Dealing with the Paradox of Quality Estimation

Sugyeong Eo* djtnrud@korea.ac.kr
Chanjun Park* bcj1210@korea.ac.kr
Hyeonseok Moon glee889@korea.ac.kr
Jaehyung Seo seojae777@korea.ac.kr
Heuiseok Lim† limhseok@korea.ac.kr
Department of Computer Science and Engineering, Korea University, Korea

Abstract

In quality estimation (QE), the quality of translation can be predicted by referencing the source sentence and the machine translation (MT) output without access to the reference sentence. However, there exists a paradox in that constructing a dataset for creating a QE model requires non-trivial human labor and time, and it may even requires additional effort compared to the cost of constructing a parallel corpus. In this study, to address this paradox and utilize the various applications of QE, even in low-resource languages (LRLs), we propose a method for automatically constructing a pseudo-QE dataset without using human labor. We perform a comparative analysis on the pseudo-QE dataset using multilingual pre-trained language models. As we generate the pseudo dataset, we conduct experiments using various external machine translators as test sets to verify the accuracy of the results objectively. Also, the experimental results show that multilingual BART demonstrates the best performance, and we confirm the applicability of QE in LRLs using pseudo-QE dataset construction methods.

1 Introduction

In the field of machine translation (MT), most of the representative metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are used to measure the quality of MT output by comparing it with the reference sentence. However, these evaluation metrics limit the amount of datasets owing to the need for a reference sentence (Specia et al., 2010). In cases where end users use MT, they do not have sufficient knowledge of the source or target languages. Specifically, In the case of low-resource languages (LRLs), people are often unfamiliar with such languages. In such cases, it is difficult to determine whether the translation results derived using MT have been translated well.

Recently, studies on quality estimation (QE) have been actively conducted to address this problem (Kim et al., 2019; Wang et al., 2020; Fomicheva et al., 2020). In QE, the source sentence and the MT output are referenced to predict the quality of translation result. QE can be used to express the quality of MT output numerically, rank the results of several MT systems (Specia et al., 2010), and inform end users on MT system's level of trust. Quality annotations resulting from the QE system also allows individuals who are unfamiliar with the translation languages to verify the quality of MT outputs (Specia et al., 2013). Additionally, post-editing efforts can be reduced by filtering out poor-quality MT outputs (Specia et al., 2009; Specia, 2011). As a result, the importance of QE research has been emphasized in the field of MT.

*These authors contributed equally.
†Corresponding author.

We found one paradox pertaining to this useful QE task. QE has an advantage in that it can make predictions about MT results without using a reference sentence. However, efforts to build datasets that require more expertise than building a parallel corpus must be made to ensure the progress of QE. These requirements also limit the construction of large QE datasets. We refer to this paradox as the *paradox of QE*, and we use methods for generating a pseudo-QE dataset to address this paradox.

Because it is difficult to obtain a parallel corpus for LRLs and hinders to build a QE dataset for such languages, there are few QE studies on LRLs, except for those provided by the Conference on Machine Translation (WMT). Based on these limitations, we conduct a study on sentence-level QE with a main focus on LRLs. We construct a pseudo-QE dataset by automatically expanding Korean-based monolingual or parallel corpora without using extra human labor.

We conduct a comparative analysis between QE models based on various multilingual pretrained language models (mPLMs), and we confirm the possibility of creating a QE model for LRLs through the experimental results. The contributions of this study are as follows:

- We point out the *paradox of QE* and to address this problem, we propose a method for automatically constructing a pseudo dataset using monolingual or parallel corpora and external machine translators without additional human labor.
- We conduct a QE study on LRLs, where previous studies on the same are rare, and we induce the various applications of QE in LRLs.
- We conduct a quantitative analysis based on various mPLMs, and conduct an empirical study using the results obtained through external machine translators, such as Google[1], Amazon[2], Microsoft[3], and Systran[4], to verify the objectivity of the translation results as we construct the pseudo dataset.

2 Proposed Method

2.1 Why Paradox?

In this section, we describe why the paradox of QE occurs at various granularity (sentence/word/document) levels of QE based on WMT20. We also describe methods for generating a pseudo-QE dataset that can address the limitations for the paradox of QE.

