Unsupervised Machine Translation Quality Estimation in Black-box Setting

Hui Huang[1*], Hui Di[2], Jin'an Xu[1†], Kazushige Ouchi[2] and Yufeng Chen[1]

[1]Beijing Jiaotong University, Beijing, China
[2]Toshiba (China) Co., Ltd., Beijing, China
{18112023,jaxu,chenyf}@bjtu.edu.cn
{dihui,kazushige.ouchi}@toshiba.com.cn

Abstract. Machine translation quality estimation (Quality Estimation, QE) aims to evaluate the quality of machine translation automatically without golden reference. QE is an important component in making machine translation useful in real-world applications. Existing approaches require large amounts of expert annotated data. Recently, there are some trials to perform QE in an unsupervised manner, but these methods are based on glass-box features which demands probation inside the machine translation system. In this paper, we propose a new paradigm to perform unsupervised QE in black-box setting, without relying on human-annotated data or model-related features. We create pseudo-data based on Machine Translation Evaluation (MTE) metrics from existing machine translation parallel dataset, and the data are used to fine-tune multilingual pre-trained language models to fit human evaluation. Experiment results show that our model surpasses the previous unsupervised methods by a large margin, and achieve state-of-the-art results on MLQE Dataset.

Keywords: Machine Translation, Unsupervised Quality Estimation, Pre-trained Language Model.

1 Introduction

In recent years, with the development of deep learning, Machine Translation (MT) systems made a few major breakthroughs and were wildly applied. Machine translation quality estimation (Quality Estimation, QE) aims to evaluate the quality of machine translation automatically without golden reference [1]. The quality can be measured with different metrics, such as HTER (Human-targeted Edit Error) [2] or DA (Direct Assessment) Score [3].

Previous methods treat QE as a supervised problem, and they require large amounts of in-domain translations annotated with quality labels for training [4][5]. However, such large collections of data are only available for a small set of languages in limited domains.

---

* Work was done when Hui Huang was an intern at Research and Develop Center, Toshiba (China) Co., Ltd., China.

† Jinan Xu is the corresponding author.

Recently, Fomicheva [6] firstly performs QE in an unsupervised manner. They explore different information that can be extracted from the MT system as a by-product of translation, and use them to fit quality estimation output. Since their methods are based on glass-box features, they can only be implemented in limited situations and demands probation inside the machine translation system.

In this work, we firstly propose to perform unsupervised QE in a black-box setting, without relying on human-annotated data or model-related features. We create pseudo-data based on Machine Translation Evaluation (MTE) metrics, such as BLEU, HTER and BERTscore, from publicly-accessible translation parallel dataset. The MTE-metrics based data are then used to fine-tune several multilingual pre-trained language models, to evaluate translation output.

To the best of our knowledge, this is the first work to utilize MTE methods to deal with QE. Our method does not involve complex architecture engineering and easy to implement. We performed experiment on two language-pairs on MLQE[1] Dataset, outperforming Fomicheva by a large margin. We even outperformed two supervised models of Fomicheva, revealing the potential of MTE-based methods for QE.

2 Background

2.1 Machine Translation Evaluation

Similar to QE, Machine Translation Evaluation (MTE) also aims to evaluate the machine translation output. The difference between MTE and QE is that MTE normally requires annotated references, while QE is performed without reference and highly relies on source sentences.

Human evaluation is often the best indicator of the quality of a system. However, designing crowd sourcing experiments such as Direct Assessment (DA) [3] is an expensive and high-latency process, which does not easily fit in a daily model development pipeline.

Meanwhile automatic metrics, for example BLEU [7] or TER [2], can automatically provide an acceptable proxy for quality based on string matching or hand-crafted rules, and have been used in various scenarios and led the development of machine translation. But these metrics cannot appropriately reward semantic or syntactic variations of a given reference [8].

Recently, after the emergence of pre-trained language models, a few contextual embedding based metrics have been proposed, such as BERTscore [8] and BLUERT [9]. These metrics compute a similarity score for the candidate sentence with the reference based on token embeddings provided by pre-trained models. Refraining from relying on shallow string matching and incorporate lexical synonymy, BERTscore can achiev higher relevance with human evaluation.

Given the intrinsic correlation nature of MTE and QE, few works have been done to leverage MTE methods to deal with the task of QE.

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

2.2 Machine Translation Quality Estimation

Despite the performance of machine translation systems is usually evaluated by automatic metrics based on references, there are many scenarios where golden reference is unavailable or hard to get. Besides, reference-based metrics also completely ignore the source segment [10]. This leads to pervasive interest on the research of QE.

Early methods referred to QE as a machine learning problem [11]. Their model could be divided into the feature extraction module and the classification module. Highly relied on heuristic artificial feature designing, these methods did not manage to provide reliable estimation results.

During the trending of deep learning in the field of natural language processing, there were also a few works aiming to integrate deep neural network into QE systems. Kim [12] proposed for the first time to leverage massive parallel machine translation data to improve QE results. They applied RNN-based machine translation model to extract high-quality feature. Fan [13] replaced the RNN-based MT model with Transformer and achieved strong performance.

After the emergence of BERT, there were a few works to leverage pretrained models on the task of QE [14][15]. Language models pre-trained on large amounts of text documents are suitable for data-scarce QE task by nature, and have led to significant improvements without complex architecture engineering.

Despite most models relied on artificial annotated data, there were also a few trials aiming to apply QE in an unsupervised manner. The most important work is Fomicheva [6], which proposed to fit human DA scores with three categories of model-related features: A set of unsupervised quality indicators that can be produced as a by-product of MT decoding; the attention distribution inside the Transformer architecture; model uncertainty quantification captured by Monte Carlo dropout. Since these methods are all based on glass-box features, they can only be applied in limited scenarios where inner exploration into the MT model is possible.

3 Model Description

3.1 Pretrained Models for Quality Estimation

Our QE predictor is based on three different pre-trained models, namely BERT [16], XLM [17], and XLM-R [18], as shown in Figure 1.

Given one source sentence and its translated result, our model concatenates them and feeds them into the pre-trained encoder. To leverage the global contextual information when doing sentence-level prediction, an extra layer of bidirectional recurrent neural network is applied on the top of the pre-trained model.

Despite the shared multilingual vocabulary, BERT is originally a monolingual model [19], pretrained with sentence-pairs from one language or another. To help BERT adapts to our bilingual scenario, where the inputs are two sentences from different languages, we implement a further pre-training step.

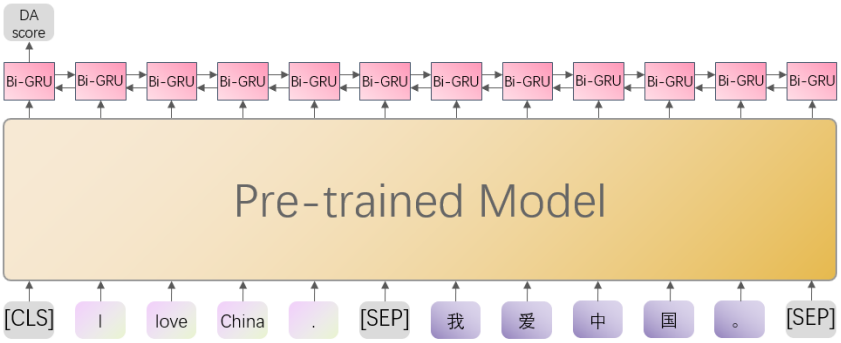

Fig. 1. Pre-trained Model for Quality Estimation.

During the further pre-training step, we combine bilingual sentence pairs from large-scale parallel dataset, and randomly mask sub-word units with a special token, and then train BERT model to predict masked tokens. Since our input are two parallel sentences, during the predicting of masked words given its context and translation reference, BERT can capture the lexical alignment and semantic relevance between two languages.

