Bad MT Systems are Good for Quality Estimation

Iryna Tryhubyshyn[⋆] tryhubyshyn@gmail.com
Aleš Tamchyna[†] ales.tamchyna@phrase.com
Ondřej Bojar[⋆] bojar@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic[⋆]
Phrase a.s., Václavské náměstí 2132/47, 110 00 Prague, Czech Republic[†]

Abstract

Quality estimation (QE) is the task of predicting quality of outputs produced by machine translation (MT) systems. Currently, the highest-performing QE systems are supervised and require training on data with golden quality scores. In this paper, we investigate the impact of the quality of the underlying MT outputs on the performance of QE systems. We find that QE models trained on datasets with lower-quality translations often outperform those trained on higher-quality data. We also demonstrate that good performance can be achieved by using a mix of data from different MT systems.

1 Introduction

Quality Estimation (QE) involves predicting the quality of a machine-translated text based on the original text and the machine translation (MT) output (Blatz et al., 2004; Specia et al., 2009). This can be done at the word, sentence, or document level.

In this paper, we focus on sentence-level QE, where the goal is to predict a score that a human assessor would attribute to the sentence. Depending on the manual evaluation process used to gather data, we can talk about different variations of the task. These include Direct Assessment QE (Graham et al., 2015), which aims to estimate the perceived quality of translation, Post-editing QE, which measures the effort required to edit the translation, and MQM QE (Lommel et al., 2014), which identifies critical errors in the translation.

Evaluating a QE system means checking how closely its predictions match manual scores on a held-out set. QE systems are closely tied to MT systems in many ways. Their performance can vary greatly depending on the MT system on which they are being evaluated. The current high-performing solutions for quality estimation are based on supervised methods, which in turn makes these QE systems dependent on the specifics of the MT systems used to create the training data. It is not clear which MT system should be used to create a QE system with the best performance. The contributions of this experimental work are as follows:

1. We examine the relationship between MT system quality and QE system performance by training QE models on datasets that consist of the same source data but different translations produced by MT of varying quality.
2. We evaluate the models on evaluation datasets from different domains and show that the QE system trained on translations from low-quality MT systems outperforms the QE system trained on translations from high-quality MT systems.

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 200–208
September 4–8, 2023, Macau SAR, China.

3. We demonstrate that QE systems trained on a mix of translations from different MT models also show good performance but do not necessarily outperform the best-performing system that is trained on the translations from one MT system.

2 Proposed approach

We investigate the impact of MT system quality on QE system performance by training QE models on datasets consisting of a fixed set of source sentences and differing in the target side which is translated by MT systems of varying quality. As there are no existing QE datasets that have the same source sentences translated by different MT systems of known performance, we create our own datasets by training MT systems and translating the same source sentences. Due to the lack of human annotators and a large amount of work required, we approximate the manual quality scores, i.e. our targets for QE are assigned automatically. The scores are assigned by calculating the similarity between the translations and reference translations available in a parallel dataset. Note that for QE itself, reference translations are not needed, only the quality judgments.

We explore the use of two automatic reference-based metrics of MT quality, namely TER (Snover et al., 2006a) and COMET (Rei et al., 2020), as the golden truth for QE training. We select these metrics because they mimic the manual targets typically used in QE tasks, and each highlights a distinct aspect of translation quality. Specifically, COMET has been trained to predict sentence-level Direct Assessment scores, while TER is a proxy for HTER (Snover et al., 2006b), which measures post-editing effort. Additionally, we conduct the evaluation of the models trained on COMET scores on available data with Direct Assessment scores to demonstrate that the relationship that holds for proxy targets also applies to real targets.

COMET is a metric based on sentence embeddings and designed to predict the quality score that a human annotator would assign. This leads us to believe that COMET reflects the overall meaning match. As a pre-trained metric, it has a high correlation with human-based scores. However, its training to directly predict DA scores is also a limitation. COMET may contain a bias towards the MT systems on which it was trained, which is the exact bias that we are trying to evaluate in our QE systems. While COMET is available in QE mode with multiple releases, it is not suitable for our purposes, since they differ in various aspects like training procedures, source data, and MTs used in training. Our focus, however, is solely on understanding the impact of the MT used in translation and using QE COMET models would not allow us to separate the MT's impact from other factors affecting the QE evaluation.

TER, on the other hand, is focused on string editing, which means a rather superficial similarity of the candidate and the reference translation. It uses the same mechanism of string comparison as HTER, so we use it as a proxy for HTER-measured post-editing effort. TER is known for having a lower correlation with translation targets. However, it is not trained on translations of any kind, so the risk of any bias towards some training data is avoided.

3 Experiments

Our experimental approach involves training QE systems on translations of varying quality, and then evaluating their performance on datasets with different target types, namely COMET and TER targets, as well as DA targets. In this section, we provide a detailed description of our experimental setup, including information on how we trained the MT and QE systems, as well as the datasets used for training and evaluation.