Paradox of QE - Sentence Level In the sentence-level direct assessment task, the MT output is evaluated based on perceived quality, which is referred to as direct assessment (DA) (Specia et al., 2020). At least three translation experts rate the quality of the MT output from zero to 100, and the system predicts the mean z-standardized DA. The dataset construction for this task requires DA annotations from at least three human experts.

The sentence-level post-editing task is configured to predict the quality score for the MT output based on the human translation error rate (HTER) (Snover et al., 2006). HTER scores are obtained through the comparison between the MT outputs and human post-edited sentences. Thus, to generate post-edited sentences for measuring HTER scores, humans must consider how minimal changes make the MT output a correct sentence, which tokens in the MT output have been mistranslated, and how to change them. Building a parallel corpus for LRLs is not easy, and hiring language experts is more difficult. These limitations make LRL-based QE studies

[1] https://translate.google.co.kr/
[2] https://aws.amazon.com/translate/
[3] https://www.microsoft.com/ko-kr/translator/
[4] https://translate.systran.net/

more challenging. The tagging process also requires post-edited sentences to be corrected by translation experts, who are quite limited in terms of human labor.

Paradox of QE - Word Level In the word-level post-editing task, the quality of the MT output is predicted using the OK label or the BAD label for each token, and the GAP tag is used in cases where there is a missing word between tokens. The tagging process also requires post-edited sentences to be modified by a translation expert. However, similar to sentence-level, the construction of a large dataset is quite limited in terms of human labor. The number of datasets released annually by the WMT is only 9K, including those on the train, validation, and test for one pair of languages.

Paradox of QE - Document Level The document-level task is configured to find translation errors in documents and estimate quality scores based on minor, major, and critical errors. In the dataset used in this task, the error part is annotated using span and span length (Specia et al., 2020). Error annotations, such as severity, word span, and specific error type are annotated through crowd-sourcing. Human labor is essential for this process because constructing a new dataset requires humans to annotate the errors. In LRL settings, the language itself is sometimes unfamiliar, making it more difficult to hire an expert that can tag translation errors in documents.

2.2 Constructing Pseudo-QE Dataset

We point out that in QE, building a dataset requires additional effort compared to the translation process. To address this issue, we propose two strategies for generating a pseudo-QE dataset for Korean, which is an LRL, and we conduct sentence-level post-editing, a sub-task of WMT.

2.2.1 Monolingual Corpus-based (M-based) Pseudo-QE Dataset Generation

The monolingual corpus-based (M-based) pseudo-QE dataset is a method for constructing a QE dataset based on round-trip translation (RTT). We generate a dataset with a three-step process based on the fact that RTT can be used to generate paraphrased sentences (Mallinson et al., 2017).

The first process involves a backward translation of the source language. In this process, we adopt Google as an external machine translator because it can easily translate large documents and is frequently used by most people. The source text generated through the backward translation process is similar to the source-side text of the parallel corpus, but there are some errors or paraphrased parts. The output of the first process is converted back into the text of the target language via the second process, which is known as forward translation. In the process of combining and traversing monolingual text using external machine translators in the target language, errors easily committed by translators are additionally attached to the plane text, and the skewed output with translation errors is generated.

In the final process, the translation error rate (TER) between the monolingual corpus and the skewed output is extracted. In other words, we consider the monolingual text as a post-edited sentence, and we measure the HTER using the generated pseudo dataset to eliminate human labor.

In this case, the pseudo dataset created through this approach may only be distorted depending on the error tendency of Google translator. Considering this situation, we use the translation results from additional representative external translators, such as Amazon, Microsoft, and Systran, as test sets to ensure that QE models trained using pseudo datasets predict the quality of translation in a general way.

2.2.2 Parallel Corpus-based (P-based) Pseudo-QE Dataset Generation

Utilizing parallel corpora and external machine translators is a method for constructing parallel corpus-based (P-based) pseudo-QE datasets.

Similar to the first step, the source-side text is entered into the external machine translator, after which it is translated to the target language. In the process of translating the source-side text to the target language, the source-side text is translated to the MT output with errors attached. In the second step, the TER is measured for each sentence using the target-side text from the parallel corpus, considering the LRL settings similar to the M-based dataset generation method.

