Quality Estimation via Backtranslation at the WMT 2022 Quality Estimation Task

Sweta Agrawal*
University of Maryland
sweagraw@umd.edu

Nikita Mehandru*
University of California, Berkeley
nmehandru@berkeley.edu

Niloufar Salehi
University of California, Berkeley
nsalehi@berkeley.edu

Marine Carpuat
University of Maryland
marine@umd.edu

Abstract

This paper describes submission to the WMT 2022 Quality Estimation shared task (Task 1: sentence-level quality prediction, Zerva et al. (2022)). We follow a simple and intuitive approach: estimating MT quality by automatically back-translating hypotheses into the source language using a multilingual MT system. Using standard MT evaluation metrics, we then compare the resulting backtranslation with the original source. We find that even the best-performing backtranslation-based scores perform substantially worse than supervised QE systems, including the organizers' baseline. However, combining backtranslation-based metrics with off-the-shelf QE scorers improves correlation with human judgments, suggesting that they can indeed complement a supervised QE system.

1 Introduction

Sophisticated approaches to MT quality estimation (QE) based on large pre-trained models and careful training regimen have enabled great progress in recent years. However, when using online MT systems, such QE technology is not yet available to users and backtranslation provides an appealingly simple strategy to estimate translation quality whether by humans or by automated sytems. Lay users often rely on backtranslation to assess MT quality in languages that they do not understand (Somers, 2005; Mehandru et al., 2022). As a result, from a user experience standpoint, using backtranslation for QE is easy to explain. Furthermore, with the increasing popularity of multilingual neural MT systems that can easily translate between multiple language pairs in any direction, backtranslations are very cheap to obtain, since they do not even require training an auxiliary MT system in the reverse translation direction.

However, the effectiveness of backtranslation for estimating the quality of MT remains unclear. In early rule-based and statistical MT systems, Somers (2005) shows that, when using automatic evaluation methods (e.g., BLEU), backtranslation cannot discriminate good MT systems from bad ones, nor between texts that are easy or hard to translate. This led him to conclude that "round trip translation [is] good for nothing". Recently, Moon et al. (2020) revisited the use of backtranslation for QE with neural systems for MT and with embedding-based similarity metrics to enable a more sophisticated comparison of the backtranslation with the source. They obtained strong results on the WMT 2019 QE task, outperforming the YISI-2 metric (Lo, 2019) on system-level evaluations, but exhibited rather low correlations on the segment-level task which is more directly aligned with how humans use BT to gauge MT quality.

The goal of our submission is to pitch a backtranslation-based QE score that can complement state-of-the-art quality estimation systems in the controlled settings of the WMT shared task (Zerva et al., 2022) and understand its reliability as a sentence-level quality estimation technique.

2 Approach

Following Moon et al. (2020), given a source sentence x and a MT hypothesis, we translate y back into the source language using an off-the-shelf multilingual model M, yielding backtranslation $\tilde{x}$. We then compare x and $\tilde{x}$ using standard machine translation evaluation metrics, and hypothesize that the distance between x and $\tilde{x}$, referred to as **BT-score(x, $\tilde{x}$),** can be an indicative of the translation quality of y.

However, MT systems are prone to making errors and are shown to hallucinate content. When the BT system makes an error, it can misguide the users in believing that the translation is a) erroneous when it is not and b) correct when the BT system magically recovers the source content. In order to improve the reliability of the BT-based QE

* equal contribution.

BT Metrics	Footprint Bytes	Params.	Development Set Pearson	Development Set Spearman	Test Set Pearson	Test Set Spearman
BLEU	0	0	0.179	0.170	0.141	0.137
chrF	0	0	0.203	0.181	0.184	0.174
BERTScore	0	177853440	0.292	0.296	0.325	0.285
Baseline[1]	2280011066	564527011	n/a	n/a	0.560	0.576

Table 1: Pearson and Spearman correlation between backtranslation-based QE metrics and Direct Assessment judgments on the WMT 2022 En-Cs task.

