Are we Estimating or Guesstimating Translation Quality?

Shuo Sun
Johns Hopkins University
ssun32@jhu.edu

Francisco Guzmán
Facebook AI
fguzman@fb.com

Lucia Specia
Department of Computing
Imperial College London, UK
l.specia@imperial.ac.uk

Abstract

Recent advances in pre-trained multilingual language models lead to state-of-the-art results on the task of quality estimation (QE) for machine translation. A carefully engineered ensemble of such models dominated the QE shared task at WMT 2019. Our in-depth analysis, however, shows that the success of using pre-trained language models for QE is overestimated due to three issues we observed in current QE datasets: (*i*) The distributions of quality scores are imbalanced and skewed towards good quality scores; (*ii*) QE models can perform well on these datasets without even ingesting source or translated sentences; (*iii*) They contain statistical artifacts that correlate well with human-annotated QE labels. Our findings suggest that though QE models might capture *fluency* of translated sentences and *complexity* of source sentences, they cannot model adequacy of translations effectively.

1 Introduction

Quality Estimation (QE) (Blatz et al., 2004; Specia et al., 2009) for machine translation is an important task that has been gaining interest over the years. Formally, given a source sentence, s and a translated sentence, $t = \phi(s)$ where ϕ is a machine translation system, the goal of QE is to learn a function f such that $f(s, t)$ returns a score that represents the quality of t, without the need to rely on reference translations.

QE has many useful applications: QE systems trained to estimate Human-mediated Translation Error Rate (HTER) (Snover et al., 2006) can automatically identify and filter bad translations, thereby reducing costs and human post-editing efforts. Industry players use QE systems to evaluate translation systems deployed in real-world applications. Finally, QE can also be used as a feedback mechanism for end-users who cannot read the source language.

Recently, language models pre-trained on large amounts of text documents lead to significant improvements on many natural language processing tasks. For instance, an ensemble of multilingual BERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019) models (Kepler et al., 2019a) won the QE shared task at the Workshop on Statistical Machine Translation (WMT 2019) (Fonseca et al., 2019), outperforming the baseline neural QE system (Kepler et al., 2019b) by 42.9% and 127.7% on the English-German and English-Russian sentence-level QE tasks respectively.

While pre-trained language models contribute to tremendous improvements on publicly available benchmark datasets, such increases in performance beg the question: Are we really learning to *estimate* translation quality? Or are we just *guessing* the quality of the test sets? We performed a careful analysis that reveals that the latter is happening, given several issues with QE datasets which undermine the apparent success on this task:

(*i*) The distributions of quality scores in the datasets are imbalanced and skewed towards high-quality translations. (*ii*) The datasets suffer from the partial-input baseline problem (Poliak et al., 2018; Feng et al., 2019) where QE systems can still perform well while ingesting only source or translated sentences. (*iii*) The datasets contain domain-specific lexical artifacts that correlate well with human judgment scores.

Our results show that though QE systems trained on these datasets can capture *fluency* of the target sentences and *complexity* of the source sentences, they are over-leveraging lexical artifacts instead of modeling *adequacy*. From the findings above, we conclude that QE models cannot generalize, and the successes in this task are overestimated.

2 Methodology

In this paper, we analyze three different instances of *sample bias* that are prevalent in QE datasets, which affect the generalization that models trained on them can achieve.

Lack of label diversity With the advent of NMT models, we have seen an increase in the quality of translation systems. As a result, a random sample of translations might have few examples with low-quality scores. Systems trained on imbalanced datasets and tested on similar distributions can get away with low error rates without paying much attention to samples with bad quality scores. To detect these issues, we analyze the labels and predicted score distributions for several models.

Lack of representative samples We want to have datasets that adequately represent both the *fluency* and *adequacy* aspects of translation. QE datasets should have a mixture of instances that model both high and low adequacy irrespective of the fluency. To evaluate if our models learn both aspects of translation quality, we run *partial input* experiments, where we train systems with only the source or target sentences and analyze the discrepancies w.r.t to the full-input experiments.

Lack of lexical diversity Most QE datasets come from a single domain (e.g., IT, life sciences), and certain lexical items can be associated with high-quality translations. Lexical artifacts are also observed in monolingual datasets across different tasks (Goyal et al., 2017; Jia and Liang, 2017; Kaushik and Lipton, 2018). For example, Gururangan et al. (2018) find that annotators are responsible for introducing lexical artifacts into some natural language inference datasets because they adopt heuristics to generate plausible hypothesis during annotation quickly. Here, we use Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009) to find possible lexical artifacts associated with different levels of HTER.

