Quality Estimation without Human-labeled Data

Yi-Lin Tuan[1], Ahmed El-Kishky[2], Adithya Renduchintala[2], Vishrav Chaudhary[2], Francisco Guzmán[2] and **Lucia Specia[3]**
[1]University of California Santa Barbara, [2]Facebook AI, [3]Imperial College London
[1]ytuan@cs.ucsb.edu, [2]{ahelk,adirendu,vishrav,fguzman}@fb.com
[3]l.specia@imperial.ac.uk

Abstract

Quality estimation aims to measure the quality of translated content without access to a reference translation. This is crucial for machine translation systems in real-world scenarios where high-quality translation is needed. While many approaches exist for quality estimation, they are based on supervised machine learning requiring costly human labelled data. As an alternative, we propose a technique that does not rely on examples from human-annotators and instead uses synthetic training data. We train off-the-shelf architectures for supervised quality estimation on our synthetic data and show that the resulting models achieve comparable performance to models trained on human-annotated data, both for sentence and word-level prediction.

1 Introduction

The adoption of Machine Translation (MT) has been increasing in areas ranging from government and finance, to even social media due to the substantial improvements achieved from Neural Machine Translation (NMT). However, even with improved performance, translation quality is not consistent across language pairs, domains, and sentences. This can be detrimental to end-user's trust and can cause unintended consequences arising from poor translations. Thus, having metrics to assess the quality of translated content is crucial to ensure that only high-quality translations are provided to end-users or downstream tasks. Quality Estimation (QE) metrics aim to predict translation quality without access to reference translations (Blatz et al., 2004; Specia et al., 2009, 2013).

State-of-the-art QE techniques have leveraged MT systems and language-specific human annotations as supervision, including direct assessment and post-editing (Kepler et al., 2019a; Fonseca et al., 2019; Sun et al., 2020). However, these annotations are costly and time-consuming, particularly for word-level QE, where each token needs a label.

Some unsupervised approaches take inspiration from statistical MT (Popović, 2012; Moreau and Vogel, 2012; Etchegoyhen et al., 2018) or apply uncertainty quantification (Fomicheva et al., 2020) for QE. However, their performance is inferior to that of supervised models. In related areas such as automatic post-editing, parallel data has been used to create synthetic post-editing data (Negri et al., 2018), however this technique only compares machine-translated sentences to references. Our approach augments MT errors with additional errors via masked language model rewriting.

We leverage noisy, mined comparable sentences obtained by weakly-supervised techniques (El-Kishky et al., 2020b). These noisy bitexts have been mined from a variety of domains such as Wikipedia (Schwenk et al., 2019a) and large web-crawls (Schwenk et al., 2019b; El-Kishky et al., 2020a; El-Kishky and Guzmán, 2020) and have been shown to be an invaluable source of training data for NMT models. Using this data is crucial to avoid data leakage between a trained NMT model and the data we use to create synthetic QE data. For each source-target sentence pair from the mined data, we apply an MT system to generate a candidate translation of the source sentence. Additionally we rewrite each target reference sentence using a masked language model to introduce errors. These two approaches generate two alternative "translations" of the source sentence. We then produce pseudo-labels for each token in these translations by edit distance alignment to the original reference sentence. This results in each translated word being pseudo-labelled as correct or incorrect, which is our synthetic QE training data. Analogously, sentence-level training data is derived as the proportion of incorrect words per sentence.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 619–625
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

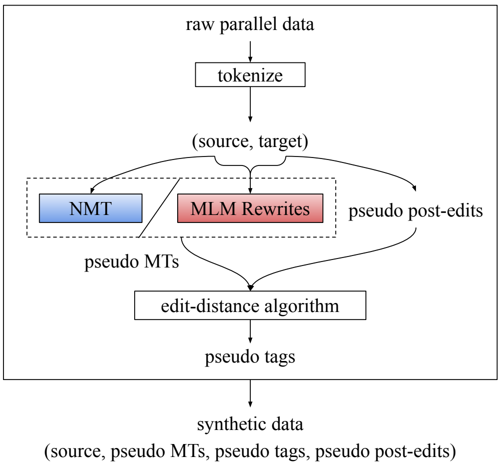

Figure 1: The pipeline to synthesize data for QE from comparable mined data.