3.1 Setup

For our experiments, we need MT systems of varying performance. We achieve this by adjusting the amount of training data used, with one MT system trained on 10 million sentence pairs

Dataset	Domain	Sentences	Words	Distinct words
CzEng	Mixed: Europarl, News commentary, Wikititles, etc.	10 000	124 481	26 466
WMT18	News	2983	55 920	12 548
Antrecorp	Student presentations by non-native English speakers	571	7 893	1 532
SAO	Presentations by officers of two supreme audit institutions	654	13 158	1 897
Khan Academy	Subtitles to math educational videos	538	4 470	871

Table 1: Datasets used in evaluation: domain, sentence and word count, vocabulary size. We report the statistics only for the source language (English). Antrecorp, SAO, and Khan Academy are parts of IWSLT dataset.

displaying superior quality compared to a second MT system trained on 1 million sentence pairs. Additionally, a mixed dataset is also created by utilizing the same source data, with translations randomly selected from both the high-quality and low-quality datasets at the 50:50 ratio.

Separate QE systems are trained for each type of target: one system is trained for direct assessment using COMET targets, and another system is trained for post-editing effort using TER targets. One system is trained on each dataset, resulting in a total of six QE systems (COMET and TER times low, high and mixed quality MT).

Training dataset. The experiments are performed on the English→Czech language pair. The MT and QE systems are trained on the authentic CzEng 2.0 dataset (Kocmi et al., 2020) using randomly selected non-overlapping parts: 10 million sentences for the MT training data and 500 thousand for the QE data.

MT systems. The MT systems trained are Transformers with base configuration in the Marian implementation (Junczys-Dowmunt et al., 2018). The default settings for the Transformer provided by the Marian package are used, only setting the pre-allocated memory space to 6500 MB for maximum possible batch size. Each system is trained on two GeForce GTX 1080 Ti GPUs. The dataset preprocessing includes normalization, tokenization, and truecasing using the Moses toolkit (Koehn et al., 2007), followed by BPE tokenization (Sennrich et al., 2016) with 32,000 merge operations.

QE systems. All our QE models use the Predictor-Estimator architecture (Kim et al., 2017) in the OpenKiwi implementation (Kepler et al., 2019) with XLM-R (Conneau et al., 2019) as the predictor. We follow the default settings for the XLM-R model adjusting certain parameters for the larger dataset size. These adjustments include setting the learning rate to 5e-6, using 1000 warm-up steps, and unfreezing the XLM-R predictor after 2000 steps. Additionally, the model is validated every 25 thousand sentences and the training process is stopped if the Pearson correlation of the predictions and the targets does not increase for 25 times in a row. The batch size of 4 with four gradient accumulation steps is used to fit the data into memory.

3.2 Evaluation datasets

The evaluation was carried out on three different datasets: one extracted from CzEng avoiding any overlap with the training data, an evaluation dataset from the WMT-2018 News Translation Task (Bojar et al., 2018), and a dataset used in the IWSLT 2020 Non-Native Speech Translation

Evaluation dataset	COMET models		TER models	
	Low-quality	High-quality	Low-quality	High-quality
CzEng	**0.638**	0.623	**0.524**	0.503
WMT18	**0.757**	0.744	**0.461**	0.435
IWSLT	**0.599**	0.594	**0.404**	0.357

Table 2: Evaluation of QE models trained on datasets generated by one MT (of a lower vs. higher quality), measured by Pearson correlation between predictions and targets. The winning model is denoted in bold. Results that are statistically significant at the 0.05 level are underlined.

Task (Ansari et al., 2020) that combines three sources of data: Antrecorp (Macháček et al., 2019), Khan Academy, and SAO. Table 1 provides information on the datasets, including the domain, size, and statistics such as the number of words and vocabulary size (distinct words) per dataset.

For the WMT-2018 dataset, we used translations obtained from MT systems that were entered into the competition. As an additional dataset, we use DA scores collected during the competition evaluations that are available only for a part of the dataset. IWSLT and CzEng were translated by various MT systems: the two explained in Section 3.1, Google Translate, and LINDAT Translation (sentence-level system).[1] Each QE evaluation dataset is then composed of translations combined from all MT systems, with two sets of targets computed using COMET and TER against the reference translations available for the respective test sets. We use the same test set for the evaluation of QE across all six QE settings.

4 Results

We evaluate the performance of our QE models by computing the Pearson correlation between their predictions and the corresponding targets. To determine whether there is a statistically significant difference in correlation between the models, we use a z-test on Fisher z-transformed correlation coefficients.

4.1 QE models derived from a single MT

Table 2 displays the evaluation results of QE models trained on translations from a single machine translation system. The "Low-Quality" column shows the results for QE models trained on the corpus with low-quality translations produced by the lower-quality MT, and the "High-Quality" column shows the results for QE models trained on the same corpus but with high-quality translations from the higher-quality MT.