2.1 Experimental Setup

We experiment with recent QE datasets from WMT 2018 and 2019. For every dataset, a Statistical Machine Translation (SMT) system or Neural Machine Translation (NMT) system was used to translate the source sentences. The translated sentences were then post-edited by professional translators. HTER scores between translated sentences and post-edited sentences were calculated with the TER[1] tool and clipped to the range [0, 1]. HTER score of 0 means the translated sentence is perfect, while 1 means the translated sentence requires complete post-editing. Since the test sets for WMT2018 are not publicly available, we randomly shuffled those datasets into train, dev, and test splits, following the ratio of approximately 8 to 1 to 1. Table 1 presents statistics of the QE datasets.

				size (K)		
Dataset	**langs**	**dom.**	**syst.**	*train*	*dev*	*test*
WMT18*	en-de	IT	SMT	21.8	2.7	2.7
		IT	NMT	11.5	1.4	1.4
	en-cs	IT	SMT	33.0	4.1	4.1
	en-lv	SCI	SMT	9.8	1.2	1.2
		SCI	NMT	11.1	1.3	1.3
	de-en	SCI	SMT	21.6	2.7	2.7
WMT19	en-de	IT	NMT	13.4	1.0	1.0
	en-ru	Tech	NMT	15.0	1.0	1.0

Table 1: Statistics of various QE datasets. WMT18* contains random splits of the publicly available training data given that the official test sets are not publicly available.

2.2 Models

BERT We experiment with a strong neural QE approach based on BERT (Devlin et al., 2019). In particular, we focus on the *bert-base-cased* version of the multilingual BERT[2]. We join the source and translated sentences together using the special SEP token and convert the vector representation of the final CLS token to score via a Multilayer Perceptron (MLP) layer. Our models perform competitively to the state-of-the-art QE models (Kepler et al., 2019a; Kim et al., 2019). However, we do not treat this as a multitask learning problem where word-level labels are also needed because this is severely limited by the availability of data. We also do not do further optimizations (e.g. model ensembling) given that our focus is on what can be learned with the current data, and not maximizing performance. Our simpler models allow us to carefully analyze and determine the effects of source and translated sentences on the performance of the models. We expect the trends to be the same as other neural QE models.

[1] http://www.umiacs.umd.edu/ snover/terp/

[2] https://github.com/google-research/bert

QUEST We also trained and evaluated SVM regression models over 17 baseline features highly relevant to the QE task (Specia et al., 2013, 2015).

3 Results and Recommendations

3.1 Imbalanced datasets

Figure 2 presents the distributions of HTER scores for QE datasets from WMT 2018 and 2019.

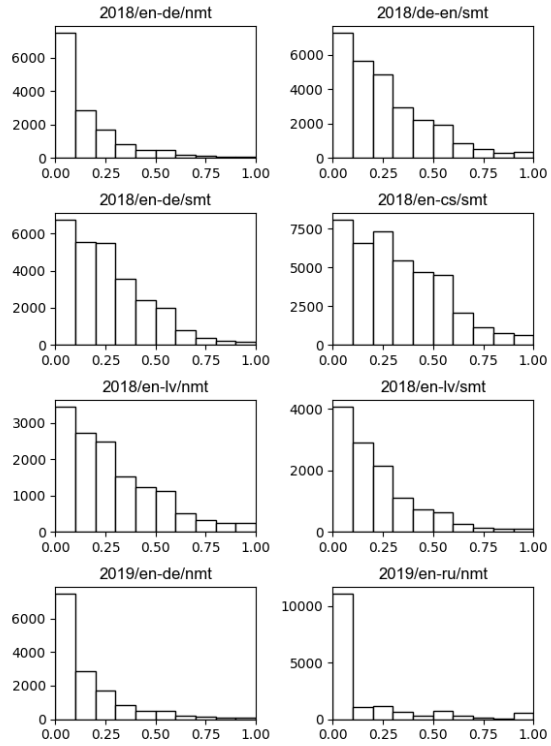

Figure 1: Histograms of HTER scores.

The distributions of quality scores are skewed towards zero, meaning most of the translated sentences require few or no post-editing. This phenomenon is especially true for the QE datasets from WMT2019, which are exclusively NMT-based, and for which the majority of the translated sentences have HTER scores of less than 0.1. When we examine the estimations from our QE models, we find that they rarely output values above 0.3, which implies that these models fail to capture sentences with low-quality scores. For example, 15.8% of the samples from the test set of WMT19 En-De have HTER scores above 0.3, yet a BERT QE model outputs scores above 0.3 on only 14.5% of those samples. In fact, our BERT model predicts scores above 0.3 on only 2.3% of the whole test set. This defeats the purpose of QE, especially when the objective of QE is to identity unsatisfactory translations.

Recommendation: To alleviate this issue, we recommend that QE datasets are balanced *by design* and that they include high-, medium- and low-quality translations. One way to ensure this would be to include models with different levels of quality.