Our **main contributions** are: (i) We explore a simple technique to effectively generate synthetic data for QE that allows for both word-level and sentence-level estimation (ii) we demonstrate that our technique performs comparably to off-the-shelf models trained on human-annotated data.

2 QE Task Description

Word-level QE has been mainly framed as the task of predicting which words in the translation need to be post-edited. As such, word-level QE aims to assign a tag for each word and gap between words in a machine-generated translation as *correct*, i.e., the word does not need editing, or *incorrect*, i.e., the words should be substituted, deleted, or inserted (tags for gaps) (Specia et al., 2020).

For word-level, we denote the tag of each word in a translation as $m_t \in \{\texttt{OK}, \texttt{BAD}\}$, where $t \in [1, T]$ and T is the length of the translation. Also, we denote the tag of each gap between two words (including the beginning and the end) as $g_t \in \{\texttt{OK}, \texttt{BAD}\}$, where $t \in [1, T+1]$.

In traditional QE, data is collected by first translating source sentences using an MT model. Second, experts post-edit these translations. Third, the post-edits and machine translations are aligned in such a way that induces the minimum edit distance between the tokens of each. Finally, each m_t is labelled as BAD if it should be deleted or substituted and each g_t is labelled as BAD if at least a word should be inserted there. Sentence-level QE labels can be generated by computing the Human-targeted Translation Error Rate (HTER) (Snover and Brent, 2001; Snover et al., 2006), which is the minimum ratio of edit operations needed to fix the translation to the number of its tokens. We explore the possibility to skip the costly human post-editing process by proposing a data synthesis pipeline, which we then test on human labelled data.

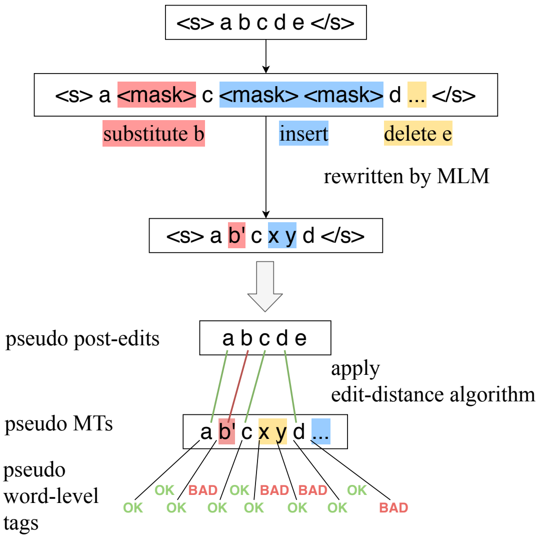

Figure 2: The rewriting process by text-infilling using a masked language model.

3 Approach to Data Synthesis

As depicted in Figure 1, we synthesize data from mined Wikipedia datasets, where each example consists of a *(source, target)* sentence pair.

We create candidate translations of source sentences in two ways: For the first approach, we apply the *NMT* model to translate each source sentence. For the second approach, we rewrite each reference target sentence using a masked language model (MLM), as shown in the *MLM Rewrites* block in Figure 1. The two approaches create two forms of translations. Then, by treating target sentences as if they were post-edited data (*pseudo post-edits*), we identify errors in each candidate translation by looking at the insertions, deletions, and substitutions between the references and generated translations.

Neural Machine Translation. For the first approach to generating synthetic data, we use a pretrained NMT model to create translations. The NMT model is the same model that was used to generate translations in the supervised data; the architecture is a standard transformer as used in (Vaswani et al., 2017; Ott et al., 2019). The process of creating synthetic QE data first involves translating each source sentence using this model and taking the output as a translation which will

Data	English-German			English-Chinese		
	size	MT bad (%)	Gap bad (%)	size	MT bad (%)	Gap bad (%)
Human annotation	7K	27.8	4.7	7K	54.2	8.4
NMT	459K	38.2	5.7	189K	49.5	6.8
MLM (word-QE)	459K	40.7	2.9	189K	53.9	8.6
MLM (sent-QE)	459K	43.1	3.3	189K	49.9	2.7

Table 1: Statistics of annotated and synthetic (NMT and MLM) data.

later be used to generate the synthetic labels. When decoding, we apply a beam of 5 following the NMT models available in Fomicheva et al. (2020) to generate a candidate translation. Next, we take the mined reference target sentence and treat it as a pseudo post-editing of the machine translation.