On all datasets, the QE models trained on lower-quality translations perform better than those trained on higher-quality translations. This phenomenon is statistically significant for all datasets except IWSLT with COMET labels. These results indicate that choosing high-quality translations for training a QE system may actually result in an inferior performance compared to training on low-quality translations. This goes against conventional wisdom and suggests that opting for a mediocre MT instead of the best-performing one may be a wiser choice when selecting data to train a QE system.

4.2 How QE models' performance is affected by the evaluated MT system

This section focuses on analyzing how the performance of QE models varies depending on the MT system that is the subject of the quality estimation. We reused the data from the previous

[1] https://lindat.mff.cuni.cz/services/translation/

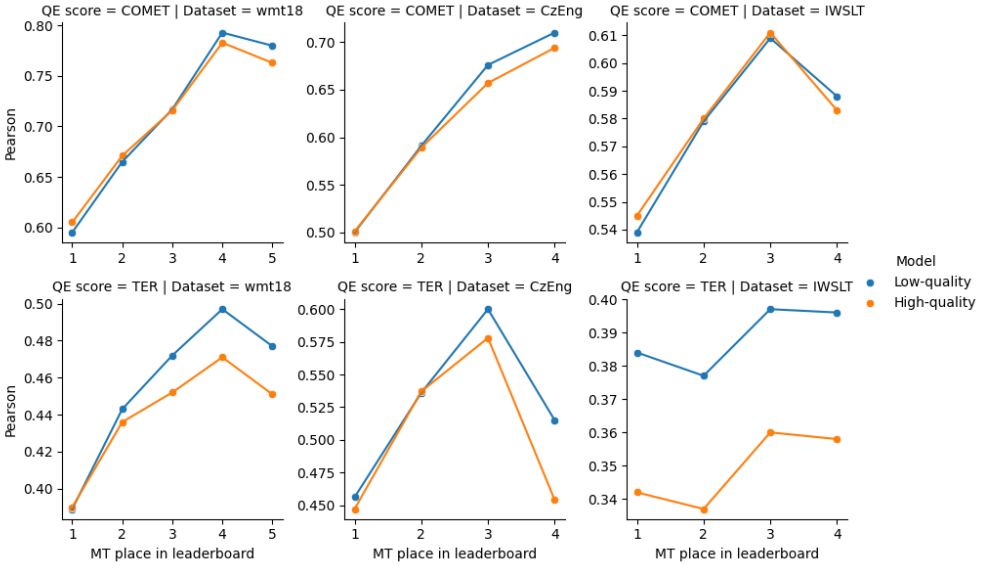

Figure 1: QE model performance for each translator separately measured by Pearson correlation between predictions and targets. On the x-axis, MT systems are sorted by their decreasing performance, with the MT that achieved the top position in the leaderboard labeled as 1, the second-best system as 2, and so on. The two lines correspond to lower and higher quality QE, i.e. QE trained on worse or better MT systems, resp.

section and evaluated the model's performance on test set translations produced by each MT system individually. We rank the evaluated MT systems by the quality of their translations using system-level COMET scores (MT evaluation results are available in Appendix A). Figure 1 shows how performance of the QE systems varies depending on the quality of the evaluated MT system. The results reveal a clear trend: the QE models' performance decreases as the quality of the evaluation dataset increases.

Interestingly, we also note that the low-quality and high-quality QE models exhibit different behaviors. The low-quality QE models (i.e. those trained on low-quality MT outputs) perform better on datasets lower on the leaderboard, but their performance deteriorates when they encounter more challenging translations of higher quality. We observe this behavior in all evaluation datasets, except for IWSLT with TER targets. On translations with higher quality, both high-quality and low-quality QE models perform on the same level, with high-quality models sometimes outperforming low-quality models. On translations with lower quality, low-quality translations QE models outperform high-quality translations QE models.

It is evident from these findings that the selection of optimal training data for QE models must take into account the intended application of the model, particularly the quality of the MT systems it will be operating on. Considering that the evaluation datasets were mostly constructed using MT systems that outperform the one used for generating translations to train lower-quality QE models, we suggest opting for data obtained from a slightly inferior translation system.

Evaluation dataset	COMET targets		DA targets	
	Low-quality	High-quality	Low-quality	High-quality
CUNI Transformer	0.570	**0.592**	0.349	**0.378**
UEDIN	0.645	**0.650**	0.427	**0.432**
online-B	**0.698**	0.693	**0.501**	0.493
online-A	**0.777**	0.767	**0.574**	0.567
online-G	**0.767**	0.754	**0.536**	0.523
Whole dataset	**0.743**	0.731	**0.524**	0.517

Table 3: Evaluation results for WMT-18 dataset with DA and COMET targets, measured by Pearson correlation between predictions and targets. For each type of target, the winning model is denoted in bold.

