Explainable Quality Estimation: CUNI Eval4NLP Submission

Peter Polák and **Muskaan Singh** and **Ondřej Bojar**
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University
<polak,singh,bojar>@ufal.mff.cuni.cz

Abstract

This paper describes our participating system in the shared task Explainable quality estimation of 2nd Workshop on Evaluation & Comparison of NLP Systems. The task of quality estimation (QE, a.k.a. reference-free evaluation) is to predict the quality of MT output at inference time without access to reference translations. In this proposed work, we first build a word-level quality estimation model, then we finetune this model for sentence-level QE. Our proposed models achieve near state-of-the-art results. In the word-level QE, we place 2nd and 3rd on the supervised Ro-En and Et-En test sets. In the sentence-level QE, we achieve a relative improvement of 8.86% (Ro-En) and 10.6% (Et-En) in terms of the Pearson correlation coefficient over the baseline model.

1 Introduction

Quality Estimation (QE) or Confidence Estimation (CE) is a task of assessing the quality of machine-translated text given the source without accessing the reference (Blatz et al., 2004; Specia et al., 2009). QE can be assessed on sentence-level, word-level granularity or even document-level (Ive et al., 2018). Sentence-level scores predict what score would a human annotator assign to the whole sentence; most commonly, direct assessment (Graham et al., 2017) or HTER (Snover et al., 2006) serve as the golden standard. Word-level QE indicates word-level errors in machine translation output or incorrectly translated words in the source. While automatic word-level scores are usually continuous, the gold truth is binary: some words are labeled as correct while some are labeled as wrong. This estimation provides an aid in the translation workflow. For instance, it can help to determine if the machine-translated sentence is good enough to be used as-is or if it requires a human translator for post-editing or translating from scratch (Kepler et al., 2019b).

In this paper, we present our submission to the shared task of Explainable quality estimation of 2nd Workshop on Evaluation & Comparison of NLP Systems (Fomicheva et al., 2021).[1] Our solution is based on the XLM-R multilingual pre-trained model (Conneau et al., 2020). We first build a word-level quality estimation model. Then we finetune this model for sentence-level QE.

We make our code publicly available.[2]

2 Related Work

In the past decade, most of the quality estimation systems depended heavily on feature engineering, linguistic information, and machine learning algorithms such as support vector machines or randomized decision trees (Specia et al., 2013, 2015). In recent years, emerging neural-based QE have been outperforming earlier approaches on leaderboards of MT quality estimation (Kepler et al., 2019a). For instance, POSTECH (Kim et al., 2017), a purely neural system based on encoder-decoder recurrent neural network (referred to as a predictor) is stacked with bidirectional RNN (referred to as an estimator).

This predictor-estimator QE system was the best-performing one in WMT 2017.[3] It was further extended in DeepQuest architecture (Ive et al., 2018). These systems required extensive pre-training, which makes them dependent on large parallel corpora and computationally expensive. To overcome this problem, cross-lingual embeddings (Ruder et al., 2019) were used to reduce the burden of deep neural network architecture. TransQuest used these cross-lingual embeddings and was the best-performing sentence-level QE model at WMT 2020 QE Shared Task (Specia et al., 2018). For the

[1] https://eval4nlp.github.io/sharedtask.html

[2] https://github.com/pe-trik/eval4nlp-2021

[3] https://www.statmt.org/wmt17/quality-estimation-task.html

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 250–255

https://doi.org/10.26615/978-954-452-056-4_024

sentence-level QE task, the model was finetuned on multilingual pre-trained representations. For the word-level QE task, the authors used direct assessment (DA) quality scores from the MLQE-PE dataset. Motivated by this work, we finetune our word-level model to yield sentence-level QE.

Overall, the tremendous progress in the field of quality estimation is achieved thanks to the annual focus of the shared task organized by WMT and thanks to the annotated data released in these tasks, leading to the development of various open-source systems such as QuEst (Specia et al., 2013), QuEst++ (Specia et al., 2015), deepQuest (Ive et al., 2018), OpenKiwi (Kepler et al., 2019b) and TransQuest (Ranasinghe et al., 2020).