We organize the dataset to enable the measurement of translation quality without additional human labor by solely using the parallel corpus. However, even in a P-based dataset, error types may appear to be biased to only one external machine translator throughout the dataset construction process. Therefore, the objectivity of how well the translation quality was measured, as in the monolingual case, was verified using test sets containing multiple translation results from external machine translators. The overall process of our proposed method is shown in Figure 1.

2.3 TransQuest-based Korean QE Model

We conduct training on the pseudo-QE dataset using TransQuest[5] (Ranasinghe et al., 2020), which is an open-source framework. Ranasinghe et al. (2020) proposes two structures: MonoTransQuest and SiameseTransQuest. We focus on the consistent high performance of MonoTransQuest, and we only utilize the former structure for learning. Three pooling strategies were experimented in MonoTransQuest, of which the output corresponding to the location of the [CLS] token was inserted into the softmax layer and the score was predicted. In addition to XLM-RoBERTa (XLM-R) used by MonoTransQuest, we use the multilingual BART (mBART) and the cross-lingual language model (XLM), which support Korean, for model performance comparison. For mBART model that is not associated with any previous studies on QE, we find it worth fully exploiting this because they are state-of-the-art models in MT, and we utilize additional noising schemes compared to those used in XLM and XLM-R models.

3 Experiments and Results

3.1 Dataset Details

In this study, we conduct experiments on the sentence-level task corresponding to sub-task 2 of the WMT20 based on various mPLMs for Korean, which is one of the LRLs. As the dataset for our experiments, we leverage two methods to build our proposed pseudo-QE training dataset. We use data from AI-HUB[6] (Park and Lim, 2020) and only the sentences of the target-side for the M-based pseudo-QE dataset.

The statistics of the dataset obtained through the two dataset generation methods are listed in Table 1. In Korean, the sum of the total token lengths of the M-based dataset is more than that of the P-based dataset, but the opposite occurs in English. In other words, when translated from the target language to Korean, the average length of the translated sentence becomes longer than that of the original source. However, when it is translated based on RTT into Korean, the number of tokens in the translated sentence tends to be smaller, even if the length of the source sentence is longer. Overall, the TER scores were distributed slightly lower on the M-based datasets.

Based on the datasets constructed using both methods, we segment the TER scores at 0.1 intervals, and count the scores that are part of each range, as shown on the left side of Figure 2. The distribution over the dataset shows that the M-based dataset is lower overall than the P-based dataset, as illustrated in Table 1. Based on these results, we explore the length distribution of the MT token over the range of the TER scores to analyze why the TER scores are low in the M-based dataset. As shown on the right side of Figure 2, both datasets are distributed with lower error rates as the token length becomes shorter in the TER score range from zero to five.

[5] https://github.com/TharinduDR/TransQuest

[6] https://aihub.or.kr/

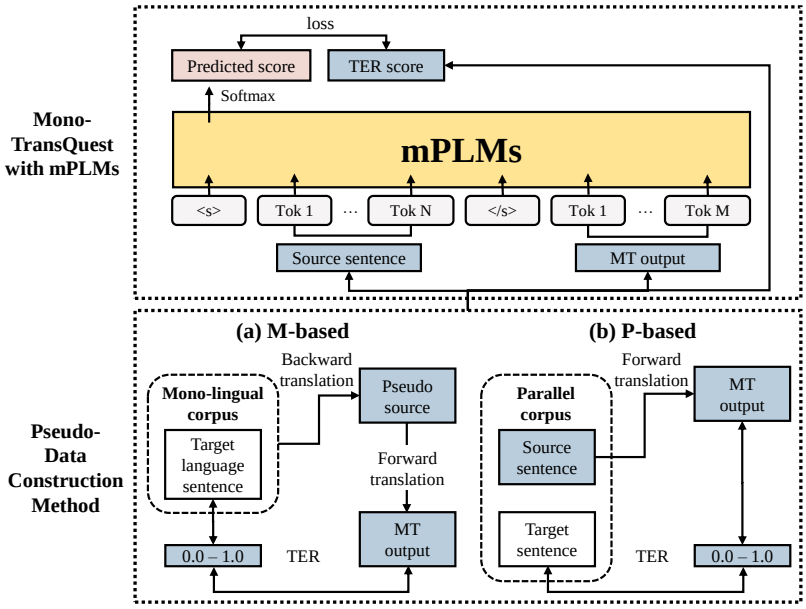

Figure 1: Overall architecture of pseudo QE dataset construction method and model training. (a) corresponds to a monolingual corpus based pseudo-QE dataset generation method, and (b) corresponds to a parallel corpus based method.