Evaluation dataset	COMET models		TER models	
	Best single MT	Mixed	Best single MT	Mixed
CzEng	0.638	**0.643**	**0.524**	0.518
WMT18	0.757	**0.764**	0.461	**0.471**
IWSLT	0.599	**0.605**	**0.404**	0.373

Table 4: Evaluation results comparing QE models trained on single-MT dataset with models trained on data mixed from different MTs. For better readability, we only show which model leads to better results.

4.3 Evaluation on DA scores

In the absence of a large-scale QE dataset labeled by humans, we have trained our QE models on proxy metrics, namely TER and COMET, and then evaluated them on datasets that also use these proxy metrics. In this section, we assess our QE models using DA scores that were generated for the WMT-18 competition to evaluate MT systems. However, these scores are only available for a subset of the data, so we compare them to results for the same subset of data with COMET targets. Table 3 shows that despite the overall lower performance on DA scores, the trend in the relationship between high-quality and low-quality QE models remains the same. The low-quality QE model performs better than the high-quality QE models, and just like with COMET labels, its performance deteriorates quicker than that of higher-performing models. As a result, high-quality models perform better only on translations from CUNI Transformer and UEDIN, which are the top MT systems in WMT-18. This evidence suggests that the relationship between lower-quality and higher-quality QE models is likely to be the same with actual human-based metrics: For standard quality MT outputs, it is safer to train QE on lower-quality MT.

4.4 QE models based on more MT systems

In this section, we investigate the effect of combining datasets created by MT systems of different qualities, compared to using datasets from a single MT (either lower or higher quality). The evaluation results are shown in Table 4. The column titled "Best single MT" displays the performance of the best QE systems trained on data from a single MT, namely the one that employs lower-quality translations. The column labeled "Mixed" presents the evaluation results for QE models trained on a combination of high-quality and low-quality translations.

Overall, the results suggest that combining stronger and weaker MT systems when prepar-

ing training data for QE may not necessarily improve QE performance. The outcome depends on the specific settings in which the models will be used. While the mixed setting shows better results, we would like to point out that adding more machine-translated datasets to the QE training data may come at a cost. If there are good translation data from one MT that yield good QE results, it may not be worth the effort to mix it with data from another MT.

5 Conclusion

Our study investigated the impact of MT quality used to train QE systems on the performance of the QE systems. We trained QE models on the datasets that consist of the same source data but different translations produced by MT systems of varying quality. The findings revealed that QE models trained on lower-quality MT translations tended to perform better than those trained on higher-quality MT outputs. Additionally, the study suggests that mixing the better and worse MT outputs for training QE models may not necessarily lead to improved QE performance, and the results may vary depending on the specific application or usage scenario.

Acknowledgement

This research was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

Ansari, E., Axelrod, A., Bach, N., Bojar, O., Cattoni, R., Dalvi, F., Durrani, N., Federico, M., Federmann, C., Gu, J., Huang, F., Knight, K., Ma, X., Nagesh, A., Negri, M., Niehues, J., Pino, J., Salesky, E., Shi, X., Stüker, S., Turchi, M., Waibel, A., and Wang, C. (2020). FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2015). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3 – 30.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019). 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, Florence, Italy. Association for Computational Linguistics.

Kim, H., Jung, H.-Y., Kwon, H., Lee, J.-H., and Na, S.-H. (2017). Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17:1–22.

Kocmi, T., Popel, M., and Bojar, O. (2020). Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lommel, A., Burchardt, A., Popović, M., Harris, K., Avramidis, E., and Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

Macháček, D., Kratochvíl, J., Vojtěchová, T., and Bojar, O. (2019). A speech test set of practice business presentations with additional relevant texts. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings*, page 151–161, Berlin, Heidelberg. Springer-Verlag.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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, Berlin, Germany. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006a). 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, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006b). 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, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Specia, L., Turchi, M., Cancedda, N., Cristianini, N., and Dymetman, M. (2009). Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

A MT systems evaluation

	MT	COMET
1	CUNI Transformer	0.800
2	UEDIN	0.720
3	online-B	0.587
4	online-A	0.321
5	online-G	0.191

Table 5: Evaluation of MT systems that compose WMT-18 dataset measured with COMET score

	MT	COMET
1	LINDAT	0.778
2	Our high-quality MT	0.729
3	Our low-quality MT	0.604
4	Google Translate	0.390

Table 6: Evaluation of MT systems that compose CzEng dataset measured with COMET score

	MT	COMET
1	LINDAT	0.629
2	Our high-quality MT	0.540
3	Google Translate	0.500
4	Our low-quality MT	0.437

Table 7: Evaluation of MT systems that compose IWSLT dataset measured with COMET score