We then compute the edit distance between MTs and pseudo post-edits. The resulted edit operations are the pseudo tags, which consist of word tags m_t and gap tags g_t. This process is illustrated in Algorithm 1.

Algorithm 1: DataSynthesis-NMT

```
Input: pairs (source, target) from mined data,
       pretrained NMT model
Output: (MTs, pseudo tags)
for each pair (source, target) do
    MTs = NMT(source)
    $\{m_t\}_{t=1}^{T}$, $\{g_t\}_{t=1}^{T+1}$ = edit_distance(MTs, target)
    pseudo tags = ($\{m_t\}_{t=1}^{T}$, $\{g_t\}_{t=1}^{T+1}$)
    return (MTs, pseudo tags)
```

Rewriting by Masked Language Model (MLM). Our second approach to creating synthetic QE training data is to introduce errors by rewriting target sentences. We inject these errors by performing *text-infilling* (Zhu et al., 2019; Lewis et al., 2019). As displayed in Figure 2, we perform text-infilling by applying three operations: (1) randomly substituting a proportion of tokens with a `<mask>` token, (2) deleting consecutive tokens, and (3) inserting additional consecutive `<mask>` tokens. We determine the lengths of consecutive deletions and insertions by drawing them from a Poisson distribution with mean $\lambda = 1$ shifted by 1 to avoid zero-length insertions or deletions. We then use a pre-trained masked language model (MLM) supplied with the source sentence as input to infill the masked reference sentence. We select multilingual BERT (Devlin et al., 2019) as it is pre-trained on Wikipedia which is in-domain to our test set. We present the target-rewriting approach in detail in Algorithm 2.

In Section 4, we will investigate the performance of QE models trained on NMT-based synthetic data, rewriter-based synthetic data, and a two-model ensemble where each model is trained on a different form of synthetic data.

Algorithm 2: DataSynthesis-Rewriting

```
Input: pairs $(S, W)$:=(source, target) from mined
       data, pretrained MLM
Input: $P_s$, $P_d$, $P_i$ as the probabilities of substitution,
       deletion, and insertion
Output: (pseudo MTs, pseudo tags)
for each pair $(S, W)$ do
    $W'$ = randomly mask tokens in $W$ by $P_s$
    $D$ = randomly mark deletion in $W'$ by $P_d$
    $W'$ = randomly delete a text span from marks $D$
      in $W'$ (length$\sim$ $Poisson(\lambda = 1) + 1$)
    $I$ = randomly mark insertion in $W'$ by $P_i$
    $W'$ = randomly insert contiguous masks from
      marks $I$ in $W'$ (length$\sim$ $Poisson(\lambda = 1) + 1$)
    rewrites = MLM_fills_in_masks($S$, $W'$)
    $\{m_t\}_{t=1}^{T}$, $\{g_t\}_{t=1}^{T}$ = edit_distance(rewrites, $W$)
    pseudo tags = ($\{m_t\}_{t=1}^{T}$, $\{g_t\}_{t=1}^{T+1}$)
    return (rewrites, pseudo tags)
```

4 Experiments and Results

We focus on data released by the WMT20 shared task on QE for predicting post-editing effort, which includes English-to-German (En-De) and English-to-Chinese (En-Zh) word-level data and their sentence-level HTER (Specia et al., 2020).[1] As the human-annotated data is sampled from Wikipedia, we choose to synthesize data from WikiMatrix (Schwenk et al., 2019a), which consists of mined Wikipedia parallel data from which we sample pairs with a LASER (Artetxe and Schwenk, 2019) margin score threshold of 1.06 to ensure high-quality pairs. We note that the original QE data is not a subset of WikiMatrix. The German and Chinese text were tokenized using the Moses[2] and Jieba[3] tokenizers, respectively. We list the statistics of the filtered Wikimatrix data as well as our resulting synthetic data in Table 1.