However, in the case where the TER score is higher than 0.5, the average token length of the P-based dataset is six to seven times higher overall compared to the M-based dataset. The graph shows that the error rate is also high when MT sentences are generally long and that longer sentences in the P-based dataset result in a negative effect on the TER scores.

	M-based Pseudo Dataset				**P-based Pseudo Dataset**			
	Train		**Valid**		**Train**		**Valid**	
	SRC	**MT**	**SRC**	**MT**	**SRC**	**MT**	**SRC**	**MT**
# of sentences	96,000	96,000	12,000	12,000	96,000	96,000	12,000	12,000
# of tokens	1,457,832	2,215,902	183,258	278,451	1,345,381	2,370,791	168,507	297,126
# of min tokens per S	1	1	1	1	3	2	3	2
# of max tokens per S	84	123	60	87	71	143	45	122
Average tokens per S	15	23	15	23	14	24.6	14	24.7
Average TER score	0.419		0.415		0.527		0.525	
Median TER score	0.417		0.417		0.524		0.523	

Table 1: Statistics of the pseudo-QE train and valid dataset. We denote the sentence as S.

	Google	**Amazon**	**Microsoft**	**Systran**
# of sentences	12,000	12,000	12,000	12,000
# of tokens	297,011	264,401	283,450	302,239
# of min tokens in a S	3	3	1	3
# of max tokens in a S	158	120	142	162
Average tokens per S	24.7	22	23.6	25
Average TER score	0.526	0.591	0.591	0.418
Median TER score	0.524	0.6	0.6	0.4

Table 2: Statistics of the pseudo-QE test sets constructed using various external machine translators

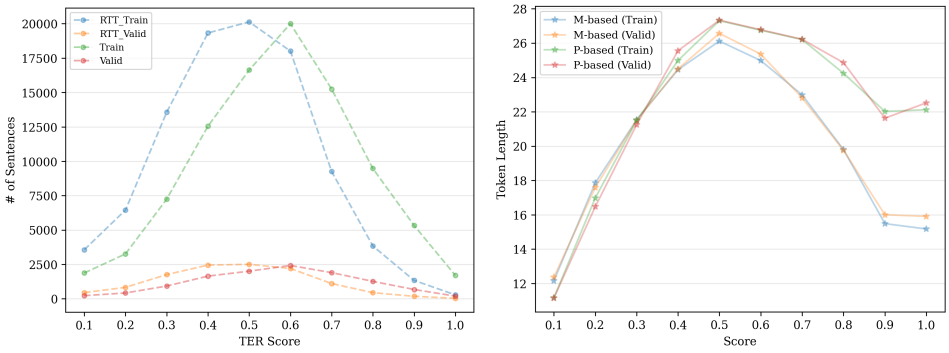

Figure 2: Number of sentences (left-side) and sentence length (right-side) according to TER score range

We build a pseudo dataset without any human labor. In addition, we leverage external machine translators by Google, Amazon, Microsoft, and Systran, to ensure objective evaluation considering the possibility of learning distortion based on the MT results. The statistics for each external machine translator are listed in Table 2. Compared to the train and validation set of the M-based dataset, the TER scores are generally higher, except for Systran. The average token length per sentence is distributed similarly, with 22 to 25 overall.

3.2 Model details

In this study, we conduct a comparative analysis by fine-tuning three representative mPLMs: XLM-R-large, XLM-MLM-100, and mBART. We compare the performance of these models to discover the performance differences that occur depending on the number of language pairs and the noising schemes in the pre-training stage. The description for each model is as follows:

- **Cross-lingual language model (XLM)**: XLM (Lample and Conneau, 2019) is a structure that extends the existing learning method of a language model for the purpose of learning multi-lingual representations. The XLM proposes a causal language model (CLM) that performs unsupervised learning on monolingual corpora, the translation language model (TLM) that implements supervised learning on parallel corpora, and the masked language model (MLM). We used XLM-MLM-100, which is a model pre-trained using a total of 100 languages, including Korean, among various XLM models.