3 Task Description

The task consists of building a quality estimation system that (1) predicts the quality score for an input pair of the source text and MT hypothesis, and (2) provides word-level evidence for its predictions.

4 Dataset Description

The dataset for the shared task consists of training, development, and test sets. The training and development sets are Estonian-English (Et-En) and Romanian-English (Ro-En) partitions of the MLQE-PE dataset (Fomicheva et al., 2020). The test set consists of sentence-level quality scores and word-level error annotations for these two language pairs. The goal of the shared task is to estimate the word quality in unsupervised settings (no training data for word-level QE whatsoever). However, participating systems can also be labeled as "unconstrained" and use word-level QE training data.

Additionally, there are zero-shot test sets for two language pairs, i.e., German-Chinese (De-Zh) and Russian-German (Ru-De), where no sentence-level OR word-level annotations were available at training time.

5 Methodology

In the proposed system, we use a pre-trained XLM-R model (Conneau et al., 2020) to obtain representations of input sentences in continuous space. XML-R is trained on large-scale multilingual CommonCrawl datasets. We have two separate models for the word-level and sentence-level quality estimation.

5.1 Data Representation

The pre-trained XLM-R model uses BPE encoding (Sennrich et al., 2016) for input tokenization. In order to input two sentences to the model (i.e., the source and the MT candidate), the sentences are concatenated with two </s> tokens[4] between them:

<s> $s_1, ..., s_m$ </s></s> $t_1, ..., t_n$ </s>.

5.2 Word-level QE

In the word-level QE, the task is to predict for each source and target word if it was translated correctly. We use the XLM-R model extended a linear layer on top of the hidden-states output (see Figure 1). We use cross-entropy loss. The BPE encoding might break a word into more tokens; this is especially true for the less frequent or misspelled words (Polák, 2020). E.g., "misstake" (with double s) might be broken down into two tokens: "_miss" and "take". Because we are interested in word-level predictions and not token-level predictions, we label only the first token of each word and put the "ignore" label to others (including the special tokens <s> and </s>).

We also experiment with an alternative representation: we put a <cls> token after each word. In this case, we ignore all labels except for <cls>. But as we will document later, this alternative does not bring any improvement.

5.3 Sentence-level QE

For the sentence-level QE, we use the XLM-R model extended with a linear layer on top of the pooled output. We finetune the model with mean square error loss. We normalize the scores to interval [0; 1].

6 Experiments and Results

In all our experiments, we use XLM-R large model[5] using Hugging Face Transformers (Wolf et al., 2020). We run all our experiments on NVIDIA RTX 3090. To find the best parameters, we run a grid search. The optimal hyper-parameters are $2e^{-5}$ learning rate, three epochs, and batch size of 16.

[4]We follow the tokenization procedure in https://huggingface.co/transformers/model_doc/xlmroberta.html#xlmrobertatokenizerfast.

[5]https://huggingface.co/xlm-roberta-large

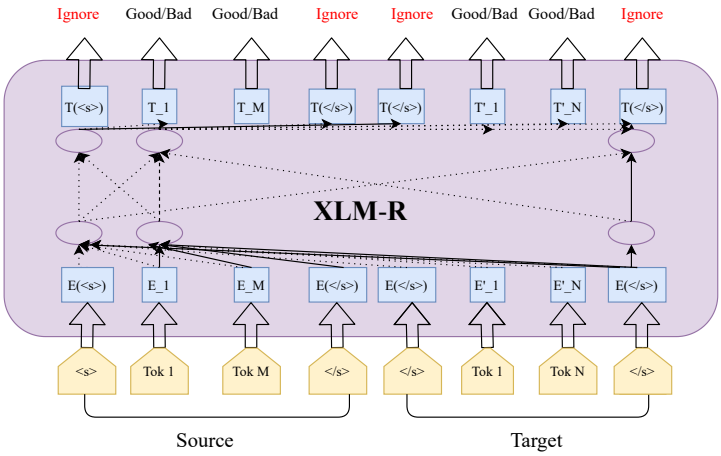

Figure 1: Our model architecture for the word-level quality estimation task.