For the off-the-shelf QE model, we choose the

[1] Available here: https://github.com/sheffieldnlp/mlqe-pe

[2] https://github.com/alvations/sacremoses

[3] https://github.com/fxsjy/jieba

Data	English-German			English-Chinese		
	MCC	F1-Ok	F1-Bad	MCC	F1-Ok	F1-Bad
Human annotation	**0.399**	0.879	**0.495**	0.525	0.820	0.659
MLM	0.332	**0.892**	0.438	0.500	0.850	0.643
NMT	0.379	0.826	0.468	0.525	**0.859**	0.660
NMT + MLM	**0.399**	0.866	0.493	**0.546**	0.835	**0.675**
Improvement (%)	+0.20	-1.40	-0.40	+4.00	+1.83	+2.43

Table 2: Results of word-level QE trained on human-annotated (7k) and synthetic data. Improvement in MCC for en-de & en-zh shows synthetic data can train word-level models comparable to human-annotated data. We report improvement comparing models trained with human-annotation vs our combined NMT+MLM synthetic data.

Data	English-German			English-Chinese		
	Pearson	MAE	RMSE	Pearson	MAE	RMSE
Human annotation	**0.394**	**0.150**	**0.187**	0.490	0.151	0.186
MLM	0.290	0.156	0.195	0.418	0.224	0.269
NMT	0.327	0.229	0.270	0.482	0.161	0.203
NMT + MLM	0.373	0.172	0.205	**0.506**	**0.148**	**0.183**
Improvement (%)	-5.50	+14.7	+9.63	+3.18	-1.79	-1.67

Table 3: Results of sentence-level HTER QE trained on human-annotated and synthetic data. For Pearson, positive improvement is better while for MAE & RMSE negative is better. We report improvement comparing models trained with human-annotation vs our combined NMT+MLM synthetic data.

multi-task predictor-estimator model (Kim et al., 2017) implemented by OpenKiwi v0.1.3 (Kepler et al., 2019b). This was the top-performing architecture for QE at WMT19 (Kepler et al., 2019a; Fonseca et al., 2019). We train the predictor on parallel MT data provided by the WMT20 QE shared task. The predictor reads in words' contextualized word representations, the estimator passes these features through a 2-layer 125-dimension bidirectional LSTM (biLSTM) and then feeds the outputs into 1-layer linear word-level classifier. The first output of the biLSTM is also fed into a multi-layer perceptron to predict a sentence-level score. For multi-task learning, we train the model with both word- and sentence-level data.

For a fair comparison, we take the pre-trained predictor provided by the WMT20 QE shared task, fine-tune the whole model on the human annotated data, and compare results to those when fine-tuned on our synthetic data. We test by comparing model predictions and held-out human-annotated QE at word and sentence-level. At the word level, we measure QE performance with Matthew's Correlation Coefficient (MCC) (Matthews, 1975) (main metric), as well as F1 scores for BAD and OK tags. At the sentence-level, we measure the sentence-level Pearson's correlation (Benesty et al., 2009), mean absolute error (MAE) and Root-mean-square deviation (RMSE).

As shown in Table 2, for word-level QE,[4] the model trained on synthetic data generated from NMT translations performs comparably to the same model trained on the original 7k human-annotated post-edits. This suggests that having human annotators post-edit each translation to create training data may be unnecessary and using reference sentences is good enough. The model trained on the MLM rewriting synthetic data generally under-performs compared to NMT generated data on MCC. However, we note that it performs better on F1 on OK tags. Therefore, we also ensemble the two models trained on each set of synthetic data through a linear combination. This yields comparable or better performance than the model trained on human-annotated data according to the main metric, MCC.

In Table 3, we compare the models trained on human-annotated data to our synthetic data for predicting sentence-level HTER scores. Again our synthetic data from NMT-generated translations outperforms MLM-rewriting data. Both underperform models trained on human-annotated data, but when combined they significantly improve and even outperform human-annotated for En-Zh. This once again suggests that the two forms of synthetic data are complementary and provide valuable signals for QE.

