this project the aim is to investigate ways to build QE models where labelled data only exists for other languages.

Task definition: The task is to predict an adequacy-oriented quality score in [0,100] (the so-called Direct Assessment, or DA score), where 0 indicates a completely incorrect and disfluent translation, and 100 indicates a perfect translation for two language pairs, English-German and English-Chinese. This could be done in various ways, but ideally you should leverage the labels for the other languages and/or NMT model information.

Datasets: The following open-source resources are available:

- Three sets of 7K training instances and 1K dev sets for different languages: Estonian-English, Romanian-English, English-Nepali annotated with z-normalised DA scores. The score to predict is in the 'z_mean' column.
- Two sets of 1K development instances for English-German and English-Chinese annotated with z-normalised DA scores. The score to predict is in the 'z_mean' column.
- Log probabilities for each translated word as given by the NMT models used to produce the translations in all the languages (the models themselves can also be made available) - part of the tar.gz files above.

- Two sets of 1K test instances for English-German and English-Chinese, where you will need to predict the z-normalised score for each instance (to be provided later).

Details of the content of the files are given at the end of this document.

Baseline and state-of-the-art performance: The baseline will be a bidirectional recurrent neural net (RNN) model that encodes pre-trained word embeddings (multilingual BERT) for both source and target sentences and predicts DA scores as output. This will be based on the Predictor-Estimator architecture in the open source OpenKiwi tool (Kepler et al., 2019: https://github.com/Unbabel/OpenKiwi/blob/master/kiwi/models/predictor_estimator.py), but where the Predictor vectors are replaced by multilingual BERT vectors (https://github.com/google-research/bert/blob/master/multilingual.md). This Estimator will be trained on the dev sentences only. This model was used by the winning submission of the WMT19 shared task on Quality Estimation (http://www.statmt.org/wmt19/pdf/54/WMT06.pdf). Since the data to be used is new, we have not yet established state of the art performance on this data.

Evaluation will be performed using Pearson Correlation and Root Mean Squared Error between the predictions and the gold-labels on a test set of 1,000 instances. Success in this task will indicate that QE models can be built for a wide range of languages without the need for labelled data in all languages. The test set will be kept blind with the evaluation run through Codalab.

Possible directions include various flavors of transfer learning from the label data in other languages, including fine-tuning on the dev set, multi-task learning, upsampling, etc.; use of better pre-trained representations; ways to automatically collect labelled data for the two languages.

Teams and workload: teams of 2-3 people would be ideal for this project. We estimate a minimum of 20-30 hours for a submission to receive a distinction-level grade.

Submission format: test set predictions submitted to Codalab (at most 100 submissions per language pair), 4 page report explaining the models used, any data collected, etc.

Additional resources: any additional corpus, pre-trained models and other resources, as well as existing software libraries can be used for the project. This will require use of GPUs (Google colab could be used).

References

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In Proceedings of the Second Conference on Machine Translation (WMT), pages 562–568.

Julia Ive, Frédéric Blain and Lucia Specia. 2018. DeepQuest: a framework for neural-based Quality Estimation. In Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera and André F. T. Martins. 2019a. OpenKiwi: An Open Source Framework for Quality Estimation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics - System Demonstrations, pages 117—122.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António Lopes and André F. T. Martins. 2019b. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In Proceedings of the Fourth Conference on Machine Translation.

README: content of the tar.gz files

For the data files, each source-target language ($sl-$tl) folder has:

- each *.tsv file (training and dev), containing the following columns:

1) index: segment id
2) original: original sentence
3) translation: MT output
4) scores: list of DA scores by all annotators - the number of annotators may vary
5) mean: average of DA scores
6) z_scores: list of z-standardized DA scores
7) z_mean: average of z-standardized DA scores
8) model_scores: NMT model score for sentence

- *.doc_ids files contain the name of the article where each original segment came from

- 'word-probas' folder, containing the following files:

-- word_probas.*.$sl$tl: word log probabilities from the NMT model for each decoded token
-- mt.*.$sl$tl: actual output of the NMT model before any post-processing (same number of tokens as log probs above)