	AUC	AP	Recall at top-K
ro-en only	0.926	0.840	0.738
ro-en & et-en	**0.933**	**0.861**	**0.770**

Table 1: Models trained on Ro-En data only and all training data performance comparison on Ro-En development set. The score differences are statistically significant (paired t-test with $p < 0.01$).

6.1 Quantitative Analysis

In this section, we provide the analysis across each level of experiments.

6.1.1 Word-level QE

First, we consider having separate models for source and target sentence word-labels prediction. We compared the separate models with a joint one, but we did not find any statistically significant evidence of difference.

We also consider the variant with `<cls>` token. Again, the model does not perform better. Additionally, we noticed a much larger GPU-memory consumption.

We were also interested in whether to use all training data (Ro-En and Et-En) or to match the test pair (two test sets are supervised with matching language pairs). We found out that it is better to include both language pairs (see Table 1). This suggests the model is learning the tasks independent of language pair and benefits from more training data.

6.1.2 Sentence-level QE

We first tried a simple approach computing the sentence-level QE using the word labels. We defined the score as 1 minus the geometric mean of the word-level scores (i.e., the probability of a word being incorrectly translated). The scores surpassed the shared task author's XMOVER-SHAP (Zhao et al., 2020) baseline with Pearson correlation coefficients of 0.501 and 0.648 compared to 0.415 and 0.638 on Et-En or Ro-En development sets, respectively.

Our second approach is again based on the XLM-R model. We tried to use the vanilla XLM-R pre-trained model as in the word-level QE task. The model failed to converge. To circumvent this, we finetuned the best-performing word-level QE model. This provided much better results (0.776 and 0.880 on Et-En or Ro-En development sets respectively).

6.1.3 Ensembling

To leverage all the models we trained during the hyper-parameter search we employ ensembling. We tried two approaches: (1) weighted geometric mean and (2) weighted arithmetic mean. The former failed to produce results as the number of models was too large (36 models).

We estimate the model weights using Bayesian optimization with target sentence AUC as an objective function. We optimized the parameters on the development sets (for each Et-En and Ro-En separately). For the zero-shot test sets (De-Zh or Ru-De), we averaged the weights obtained for both development sets (Et-En and Ro-En).

Table 2 documents slight improvement using ensembling compared to the best model.

6.2 Qualitative Analysis

We were interested to see, how well the model performed on unseen language pairs. We had no bilingual speaker of the provided language pairs.

		AUC	AP	Recall at top-K
et-en	best model	0.884	0.819	0.722
	ensemble	**0.892**	**0.829**	**0.729**
ro-en	best model	0.933	0.861	0.770
	ensemble	**0.939**	**0.871**	**0.782**

Table 2: Comparison of the best model and ensemble. Results are word-level target side QE on development sets.

	/Salvadorans/	/protested/	/against/	/bitcoin/
1.	Salvádorčania	protestovali	proti	bitcoinu
	0.40	0.02	0.02	0.02
	Salvadorianer	protestierten	gegen	Bitcoin
	0.27	0.01	0.02	0.00
2.	Salvádorčania	protestovali	proti	bitcoinu
	0.49	0.03	0.14	0.04
	Salvadorianer	protestierten	**für** /for/	Bitcoin
	0.44	0.04	0.27	0.02
3.	Salvádorčania	protestovali	proti	bitcoinu
	0.98	0.03	0.03	0.04
	Somalis	protestierten	gegen	Bitcoin
	0.97	0.02	0.02	0.01

Table 3: Examples of unsupervised Slovak-German MT QE. Each number below a word represents a probability of being incorrectly translated generated by our best-performing model. Words in slashes present English translations. Words in bold denote translation errors.