Metrics	En-Cs (DA) Dev	En-Cs (DA) Test	En-Ru (MQM) Dev	En-Ru (MQM) Test	Zh-En (MQM) Dev	Zh-En (MQM) Test
[1] BT-BERTScore	0.296	0.285	0.262	0.210	0.151	0.249
[2] Comet-Src	0.461	0.519	0.505	0.383	0.213	0.223
Multiply([1], [2])	**0.467**	**0.523**	**0.512**	**0.390**	**0.216**	**0.257**
Baseline[2]	n/a	0.560	n/a	0.330	n/a	0.164

Table 2: Spearman correlation between QE metrics and human judgments on the WMT 2022 Sentence Level Quality Estimation task: Combining BT-BERTScore and Comet-Src improves correlation with human judgments across the board.

metrics, **BT-score**(x, $\tilde{x}$), and to understand whether they can complement off-the-shelf QE scorers that directly estimate the quality of a source sentence and a MT hypothesis, **FT-score**(x, y), we also propose to combine the two evaluation methods using a simple multiplication ("AND") operation.

Back-translation Model The backward translations were generated from Facebook's mBART-50 Many-to-One and One-to-Many multilingual machine translation (MMT) models. The MMT model can translate between 49 languages into and out of English, and uses 12 layers with 1,024 sized embeddings, 4,096 feedforward neural network (FNN) embedding dimensions, and 16 heads for both encoder and decoders.[3]

MT Evaluation Metrics We experiment with model-free and model-based evaluation metrics. We apply the following sentence-level scores to compare detokenized backtranslations $\tilde{x}$ with the source x:

- BLEU: we use the Sacrebleu implementation of sentence-level BLEU, with an exponential decay smoothing.[4] (Papineni et al., 2002)
- chrF: we use the Sacrebleu implementation of the chrF score, which takes a maximum character n-gram order count of six and calculates the number of ngram overlap between hypothesis and reference n-grams. (Popović, 2015)
- BERTScore: we compute the F-score based on wordpiece-level embedding similarities of, weighted by inverse document frequency (idf), using BERT as the embedding model (Zhang et al., 2019).[5].

We use the publicly available QE metric, Comet-Src ("wmt21-comet-qe-mqm") to compute FT-score(x, y).

3 Official Results using BT-based Metrics

We evaluate our approach on the English-Czech sentence-level quality prediction subtask. As our approach is unsupervised, we do not use the training data provided by the organizers. We report results obtained on the development and test sets, using the Pearson and Spearman correlations with human judgments of quality.

[3] https://huggingface.co/facebook/mbart-large-50-many-to-one-mmt/, https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt/

[4] https://github.com/mjpost/sacrebleu

[5] https://pypi.org/project/bert-score/

	$DA >= -1$	$DA < -1$	$DA >= 0$	$DA < 0$	$DA >= 1$	$DA < 1$
BT-BERTScore	0.197	**0.230**	0.133	0.222	0.022	0.235
Comet-Src	**0.397**	0.139	**0.337**	**0.313**	**0.139**	**0.413**

Table 3: En-Cs segment-level correlation in different quality buckets according to the direct assessment scores.

Sample	**Development Set**			
	z-mean	BT-BLEU	BT-chrF	BT-BERT
Source: Arif Lohar briefly went into acting in punjabi movies before returning to his music career at the age of 22 . **Output:** Arif Lohar krátce začal hrát v Punjabi filmech , než se v roce 22 vrátil ke své hudební kariéře . **BT Source:** Arif Lohar briefly began acting in Punjabi films before returning to his musical career in the year 22.	-1.486	20.95	62.57	0.949
Source: Promulgate Thai Royal and noble titles back and return the title to politician who was canceled . **Output:** Promulgate Thajské královské a šlechtické tituly zpět a vrátit titul politici , který byl zrušen . **BT Source:** Promulgate Thai royal and noble titles back and return the title of politician that was abolished.	-1.781	48.34	73.94	0.959
Source: Ika-6 na utos ; re - runs ; aired on gma life tv for the first time ; replacing I heart davao . **Output:** Ika-6 na utos ; re - runs ; poprvé vysíláno na gma life TV ; nahrazuje I heart davao . **BT Source:** Ika-6 on utos; re-runs; first broadcast on gma life TV; replaces I heart davao.	-2.935	18.00	53.63	0.941

Table 4: Three randomly sampled sentences from the bottom 5% according to DA scores.