- **XLM-RoBERTa (XLM-R)**: XLM-R (Conneau et al., 2019) significantly increases the number of datasets and conducts pre-training by applying only MLM among the learning methods of XLM. XLM-R faces the curse of multilinguality because it increases the number of training datasets and extends the number of languages. The curse of multilinguality refers to a situation in which the addition of languages improves the performance of LRLs, which have similar linguistic features with high-resource languages, initially by high-resource languages. However, at some point, the performance of both the high-resource languages and the low-resource languages is reduced when the model capacity is fixed. This is because the number of languages increases and the capacity of high-resource languages within the model decreases. By greatly expanding the model capacity, it is possible to improve the performance of low-resource languages and maintain the performance of high-resource languages.

- **multilingual BART (mBART)**: mBART (Liu et al., 2020) is a multilingual extension of BART (Lewis et al., 2019). BART adds noise from sentence permutations, token masking,

token deletion, and text infilling, and document rotations to restore them to a completely original sentence based on the structure of transformer. mBART does not utilize all the noise schemes used in BART. However, it learns by employing text infilling that replaces the span length with one [MASK] token according to the Poisson distribution in sentences and the sentence permutation that shuffles the order of sentences. mBART supports the efficient learning of LRLs by matching the dataset rates between low-resource and high-resource languages. In other words, mBART applies up-down sampling method that increases the number of datasets by copying the same in LRLs and by removing parts of the datasets in high-resource languages.

We fine-tune the pre-trained models by leveraging the framework of the Huggingface model (Wolf et al., 2019). Based on the framework provided by this model, we implement sub-word tokenization, and include the position and language embeddings for XLM. As a loss function for model learning, we use the mean square error (MSE) loss.

3.3 Main Results

As shown in Table 3 and Table 4, according to the tests conducted using datasets built using various external translators, the performance differences based on the Pearson correlation coefficient between the external translators differ by 0.193 on the M-based datasets and 0.052 on the P-based datasets. Specifically, there is no significant difference in performance (0.048), except for the results of the experiments conducted using the Systran translator on the M-based datasets. Therefore, we can conclude that that the performance difference between the external translators is not significant.

	Google			Amazon			Microsoft			Systran		
Model	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**
XLM-R	0.236	0.175	0.223	0.307	0.194	0.237	0.278	0.198	0.245	0.076	0.189	0.232
mBART	**0.334**	0.174	0.221	**0.382**	0.199	0.240	**0.360**	0.202	0.247	**0.189**	0.183	0.226
XLM-MLM-100	0.156	0.185	0.234	0.212	0.215	0.257	0.150	0.218	0.265	-0.042	0.188	0.232

Table 3: Results of the M-based pseudo-QE dataset

	Google			Amazon			Microsoft			Systran		
Model	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**	**Pearson**	**MAE**	**RMSE**
XLM-R	0.346	0.157	0.197	0.366	0.158	0.197	0.358	0.164	0.204	0.261	0.194	0.234
mBART	**0.450**	0.146	0.186	**0.445**	0.146	0.184	**0.450**	0.151	0.189	**0.398**	0.185	0.226
XLM-MLM-100	0.285	0.160	0.201	0.269	0.168	0.207	0.259	0.172	0.213	0.272	0.191	0.231

Table 4: Results of the P-based pseudo-QE dataset

3.3.1 Experimental results of M-based Pseudo-QE Model

We conduct a comparative analysis on the models trained using the M-based dataset. The experimental results are similar to those listed in Table 3, and they show the differences in performance in the order of the mBART, XLM-R, and XLM-MLM100 models.

Interpreting Results on Language Capacity The results show that the number of language pairs used in pre-training is not proportional to performance. Although mBART is trained using 25 language pairs, it performs better than XLM-MLM 100 and XLM-R, which are used to conduct pre-training in 100 language pairs. This shows that abundant language pairs do not necessarily benefit the QE of LRLs.

Interpreting Results for the Noising Scheme In XLM-R and XLM-MLM100, only MLM is utilized in the pre-training stage. mBART adds sentence permutation and text infilling during the pre-training process, thereby demonstrating the highest performance. Therefore, we can infer that the additional noising schemes for mBART are the critical factors that result in better results. Liu et al. (2020) also demonstrate that additional strategies for noising schemes are beneficial, and that model capability depends heavily on pre-training methods rather than the number of language pairs.