3.2 Lexical artifacts

Table 3 shows some examples of the domain-specific lexical artifacts we found in en-de and en-cs datasets, although other datasets exhibit similar issues. Around 37% of translated sentences in En-De datasets contain the double inverted comma, and more than 70% of these sentences require little to no post-editing. A QE system can get strong performance simply by associating any translated sentences containing double inverted commas with low HTER scores.

These lexical artifacts are introduced when the lack of diversity in labels interacts with a lack of diversity in vocabulary and sentences. For example, the En-De dataset, which was sampled from an IT manual, contains many repetitive sentences similar to *"Click X to go to Y"*.

Recommendation: We can mitigate this problem by sampling source sentences from various documents across multiple domains.

3.3 Partial-input baselines

In Table 2 we report the average Pearson correlation over five different training runs of the same model.

We observe that the QE systems trained on partial inputs perform as well as systems trained on the full input. This is especially true for the systems that use BERT, which achieve 90% or more of the full performance on five out of eight test sets by only considering the target sentence. Additionally, QE systems trained on only source sentences consistently perform at the correlation of around 0.4. The partial-input problem is less significant on the feature-based SVM models, where only the partial-input systems trained on WMT18 SMT have higher than 85% performance. The strong performances on partial-inputs show that these datasets are *cheatable*, and QE systems trained on them would not generalize well (Feng et al., 2019). The partial-input baseline problem is also evident in the top-performing QE system from WMT 2019 (Kepler et al., 2019a): rather than using both

Dataset	langs	syst	SVM + 17 features			BERT		
			ρ	src (%)	tgt (%)	ρ	src (%)	tgt (%)
WMT18*	de-en	SMT	0.342	62.3%	57.6%	0.697	62.0%	81.2%
	en-cs	SMT	0.398	57.3%	79.9%	0.609	**88.2**%	**96.1**%
	en-de	NMT	0.290	63.4%	78.6%	0.456	**92.5**%	**88.4**%
		SMT	0.326	**113.2**%	**100.0**%	0.597	71.2%	**100.3**%
	en-lv	NMT	0.273	52.4%	60.8%	0.621	68.8%	77.3%
		SMT	0.311	38.6%	51.5%	0.509	82.5%	**93.9**%
WMT19	en-de	NMT	-	-	-	0.423	**94.6**%	**90.5**%
	en-ru	NMT	-	-	-	0.439	75.2%	**95.9**%

Table 2: Pearson correlation (ρ) between predictions from various QE models and gold HTER labels, and the percentage of performance obtained by presenting the model with partial input from only the source (src) or target (tgt) sentences. In **bold** we highlight instances with higher than 85% performance. Results for QUEST with the WMT19 data are omitted as feature sets for those datasets are not publicly available.

Dataset	markers	prev. (%)	H<0.1 (%)
WMT18/19 en-de	”	37.1	73.6
	>	7.1	88.8
	wählen	21.1	78.0
	klicken	13.2	82.8
WMT18 en-cs	gt	4.8	43.2
	&	4.8	43.0
	go	5.8	22.9
	www	0.8	43.9

Table 3: Top 4 lexical items ranked by NPMI for HTER in the range [0.0 - 0.1) and the prevalence % of sentences containing these words and with HTER (H) score of less than 0.1.

Dataset	langs	syst.	ρ_{test}	ρ_{adv}
WMT18*	en-de	SMT	0.597	0.030
		NMT	0.456	-0.017
	en-cs	SMT	0.609	0.047
	en-lv	SMT	0.509	0.012
		NMT	0.621	0.030
	de-en	SMT	0.697	0.014
WMT19	en-de	NMT	0.423	0.002
	en-ru	NMT	0.439	-0.036

Table 4: Pearson correlations on the original test sets (ρ_{test}) and adversarial test sets (ρ_{adv}) for the BERT-based models.

source and translated sentences, they obtain the best results on the word-level QE task by ignoring source sentences when making predictions on translated sentences and vice versa. Such counter-intuitive phenomenon violates the assumption that the quality scores of translated sentences are dependent on both the source and target sentences.

Recommendation: When designing and annotating QE datasets, we suggest using a metric that intrinsically represents both *fluency* and *adequacy* as labels, such as *direct assessments* (Graham, 2015) and ensure we have enough representation instances with high and low adequacy and fluency.

4 Discussion

Our results suggest that source sentences or translated sentences alone might already contain cues that correlate well with human-annotated scores in the QE datasets. Given this, it is highly unlikely that these QE models figure out how to model inter-dependencies between source and translated sentences, which usually require several levels of linguistic analysis. We hypothesize that QE models rely on either the complexity of source sentences or the fluency of translated sentences, but not on adequacy, to make their predictions. To test this, we create adversarial test sets across all language directions by randomly shuffling the source sentences and changing HTER scores to 1.0. A good model should be able to assign high HTER scores to mismatched pairs.