As can be seen in Table 4, BERTScore provides a better correlation with human judgments than BLEU and chrF consistently on the development and test sets. This is expected since the underlying BERT model provides a more semantically informed comparison than n-gram metrics. However, the backtranslation metrics yield low correlation scores overall, underperforming the organizer's baseline on the test set.

Our results are complementary to Moon et al. (2020) in that they suggest that BT-based metrics might be better suited to ranking diverse outputs from systems of varying overall quality, than those from a single MT system, i.e. at predicting quality assessments at the segment level.

4 Can BT-based scorers complement existing QE metrics?

While standalone evaluation using BT-based scoring significantly lags behind supervised SOTA QE baselines, we evaluate whether BT-based metrics can provide reliable complementary judgments to a supervised off-the-shelf QE scorer in Table 2. We combine the best BT-based scorer, BT-BERTscore and a standard QE scorer, Comet-Src using a simple multiplication operation. On three sentence level quality estimation tasks: En-Cs (DA), En-Ru (MQM) and Zh-En (MQM), combining both BT and QE scores result in improved correlation across the board over individual metrics, outperforming baselines on both En-Ru and Zh-En.

In order to better understand the source of this improvement, we divide the En-Cs development dataset into different buckets based on the direct assessment scores and report correlation on the resulsubsets in Table 3. On very bad quality translations, i.e. $DA <= -1$, BT-BERTScore exhibits a higher correlation than Comet-Src, suggesting that it is able to more reliably distinguish between bad translations than Comet-Src, hence complementing the QE metric.

5 Qualitative Analysis on En-Cs

In Table 2, we randomly sampled three sentences from the lowest 5% of the human direct assessment scores from the development set data and report the corresponding BT-BLEU, BT-chrF, and BT-BERTScores. The outputs depict how the forward translation output can be of poor quality, as indicated by the human direct assessment scores. However, the semantic similarity between the source and the back-translated source can still suggest that the forward translation is correct. When we apply machine translation to other domains, this can be problematic and misleading since users may mistakenly impart higher trust levels when using backtranslation techniques. From the same table, we can also observe that the automatic metrics cannot capture salient errors as suggested by the high scores generated by the automatic metric for the second example ("who was canceled" vs "that was abolished"). This finding is in line with prior work that has shown a positive correlation between *human evaluations* conducted on input sentences and translated outputs with *human evaluations* on input sentences and round-trip sentences (Aiken and Park, 2010). These results together call for a more systematic assessment of the role of backtranslation in lay users perceptions of MT quality.

6 Conclusion

We evaluated backtranslation-based unsupervised quality estimation systems on the sentence-level quality estimation task. Our results show that backtranslation bases scorers fall substantially behind supervised models such as the organizers' baseline. However, they can complement off-the-shelf QE metrics in distinguishing bad translations. Qualitative analysis on En-Cs indicates that while backtranslation can be a poor indicator of translation quality, the automatic metrics derived using the source and the backtranslated source might also add to the unreliability of the scorer. This suggests that more investigation is needed to determine whether backtranslation can be used effectively for QE in practical systems, whether for automatic quality estimation or to provide quality feedback to human users.

7 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2147292.

References

Milam Aiken and Mina Park. 2010. The efficacy of round-trip translation for mt evaluation. *Translation Journal*, 14(1):1–10.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, South Korea)(FAccT'22). Association for Computing Machinery, New York, NY, USA*.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Harold Somers. 2005. Round-trip Translation: What Is It Good For? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133, Sydney, Australia.

Chrysoula Zerva, Frederic Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, Andre F. T. Martins, and Lucia Specia. 2022. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.