Therefore, we assembled our own Slovak-German test set (we used Google Translate to obtain the translation and we introduced artificial errors). We draw three examples in Table 3. We see the model also works on an unseen language pair (recall the XLM-R model was pre-trained on both languages, but we use different languages for the QE task finetuning).

In the first example (a correctly translated sentence), the model labels the words correctly (although we see a bit of uncertainty over the first, less frequent word). The second example changes the meaning (protested *for* instead of protested *against*). We see some hesitation in the third word, still, the predicted probabilities are incorrect. We hypothesize this may be due to the fact the model prefers predicting based on co-occurrences, not necessarily on the meaning (Kim et al., 2019). In the third example, the model correctly detects the first mistranslated word both in the source and target sentence.

6.3 Comparative analysis of our submission with other submissions.

In the shared task, 11 teams from different organizations participated. Out of the 11 different submissions, two teams submitted to the unconstrained track. The rest of the teams submitted to the constrained track as they did not use any supervision at the word level. Our submission was also in the unconstrained category. We report all the results from the leaderboard[6] in Table 4.

7 Conclusion

The paper describes our submission to the shared task, EVAL4NLP, co-located with EMNLP. First, we built a word-level quality estimation model. Then we finetuned it for obtaining sentence-level QE. In the word-level QE, we placed 2nd and 3rd on the supervised Ro-En and Et-En test sets. In the sentence-level QE, we achieved a relative improvement of 8.86% (Ro-En) and 10.6% (Et-En) in terms of the Pearson correlation coefficient over the baseline model.

Acknowledgements

The work was supported by the grant 825303 "Bergamot" of European Union's Horizon 2020 research and innovation programme, 19-26934X "NEUREM3" of the Grant Agency of the Czech Republic, and START/SCI/089 (Babel Octopus: Robust Multi-Source Speech Translation) of the START Programme of Charles University.

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 *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The

[6] https://competitions.codalab.org/competitions/33038#results

	Source			Target		
	AP	**AUC**	**Recall at top-K**	**AP**	**AUC**	**Recall at top-K**
Et-En (supervised)						
Our Submission	0.851	0.928	0.761	0.841	0.917	0.751
Average of all teams	0.496	0.610	0.397	0.651	0.759	0.548
Best of all teams	0.855	0.935	0.771	0.852	0.922	0.762
Ro-En (supervised)						
Our Submission	0.808	0.934	0.696	0.830	0.935	0.730
Average of all teams	0.445	0.603	0.344	0.653	0.788	0.544
Best of all teams	0.851	0.947	0.752	0.869	0.946	0.778
Ru-De (unsupervised)						
Our Submission	0.668	0.802	0.562	0.610	0.759	0.502
Average of all teams	0.421	0.562	0.332	0.494	0.669	0.389
Best of all teams	0.804	0.922	0.709	0.829	0.927	0.736
De-Zh (unsupervised)						
Ours	0.423	0.618	0.268	0.435	0.611	0.303
Average of all teams	0.279	0.468	0.160	0.406	0.603	0.279
Best of all teams	0.645	0.847	0.509	0.679	0.849	0.571

Table 4: Comparative results from the leaderboard. The scores were evaluated against error labels resulting from manual annotation. Missing word errors were ignored in this track. The main metrics for evaluation were AUC and AUPRC scores for word-level explanations.

eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019a. Unbabel's participation in the wmt19 translation quality estimation shared task. *arXiv preprint arXiv:1907.10352*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019b. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

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*, pages 562–568.

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019. Qe bert: bilingual bert using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 85–89.

Peter Polák. 2020. Spoken Language Translation via Phoneme Representation of the Source Language. Master's thesis, Charles University, Faculty of Mathematics and Physics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest at wmt2020: Sentence-level direct assessment. *arXiv preprint arXiv:2010.05318*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

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 the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120.

Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. Quest-a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *EAMT*, volume 9, pages 28–35.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.