Interpreting Results for the Tokenization Method Korean is classified as an agglutinative language based on morphological characteristics. Depending on the characteristics of agglutinative languages, a single word may consist of just one word. However, there are some cases in which a substantive (noun, pronoun, numeral) and a post-positional particle appear together or a stem and an ending co-occur. Recent studies have shown that the tokenization method is an important approach that considers morphemes because they have a variety of meanings determined by the post-positional particle (Park et al., 2020, 2021).

mBART and XLM-R employ SentencePiece (Kudo and Richardson, 2018), and XLM-MLM uses byte pair encoding (BPE) (Sennrich et al., 2015). Among them, mBART applies morphological segmentation by considering the agglutinative characteristics of Korean, which can be interpreted as one of the reasons for enhancing the understanding of source text. In the case of the BPE used by XLM, the criteria for pre-tokenization are ambiguous in Korean, and they construct vocabularies in a greedy way. Therefore, there is a high probability of proceeding with incorrect sub-word segmentation. By using the XLM tokenizer, '⟨/w⟩' tokens are attached to the end of every syllable as well as the complete separation of syllables into consonants and vowels. Accordingly, it can be interpreted that the words are completely separated through the use of syllable units, thereby resulting in the poor understanding of the entire sentence and demonstrating the low performance of XLM.

3.3.2 Experimental results of P-based Pseudo-QE Model

Furthermore, we conduct a comparative analysis on models trained using the P-based dataset. As shown in Table 4, mBART and XLM-MLM-100 demonstrate the highest performance and the lowest performance, respectively, for all test sets. This difference in performance can be considered similar to that obtained in the previous analysis. Considering the construction of the dataset, we establish that the overall capability improves when the model is trained using a P-based dataset rather than an M-based dataset. Moreover, it is certain to obtain more desirable results, as they pertain to the measurement of the TER, by comparing the translation of the source sentences in parallel corpora with target sentences, rather than building datasets based on RTT. This result is attributed to the higher intimacy of the test set as a result of translating source sentences into multiple external machine translators and P-based datasets. In contrast, despite the same number of training sentences used in P-based datasets and M-based datasets, the Pearson correlation coefficients differed by a range of 0.063 to 0.209. Because the M-based dataset allows for much more datasets to be added compared to parallel corpora, learning using M-based datasets can also be expected to achieve sufficient performance gains.

4 Conclusion and Future Work

This study points out a paradox in terms of the construction of data for QE tasks. To address this limitation, we propose two methods for generating a pseudo dataset. First, considering the limitations of data construction in low-resource language settings, we generate an RTT-based pseudo-QE dataset using monolingual corpora, and second, we construct pseudo data using parallel data. The experiments are conducted using mPLMs that support Korean, and mBART demonstrated the highest performance. By conducting tests using various external

machine translators, we further confirm that the model trained using a pseudo dataset is not significantly skewed on a specific external translator. Therefore, by leveraging pseudo-QE generation methods, we confirm that QE is also available in LRLs, and induce the use of various applicability of QE in LRLs. In our future studies, as we have seen the possibility of sufficient performance improvement for the result of experimenting with monolingual corpora, we plan to conduct further experiments to expand the amount of data to large-scale. We also plan to expand the proposed methodology to various language pairs and conduct detailed verification of the proposed methodology.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation) and the MSIT, Korea, under the ICT Creative Consilience program(IITP-2021-2020-0-01819) supervised by the IITP. Additionally, this work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

References

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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*.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Kim, H., Lim, J.-H., Kim, H.-K., and Na, S.-H. (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.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Park, C., Eo, S., Moon, H., and Lim, H.-S. (2021). Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 97–104.

Park, C. and Lim, H. (2020). A study on the performance improvement of machine translation using public korean-english parallel corpus. *Journal of Digital Convergence*, 18(6):271–277.

Park, C., Yang, Y., Park, K., and Lim, H. (2020). Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Citeseer.

Specia, L. (2011). Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzmán, F., and Martins, A. F. T. (2020). Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (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, Sofia, Bulgaria. Association for Computational Linguistics.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Wang, M., Yang, H., Shang, H., Wei, D., Guo, J., Lei, L., Qin, Y., Tao, S., Sun, S., Chen, Y., et al. (2020). Hw-tsc's participation at wmt 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.