[4] The results reported in Tables 2 and 3 are evaluated on the test set provided (test20).

size	MCC	F1-BAD	F1-OK
English-German			
100k	37.72	46.23	83.99
200k	38.45	46.79	84.27
All (459k)	38.68	46.78	83.85
English-Chinese			
50k	53.07	66.58	83.65
100k	53.88	67.13	84.10
All (189k)	53.42	66.86	84.47

Table 4: Ablation study of synthetic data amounts.

5 Discussion

In this section, we further analyze how the quantity of synthetic data impacts performance, and what types of errors are represented in each of the MLM and NMT portions of the synthetic data.

5.1 Amount of Synthetic Data

As previously observed, the amount of synthetic data is orders of magnitude larger than the amount of human-annotated data. It begs the question: How much benefit do we get from smaller amounts of synthetic data? To analyze how the quantity of synthetic data affects QE performance, we conduct an ablation study of word-level QE.[5] As shown in Table 4, using only about half of the synthetic data generated (200k for En-De and 100k for En-Zh) is comparable to using the full generated set. While this suggests an upper-bound in performance to training on synthetic data. The ablation also suggests that this synthetic process can yield good performance with even a small amount of synthetic data.

5.2 Error Analysis

In addition to the performance, we posit that there are essential differences between MLM and NMT synthetic data. To test that, bilingual volunteers qualitatively analyzed the types of mistakes from MLM rewrites vs traditional NMT translations. The major reported differences in error types are:

1. Deletions from NMT translations appear more natural and do not destroy the sentence fluency. However, deletions in MLM rewrites are more destructive (e.g., "new york restaurants" vs "new restaurants" The semantics is changed).

2. Most incorrect insertions or deletions from NMT translations are due to re-ordering words. (e.g., "on 2020 in california" vs "in california on 2020") However insertions with MLM-rewrites introduces seemingly random words.

3. NMT translations often have semantically distant word substitutions. However, MLM-rewrites tend to substitute similar words (e.g., "strong tea" vs "powerful tea").

In summary, NMT translations and MLM-rewrites appear to generate different types of errors – the former leads to more subtle errors while the latter often introduces more catastrophic errors. Since a high-quality QE model should be able to detect both types of errors, ensembling the models trained on these two forms of synthetic data indeed is expected to outperform using only one form of synthetic data.

6 Conclusions and Future Work

In this work we devise a technique for building word and sentence-level QE models by creating synthetic training data. By training an off-the-shelf model on our synthetic data, we achieve performance comparable to and often better than training on human-annotated data. This technique for data synthesis can be invaluable if human annotation is difficult to come-by, for example when dealing with low-resource scenarios.

This work can be extended in various ways. While we investigate the scenario of utilizing solely synthetic data, further work can study the effects of augmenting human-labeled data with synthetic data. Further work can analyze the efficacy of this technique into low-resource language pairs where such human-annotation is difficult to obtain. Additionally, instead of a simple MLM re-writer, adversarial training to generate and detect errors could provide more realistic synthetic data.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

[5] The ablation study is only trained on word-level data.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020a. A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. *arXiv preprint arXiv:2002.00761*.

Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk. 2020b. Searching the web for cross-lingual parallel data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2417–2420.

Thierry Etchegoyhen, Eva Martínez Garcia, and Andoni Azpeitia. 2018. Supervised and unsupervised minimalist quality estimators: Vicomtech's participation in the wmt 2018 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 782–787.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *TACL*.

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.

Fabio 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. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019b. Openkiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Erwan Moreau and Carl Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. In *7th Workshop on Statistical Machine Translation*, page 120.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. Escape: a large-scale synthetic corpus for automatic post-editing. *arXiv preprint arXiv:1803.07274*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Maja Popović. 2012. Morpheme-and pos-based ibm1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Matthew Snover and Michael R Brent. 2001. A bayesian model for morpheme and paradigm identification. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 490–498.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764. Association for Computational Linguistics.

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. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Shuo Sun, Marina Fomicheva, Frédéric Blain, Vishrav Chaudhary, Ahmed El-Kishky, Adithya Renduchintala, Francisco Guzmán, and Lucia Specia. 2020. An exploratory study on multilingual quality estimation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 366–377. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling. *arXiv preprint arXiv:1901.00158*.