In contrast, XLM and XLM-R are multilingual models by nature, which receive two sentences from different languages as input during training, that means a further pre-training step is redundant. The training strategies and data of XLM and XLM-R are designed distinctly, which are explained in detail in their papers.

3.2 MTE-based QE Data

Despite sentence-pairs with source and machine-translated text readily accessible (for which we only need to translate source text into target language using a MT system), the absence of DA scores becomes our biggest challenge. Even in supervised scenario, human-annotated DA scores are still scarce and limited [5]. Therefore, we propose to use MTE metrics to fit human assessment, thus creating massive pseudo data for the training of the QE system. Our approach can be described as follows:

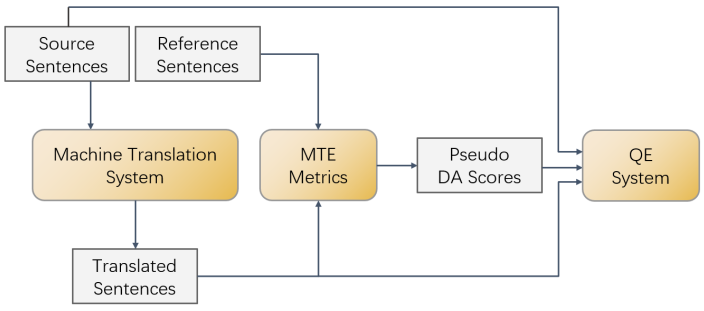

Fig. 2. MTE-metrics based QE training procedure.

Firstly, we decode source sentences in parallel corpus into target language. Secondly, we use automatic MTE-metrics to evaluate the quality of output sentences based on references. In this step we do not need any human annotation or time-consuming training. The MTE based evaluation can give a roughly accurate quality assessment, and can be used as substitution to human-annotated DA scores. Thirdly, the pseudo DA scores, combined with source and translated sentence pairs, are used to train our QE system.

We tried three different MTE metrics to fit DA evaluation, namely TER [2], BLEU [7], and BERTscore [8].

TER uses word edit distance to quantify similarity, based on the number of edit operations required to get from the candidate to the reference, and then normalizes edit distance by the number of reference words, as shown in Equation 1.

$$TER = \frac{\#\ of\ edits}{average\ \#\ of\ reference\ words} \tag{1}$$

BLEU is the most widely used metric in machine translation. It counts the number of n-grams that occur in the reference sentence and candidate sentence. Each n-gram in the reference can be matched at most once, and very short candidates are discouraged using a brevity penalty, as shown in Equation 2.

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n), BP = \begin{cases} 1 & if\ c>r \\ e^{(1-r/c)} & if\ c\leq r \end{cases} \tag{2}$$

where p_n denotes the geometric average of the modified n-gram precisions, w_n denotes positive weight for each token, c denotes the length of the candidate translation and r denotes the effective reference corpus length.

BERTscore calculates the cosine similarity of a reference token and a candidate token based on their contextual embedding provided by the pre-trained model. The complete score matches each token in reference to a token in candidate to compute recall, and each token in candidate to a token in reference to compute precision, and then combine precision and recall to compute an F1 measure, as displayed in the following equations.

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^{\mathrm{T}} \hat{x}_j \tag{3}$$

$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^{\mathrm{T}} \hat{x}_j \tag{4}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{5}$$

where x and $\hat{x}$ denote the contextual embedding for each token in reference and candidate sentences, respectively.

With the ability of matching paraphrases and capturing distant dependencies and ordering, BERTscore is proved to be highly correlated with human assessment [8].

4 Experiment

4.1 Setup

Dataset. The dataset we use is MLQE Dataset [6], which contains training and development data for six different language-pairs. We performed our experiments mainly on two high-resource languages (English–Chinese and English–German). Since we want to solve the problem in unsupervised setting, we only used the 1000 sentence-pairs from the development data for each direction respectively.

To train our own MT model, we use the WMT2020 English-Chinese and English-German data[2], which contains roughly 10 million sentence-pairs for each direction after cleaning (a large proportion is reserved to generate QE data).

Fomicheva also provide the MT model which was used to generate their QE sentence pairs, thus we have two different MT models to use. We will explain the influence of different MT models in the next section.

For fine-tuning pre-trained models, we used the reserved data from WMT2020 English-Chinese and English-German translation, and randomly sampled 500k sentence pairs for each direction to create MTE-based QE data.

Baseline. Sine there are few works done in the area of unsupervised QE, we mainly make comparison with Fomicheva. They proposed 10 methods which can be categorized as three sets, among them we display their top-two results in each direction, namely D-Lex-Sim and D-TP for English-Chinese, and D-TP and Sent-Std for English-German.

We also make comparison with supervised methods, including PredEst models using the same parameters in the default configurations provided by Kepler [14], and the recent SOTA QE system BERT, augmented with two bidirectional RNN [15]. These two models are trained with the provided 7000 training pairs.

4.2 Experiment Results

As shown in Table 1 and Table 2, our approach surpasses Fomicheva with their best-performance methods by a large margin on both directions, verifying the effectiveness of MTE-based QE data. We even outperform BERT-BiRNN trained in supervised manner on both directions.

Although the supervised training data provided is limited and our best results are achieved by XLM rather than BERT (we will explain this in next section), the result is still very fascinating.

The glass-box features, although thoroughly explored by Fomicheva, seem unhelpful compared with MTE-metrics based methods. These features are no more than statistic cues regulated by the machine translation model. If we rely on the same MT

[2] http://www.statmt.org/wmt20/translation-task.html

model to evaluate the translation, then we will be constrained by itself and unable to cope with various phenomena.

Table 1. Experiment results on English-Chinese MLQE Dataset.

Language Direction	Method	Pearsonr	Spearman
English-Chinese	PredEst	0.190	—
	BERT-BiRNN	0.371	—
	D-Lex-Sim	0.313	—
	D-TP	0.321	—
	TER	0.3919	0.4116
	BLEU	0.3668	0.3941
	BERTscore-precision	0.4254	0.4347
	BERTscore-F1	**0.4288**	0.4373

Table 2. Experiment results on English-German MLQE Dataset.

Language Direction	Method	Pearsonr	Spearman
English-German	PredEst	0.145	—
	BERT-BiRNN	0.273	—
	D-TP	0.259	—
	Sent-Std	0.264	—
	TER	0.2589	0.2828
	BLEU	0.2637	0.2931
	BERTscore-precision	**0.3124**	0.3327
	BERTscore-F1	0.3089	0.3284

Moreover, we can also conclude that when fine-tuning pre-trained models for QE task, the quantity of data is more important than the quality of data, as shown in Figure 3. Although our data is generated purely based on automatic metrics rather human annotators, we can still surpass supervised systems trained only with clean data.

Among our three methods, BERTscore-based methods achieve better results than statistical metrics-based methods, which is reasonable since BERTscore is proved to better correlate with human assessment. More accurate MTE metrics could lead to more natural pseudo data, therefore enable the QE model to perform better.

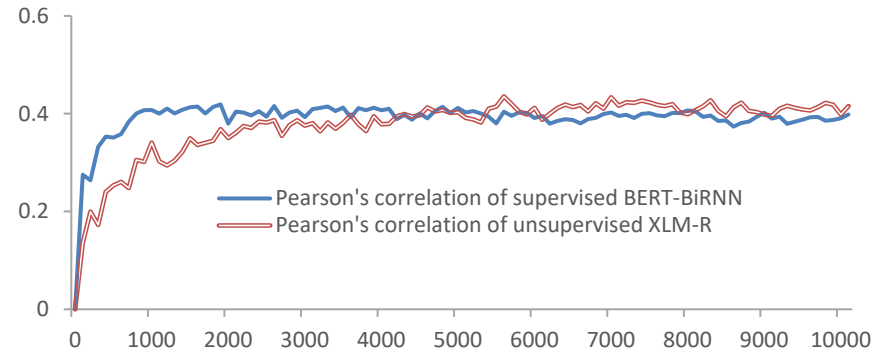

Fig. 3. The variation of Pearson's correlation coefficient with the increase of the training step. Although the supervised model could generate better results in the first few steps, as the unsupervised model receives more data after more steps, it would outperform the supervised model.

5 Analysis

5.1 Is BERT always the Best?

Despite the overwhelming results BERT has accomplished on multiple datasets, our scenario demands the ability to process bilingual input, while BERT is originally a monolingual model, treating the input as either being from one language or another.

In contrast, XLM and XLM-R are multilingual models by nature, pre-trained with bilingual inputs from different languages. Since QE task aims to evaluate the translation based on the source sentence from another language, XLM and XLM-R should be more suitable. Experiment results in Table 3 verify our hypothesis.

Table 3. Experiment results on MLQE Direct Assessment data.

Language Direction	Method	Pretrained Model	Pearonr	Spearman
English-Chinese	BERTscore-precision	BERT	0.3255	0.3295
		BERT(further-trained)	0.3827	0.3895
		XLM	**0.4254**	0.4347
		XLM-R	0.4170	0.4227
	BERTscore-F1	BERT	0.3271	0.3329
		BERT(further-trained)	0.3836	0.3889
		XLM	0.4110	0.4221
		XLM-R	**0.4288**	0.4373

Even augmented by further pre-training steps with bilingual input in our experiment, BERT is still not competitive in multilingual scenarios. Multilingual pre-trained models are more suitable than BERT on QE task.

5.2 Is Black-box Model Necessary?

While we cannot explore the internal structure of MT model in black-box setting, the input and output of the model are still available. Therefore, when creating source-translation sentence pairs, we can choose to use our own model or the provided black-box model.

Nowadays, the neural-based (especially Transformed-based) MT architecture has dominated the machine translation area [20]. Different NMT systems trained with similar data may behave similarly to the same input [21].

Therefore, even with another model trained with slightly different data, the generated translation may still have similar error distribution. Experiment results displayed in Table 4 verify our hypothesis.

Table 4. Results of different data generated by different MT models.

Language Direction	Method	MT Model	Pearonr	Spearman
English-Chinese	TER	Ours	0.3671	0.3752
		Provided	0.3919	0.4116
	BLEU	Ours	0.3485	0.3619
		Provided	0.3668	0.3941
	BERTscore-precision	Ours	0.3853	0.3998
		Provided	0.4254	0.4227
	BERTscore-F1	Ours	0.3995	0.4133
		Provided	0.4288	0.4373

While the data generated by the provided model does obtain higher correlation, the result obtained by our own model is yet competitive. When creating MTE-based QE data, the provided model can benefit a lot, but if it is not available, we can simulate its error distribution with similar architecture and similar training data.

5.3 Where is the limitation of QE?

In this section, we would like to perform a case-study based on our results on development set. Since the distributions of our system's output and the real-world QE scores differ a lot, as shown in Figure 1, we mainly compare the ranking for the same sentence in different methods. Namely, we would rank the whole development set according to scores provided by our system and the golden label, and compare the discrepancy of ranking for the same sentence in different systems.

In summary, there are two problems impede the performance of our model.

Firstly, our model relies too much on the syntactic consistency while ignoring semantic understandability to evaluate a translation. Given a translated sentence with syntactically consistent structure, our model would assign a very high score even when the translation is semantically erroneous.

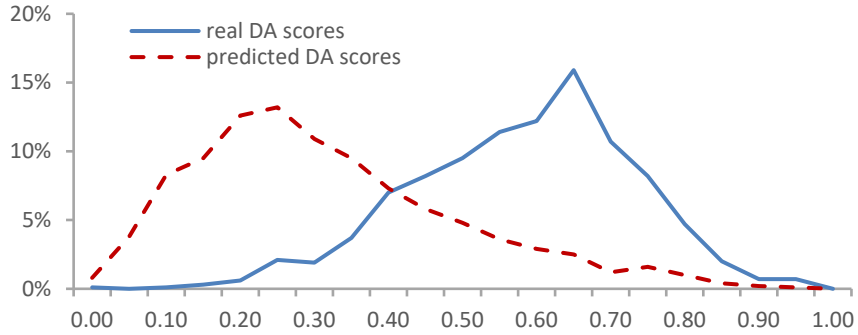

Fig. 4. Distribution of DA scores on development set. Solid line denotes the output of our system, and dashed line denotes the golden labels.

Table 5. Wrong prediction caused by syntactical inconsistency.

Source	Translation 1	Translation 2
A snob, a sneak and a coward, with very few redeeming features.	一个卑鄙的人, 一个偷偷摸摸的人, 一个懦弱的人, 几乎没有什么可取之处。(ranking 993 of 1000)	一个卑鄙, 一个偷偷摸摸, 一个懦弱, 几乎没有什么可取之处。(ranking 837 of 1000)
Others befriended and watched over the peasantry;	另一些人亲密无间地守护着农民；(ranking 763 of 1000)	另一些人做朋友并且守护着农民；(ranking 631 of 1000)

As shown in Table 5, although Translation 2 is much better than Translation 1, our method would still assign a higher evaluation score for Translation1 since the syntactic structure is more consistent.

This problem originates from pre-trained models themselves, as it is very likely for pre-trained models to rely on spurious statistical cues when doing prediction [22], while not really understand the sentence meaning. Most sentence pairs with a consistent syntactic structure are assigned with a higher score in our training data, which is captured by our model and used as an inappropriate criterion for evaluation.

The second problem is that our system fails to detect erroneously translated words, especially when prior knowledge is in need.

Table 6. Wrong prediction caused by mistranslated words.

Source	Translation
In 586 BCE, King Nebuchadnezzar II of Babylon conquered Judah.	巴比伦国王尼布查德尼扎尔二世征服了犹大. (ranking 12 of 1000)
Roman satirists ever after referred to the year as "the consulship of Julius and Caesar."	罗马讽刺家后来把这一年称为 "朱利叶斯和凯撒的领馆" 。(ranking 225 of 1000)

As shown in Table 6, for the first sentence, the provided model mistranslated the word *Judah*, which is a country, as a name. And in the second sentence, the word *consulship*, which refers to a period, is mistranslated as a building. To understand why these words are mistranslated, you may need related history knowledge.

The mistranslation of these key information makes the whole sentence beyond understanding, but since there is no grammatic error and the syntactic structure is appropriate, our model refers to them as good translations.

For the first problem, we believe it can be alleviated by strategically picked training samples, with more sentence-pairs syntactically inconsistent but semantically correct. We will leave this as our future work.

Since both QE model and MT model are based on deep-learning, QE can barely solve these problems which MT model cannot solve. More training data may help to alleviate this problem, but can hardly solve it, as more training data does not really introduce structured prior knowledge. We believe this is the limitation of QE.

6 Conclusion

Machine translation quality estimation (Quality Estimation, QE) aims to evaluate the quality of machine translation automatically without reference provided. Despite it has attracted a lot of research interest recent years, few works have been done to deal with QE in an unsupervised manner.

In this paper, we have devised an unsupervised approach to QE where we do not rely on any glass-box features. We create massive pseudo data based on automatic machine translation evaluation (MTE) metrics such as BLEU, TER and BERTscore, from publicly accessible machine translation parallel dataset. Then we use the MTE-based QE data to fine-tune multilingual pre-trained models, to predict direct assessment (DA) scores. Our approach surpassed previous unsupervised methods by a large margin, and even surpassed supervised methods, proving the effectiveness of incorporating MTE metrics into QE.

Despite the lack of human-annotated DA scores, the MTE metrics can provide a highly reliable evaluation for machine translated sentences, and enable us to perform QE in an unsupervised way. We will continue to explore the application of MTE in QE models, and try to reach the limitation of deep-learning based QE.

Acknowledge

This work is supported by the National Natural Science Foundation of China (Contract 61976015, 61976016, 61876198 and 61370130), and the Beijing Municipal Natural Science Foundation (Contract 4172047), and the International Science and Technology Cooperation Program of the Ministry of Science and Technology (K11F100010), and Toshiba (China) Co., Ltd.

References

1. John, B., Erin, F., George, F., Simona, G., Cyril, G., Alex, K., Alberto, S., Nicola, U.: Confidence estimation for machine translation. In: Proceedings of the International Conference on Computational Linguistics, page 315 (2004).
2. Matthew, S., Bonnie, D., Richard, S., Linnea, M., John, M.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas, Vol. 200, No. 6 (2006).
3. Yvette, G., Timothy, B., Alistair, M., Justin, Z.: Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering, pages 1–28 (2015).
4. Hyun, K., Jong-Hyeok L., Seung-Hoon N.: Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers, pages 562–568 (2017).
5. Erick, F., Lisa, Y., André, M., Mark, F., Christian, F.: 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 (2019).
6. Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzman, F., Fishel, M., Aletras, N., Chaudhary, V., Specia, L.: Unsupervised Quality Estimation for Neural Machine Translation. arXiv preprint arXiv:2005.10608 (2020).
7. Kishore, P., Salim, R., Todd, W., Wei-Jing, Z.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computa-tional Linguistics, pp. 311-318 (2002).
8. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019).
9. Sellam, T., Das, D., Parikh, A. P.: BLEURT: Learning Robust Metrics for Text Generation. arXiv preprint arXiv:2004.04696 (2020).
10. Lucia, S.: Exploiting objective annotations for measuring translation post-editing effort. In: Proceedings of the 15th Conference of the European Association for Machine Translation, pp. 73–80 (2011).
11. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., ... & Logacheva, V.: Findings of the 2017 conference on machine translation. In: Proceedings of the Second Conference on Machine Translation, pp.169–214 (2017).
12. Kim, H., Jung, H.-Y., Kwon, H., Lee, J. H., Na, S.-H.: Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 17(1):3 (2017).
13. Kai, F., Bo, L., Fengming, Z., Jiayi W.: "Bilingual Expert" Can Find Translation Errors. arXiv preprint arXiv:1807.09433 (2018).
14. Kepler, F., Trénous, J., Treviso, M., Vera, M., Góis, A., Farajian, M. A., ... & Martins, A. F.: Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. arXiv preprint arXiv:1907.10352 (2019).
15. Frédéric, B., Nikolaos, A., Lucia, S.: Quality in, quality out: Learning from actual mistakes. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (2020).
16. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
17. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019).

18. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019).
19. Pires, T., Schlinger, E., Garrette, D. How multilingual is Multilingual BERT?. arXiv preprint arXiv:1906.01502 (2019).
20. Barrault, L., Bojar, O., Costa-Jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... & Monz, C.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 1-61 (2019).
21. Ma, Q., Wei, J., Bojar, O., Graham, Y.: Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 62-90 (2019).
22. Niven, T., Kao, H. Y.: Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355 (2019).
23. Tandon, N., Varde, A. S., de Melo, G.: Commonsense knowledge in machine intelligence. ACM SIGMOD Record, 46(4), 49-52 (2018).
24. Zhang, J., Liu, Y., Luan, H., Xu, J., Sun, M.: Prior knowledge integration for neural machine translation using posterior regularization. arXiv preprint arXiv:1811.01100 (2018).