In Table 4, we show the Pearson correlations on the adversarial sets. As expected, our QE models perform poorly, getting correlations close to zero. The results confirm our suspicion: systems trained on these datasets fail to model adequacy. They assign high scores to fluent translations or source sentences with low complexity, regardless of whether these translated sentences are semantically related to their corresponding source or translated sentences.

5 Conclusions and future work

In this work, we presented our analysis of QE datasets used in recent evaluation campaigns. Although recent advances in pre-trained multilingual language models significantly improve performances on these benchmark QE datasets, we highlight several instances of *sampling bias* embedded in the QE datasets which undermine the apparent successes of the newer QE models. We identified (*i*) issues with the balance between high- and low- quality instances (*ii*) issues with the lexical variety of the test sets and (*iii*) the lack of robustness to partial input. For each of these problems, we proposed recommendations.

Upon the submission of this paper, we implemented the proposed recommendations and created a new dataset for quality estimation. We believe addresses the limitations in current datasets. More specifically, we collected data for six language pairs, namely two high-resource languages (English–German and English–Chinese), two medium–resource languages (Romanian–English and Estonian–English), and two low-resource languages (Sinhala–English and Nepali–English). Each language pair contains 10,000 sentences extracted from Wikipedia and translated by state-of-the-art neural models, manually annotated for quality with direct assessment (0-100) by multiple annotators following industry standards for quality control. An example is shown in figure 3.

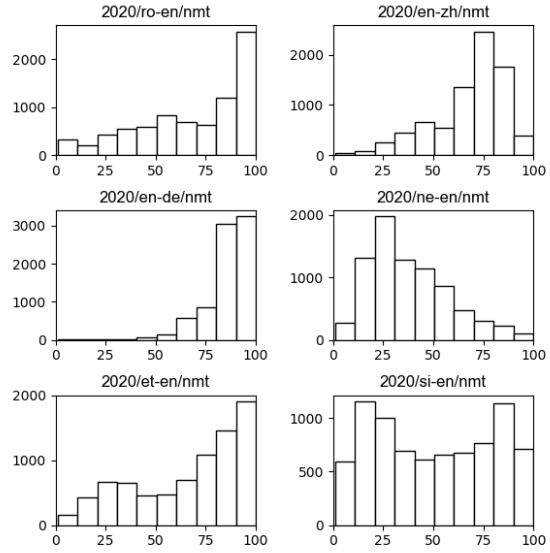

Figure 2: Histograms of DA scores for the MLQE dataset.

The selection of languages with varying degrees of resource availability leads to more diverse translation quality distributions (particularly for the medium-resource languages), mitigating the issue of imbalanced datasets.

The choice of data source – articles in a multitude of topics from Wikipedia – will lead to more diverse vocabulary and constructs, mitigating the issue of lexical artifacts. The lexical diversity of our new dataset is further supported by its average type-token ratio (TTR)[3] of 0.166, which is a 417% increase from the average TTR of the QE dataset from WMT 2018 and a 259% increase from the average TTR of the QE dataset from WMT 2019. The annotation according to direct assessment, which balances between adequacy and fluency, will mitigate the problems associated with the sampling bias and the lack of balance between low and high-quality translations.

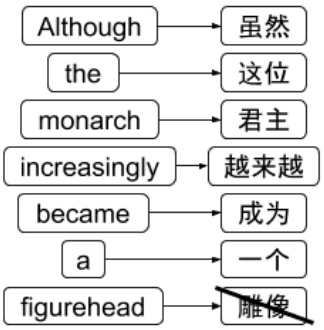

Figure 3: An English-Chinese sentence pair from our new dataset. The translated Chinese sentence is fluent but inadequate because the final token is mistranslated to ”statue” instead of ”figurehead”, and thus the original semantic meaning of the source sentence is changed. Our annotators collectively assigned it a low score of 24.0. However, HTER would misclassify it as a good translation since there is only one token that requires post-editing.

This dataset, named MLQE, has been released to the research community[4] and will be used for the WMT2020 shared task on Quality Estimation.[5] In future work, we will test the partial input hypothesis on this data and hope it will be useful to further research in quality estimation, leading to more reliable models.

[3]This is the average TTR of English sentences from the train and dev set of every language direction.

[4]https://github.com/facebookresearch/mlqe

[5]http://www.statmt.org/wmt20/quality-estimation-task.html

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.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. *arXiv preprint arXiv:1905.05778*.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Fábio 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*.

Fábio 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, 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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

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

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, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems.