Direct Exploitation of Attention Weights for Translation Quality Estimation

Lisa Yankovskaya and **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{lisa_y,mark}@tartunlp.ai

Abstract

The paper presents our submission to the WMT2021 Shared Task on Quality Estimation (QE)[1]. We participate in sentence-level predictions of human judgments (`Task 1`) and post-editing effort (`Task 2`). We propose a glass-box approach based on attention weights extracted from machine translation systems. In contrast to the previous works, we directly explore attention weight matrices without replacing them with general metrics (like entropy). We show that some of our models can be trained with a small amount of a high-cost labelled data. In the absence of training data our approach still demonstrates a moderate linear correlation, when trained with synthetic data.

1 Introduction

Quality Estimation (QE, Blatz et al., 2004; Specia et al., 2009) is an essential part of the machine translation (MT) pipeline, which estimates the quality of the translation output without relying on any reference.

Unlike the previous year, three QE sentence-level tasks were presented in the WMT2021 Shared Task (Specia et al., 2021). The goal of `Task 1` is to predict direct assessments (DA), i.e. human judgments of translation quality (Graham et al., 2015), whereas in `Task 2`, the task is to estimate the post-editing effort required to obtain a correct translation which is measured by the HTER metric (Snover et al., 2006). The goal of `Task 3` is to determine if the translation output contains at least one critical error.

We propose a lightweight glass-box approach that can be applied to `Task 1` and `Task 2`. The approach is based on using the encoder-decoder attention weight matrices as input features for supervised translation quality estimation. Next we describe our approach (Section 2) and evaluate it experimentally (Sections 3–5).

[1] http://www.statmt.org/wmt21/quality-estimation-task.html

2 Approach

There are several QE models based on attention weights of neural MT systems described earlier (Yankovskaya et al., 2018; Fomicheva et al., 2020a,b). Their main idea is to compute the entropy of encoder-decoder attention weights for each target token and then average these entropies to get a sentence-level metric:

$$Entropy = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_{ji}\log\alpha_{ji},$$

where α represents attention weights, I is the number of target tokens and J is the number of source tokens.

Yankovskaya et al. (2018) work with attention weights extracted from LSTM (Hochreiter and Schmidhuber, 1997) MT systems. As LSTM has only one attention matrix, the approach of computing entropies is straightforward. However, neural MT models based on Transformer (Vaswani et al., 2017) have several layers and heads, so the number of computed entropies equals [Layers $\times$ Heads] for each sentence, which introduces some difficulties in this approach. To overcome it, Fomicheva et al. (2020a) summarise entropies by taking the average or minimum value to get an unsupervised attention-based QE metric. Fomicheva et al. (2020b) use the obtained entropies as features of regression models.

In this article, we propose another approach of using attention weights obtained from Transformer MT models: instead of summarising them into one metric, we feed all encoder-decoder attentions weights into a convolutional neural network (CNN) to get a QE score. We test the approach in a supervised setting, and also show that it can be applied in a zero-shot scenario when training data for the required language pair is not available.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 955–960
November 10–11, 2021.

3 Data

This year, three sentence-level tasks are available: predicting human judgments, post-editing effort and critical errors. In this work, we have focused on the first two tasks.

`Task 1` and `2` include eleven language pairs, seven of which have training (7 000 sentences), development (1 000 sentences) and two test sets (WMT2020 and WMT2021, 1 000 sentences each). For the other four languages only test sets (1 000 sentences) are available, which is called the zero-shot subtask. WMT2020 test set includes gold-labels whereas WMT2021 is the usual blind test without labels before submission.

To test our approach, we have focused on two language pairs with training data: English-German (En-De) and Estonian-English (Et-En), as well as one language pair without training data: English-Czech (En-Cz).

Besides data provided by the shared task organizers, we used additional parallel corpora to train CNN networks: the OpenSubtitles (Lison and Tiedemann, 2016), JRC Acquis (Steinberger et al., 2006), EuroParl (Koehn, 2005), DGT and EMEA (Tiedemann, 2012) corpora.

4 Settings

To compare the performance of our approach with CNN models to a previous baseline, we also ran experiments with models based on machine learning algorithms with entropies as input features (ML-Ent).

Below we present the experimental settings which we used for training ML-Ent and CNN models.

4.1 Machine Learning models with entropies as input features (ML-Ent)

There are two machine learning methods that we used. Random Forest (Ho, 1995) was chosen as a relatively easy and fast approach. We used the `sklearn`[2] library, set a randomized search on the hyperparameters and performed 5-fold cross-validation.

The second method is ensemble building based on (Caruana et al., 2004). The main idea behind the method is doing a greedy search over all trained models to find such models that would improve the ensemble's performance when added. We used the `mljar`[3] library, Random Forest and CatBoost (Prokhorenkova et al., 2018) algorithms, set Pearson as the evaluation metric and ran 5-fold cross-validation.

For both models and both tasks, we combined the proposed training and development sets (8 000 sentences in total) and used `[Heads × Layers]` (in our case 48) entropies for each translation as input.

4.2 CNN-based models

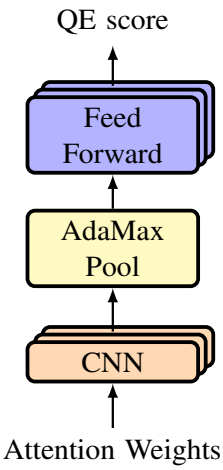

Figure 1: The proposed architecture of the QE model.

The base architecture of proposed CNN models is presented on Figure 1. The model's input is attention weights with shape (`[Heads × Layers]`, number of the source tokens, number of the target tokens). The number of `[Heads × Layers]`[4] is constant for all weights obtained from the same system, whereas the number of source and target tokens of each sentence can vary noticeably. To reduce the amount of padding added to each batch, we sort all sentences by the number of source/target tokens (max(src, tgt)) and only after that form a batch. Each CNN-based model consists of two or three CNN blocks, each of them comprises 2D-CNN, Batch Normalization, MaxPooling and Dropout Layers. We use `Relu` as the activation function. To handle the variable size of input batches, we use the Adaptive Max pooling layer. The last block of the model consists of three feed-forward layers. As a result, the model is trained to produce the desirable score: DA or HTER. We optimised our neural models with Adam (Kingma

[2] `https://scikit-learn.org/stable`

[3] `https://supervised.mljar.com/`

[4] 8 × 6 for En-Et and En-De, and 16 × 12 for En-Cs NMT systems

	En-De		Et-En		En-Cs
	wmt20	wmt21	wmt20	wmt21	wmt21
ML-Ent-RF	0.373	0.301	0.499	0.455	
ML-Ent-Ensemble	**0.395**	0.341	0.517	0.48	
CNN-DA	0.22	0.21	0.518	0.464	
CNN-BLEURT	0.383	0.357	0.577	0.526	0.299
CNN-BLEURT+	0.381	**0.369**	**0.599**	**0.547**	

Table 1: Pearson correlation coefficients between human DA scores and predicted values for WMT2020 and WMT2021 test sets (`Task 1`).

	En-De		Et-En	
	wmt20	wmt21	wmt20	wmt21
ML-Ent-RF	0.389	0.519	0.505	0.534
ML-Ent-Ensemble	0.408	**0.531**	0.519	**0.561**
CNN-HTER	**0.430**	0.503	**0.580**	0.549
CNN-HTERart	0.334	—	0.482	—

Table 2: Pearson correlation coefficients between HTER scores and predicted values for WMT2020 and WMT2021 test sets (`Task 2`).

and Ba, 2015).

`Task 1`**:** To predict DA scores, we considered three models with different training sets:

CNN-DA: we use human-labelled data provided by the shared task organizers: 7 000 for training set and 1 000 for development set;

CNN-BLEURT: we experiment with pre-training on synthetic data and for that we compute the BLEURT (Sellam et al., 2020) score for randomly chosen 300 000 sentences and use them as labels for training and development tests. We have chosen BLEURT to get artificial labels due to its good agreement with human judgments (Mathur et al., 2020);

CNN-BLEURT+: we fine-tune the model **CNN-BLEURT** on data provided by the organizers.

`Task 2` evaluates the proposed QE models for post-editing purposes.

CNN-HTER: we train a model with data provided by the shared task organizers;

CNN-HTERart: we use synthetically computed HTER between translation and reference. Though the preliminary experiments showed a poor performance compared to **CNN-HTER**, but this setting might be used in the absence of the human annotated training data.

5 Results

Below we present the obtained results and discuss the most interesting observations. To assess the performance of sentence-level QE models, Pearson correlation coefficient is used.

Table 1 shows results for the `Task 1`. For Et-En language pair, both CNN-BLEURT models show better results compared to ML-Ent models and CNN model trained only on DA score. For En-De, results are mixed. The CNN-DA model shows abysmal performance compared to both CNN-BLEURT and ML-Ent models. In contrast to Et-En, we can see that the performance of CNN-BLEURT and ML-Ent models is comparable.

Results for `zero-shot` En-Cs are not impressive (Table 1). One of the possible reasons for that is not using enough synthetic training data: while there are 300 000 examples for experiments with En-De and Et-En, we only use 50 000 for En-Cs.

The essential advantage of using CNN-BLEURT models is that they might be used for `zero-shot` settings when a training dataset is not available. However, the building and tuning of the neural network is not an easy task compared to ML-Ent models. The benefits of last ones are relatively fast training and fewer parameters that need to be tuned.

Table 2 presents results for the `Task 2`. We can see that for both languages, the results of ML-Ent and CNN-HTER models are pretty similar and

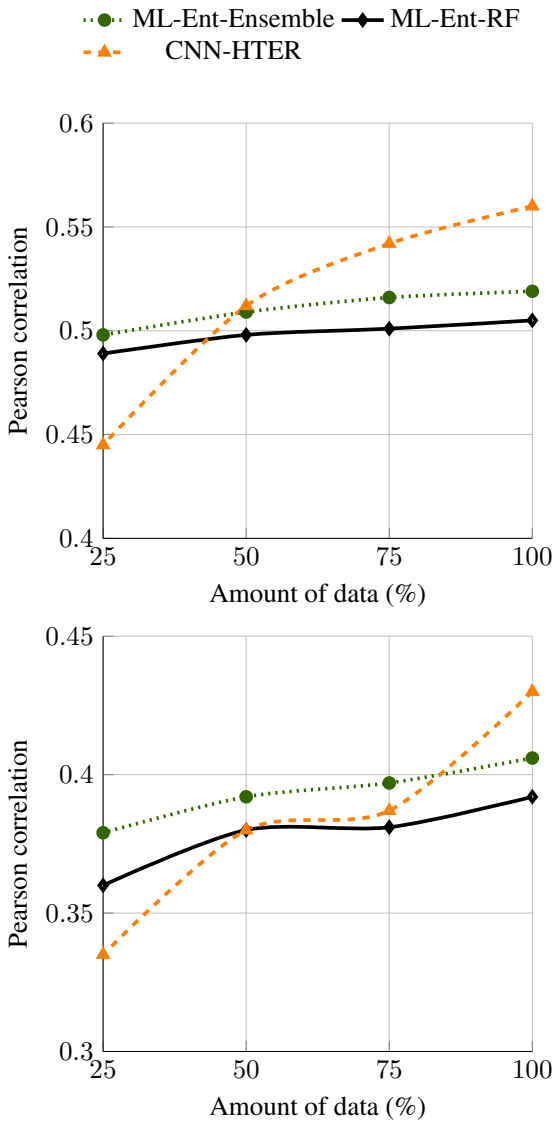

Figure 2: Pearson correlation coefficient between predicted values of WMT2020 test set and HTER scores for Et-En (top) and En-De (bottom) language pairs.

show a moderate correlation. As we mentioned in the previous chapter, the performance of CNN-HTERart is not as good as CNN-HTER, that is why we focus in this chapter only on CNN-HTER model.

Features of ML-Ent models are identical across DA- and HTER-models as well as CNN-DA and CNN-HTER share the same input features. Computed correlation coefficients of HTER- and DA-models are comparable for Et-En language pair. Nevertheless, we see a completely different picture for En-De: coefficients of DA-models are noticeably lower compared to HTER-models. As discussed in (Fomicheva et al., 2020a), the low results of DA models for En-De language pair might be caused by highly-skewed distribution of DA scores, as most translations have high quality scores.

Getting DA scores as well as HTER scores is a time-consuming and expensive task, so the less data you need, the better. To examine how labelled data we need to train models, we ran 10 tests for each examined amount of data (25%, 50%, 75%) and averaged the obtained correlation coefficients. According to our experiments, both discussed approaches, ML-Ent and CNN-HTER/DA, show comparable high performance even with a small amount of training/validation data. As Figure 2 shows, the performance of ML-Ent models for Et-En (top) and En-De (bottom) language pairs slightly worsens with decreased amounts of training data. The performance of CNN-HTER models decreases more noticeably, but still remains quite high. Especially in case of the En-Et language pair, all models demonstrate a moderate linear correlation with post-editing effort even with using 2000 training/validation examples (1750 for training and 250 for validation).

Raganato et al. (2018); Voita et al. (2019) showed that different layers play different roles in the attention mechanism. To examine it from the QE point of view, we compared CNN-HTER models with attention weights extracted from the first three layers, the last three layers and all six layers. According to Table 3, the performance of

	Et-En	En-De
all layers	0.580	0.43
first 3 layers	0.490	0.136
last 3 layers	0.536	0.43

Table 3: Pearson correlation coefficients between predicted values of WMT2020 test set and HTER scores. Results of three settings of CNN-HTER model are presented: with attention weights obtained (1) from all layers, (2) from the first three layers and (3) from the last three layers.

the models with last layers is comparable to the "all layers" models, whereas the difference between models with first layers and "all layers" models is more noticeable. While the performance gap between different models is not so noticeable for Et-En, then for En-De the difference is significant and even more, the lower layers do not provide any "useful" information to the model.

6 Conclusions

We presented sentence-level quality estimation models based on attention weights. The proposed models demonstrated a moderate linear correlation with human judgments as well as with required post-editing effort. The described models can be used as a cost-effective and light-weight QE approach in the machine translation pipeline. Results of empirical evaluation show a good performance even with a small amount of training data, as well as moderate performance in the absence of training data ("zero-shot" settings).

Acknowledgements

Lisa Yankovskaya and Mark Fishel were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). The authors also thank the Unversity of Tartu's High-Performance Computing Center for providing the computing infrastructure (University of Tartu, 2018).

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 315. Association for Computational Linguistics.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18.

Marina Fomicheva, Shuo Sun, Frédéric Blain Lisa Yankovskaya, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020a. Unsupervised quality estimation for neural machine translation.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020b. Bergamot-latte submissions for the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Tin Kam Ho. 1995. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Lei Ba. 2015. J. adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

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

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Alessandro Raganato, Jörg Tiedemann, et al. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and

Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

University of Tartu. 2018. Ut rocket cluster, https://doi.org/10.23673/ph6n-0144.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2018. Quality estimation with force-decoded attention and cross-lingual embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 816–821.