MMT's Submission for the WMT 2023 Quality Estimation Shared Task

Yulong Wu[1], Viktor Schlegel[1, 2], Daniel Beck[3] and Riza Batista-Navarro[1]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] ASUS Intelligent Cloud Services (AICS), Singapore
[3] School of Computing and Information Systems, University of Melbourne, Australia
{yulong.wu, riza.batista}@manchester.ac.uk
viktor_schlegel@asus.com, d.beck@unimelb.edu.au

Abstract

This paper presents our submission to the WMT 2023 Quality Estimation (QE) shared task 1 (sentence-level subtask). We propose a straightforward training data augmentation approach aimed at improving the correlation between QE model predictions and human quality assessments. Utilising eleven data augmentation approaches and six distinct language pairs, we systematically create augmented training sets by individually applying each method to the original training set of each respective language pair. By evaluating the performance gap between the model before and after training on the augmented dataset, as measured on the development set, we assess the effectiveness of each augmentation method. Experimental results reveal that synonym replacement via the Paraphrase Database (PPDB) yields the most substantial performance boost for language pairs English-German, English-Marathi and English-Gujarati, while for the remaining language pairs, methods such as contextual word embeddings-based words insertion, back translation, and direct paraphrasing prove to be more effective. Training the model on a more diverse and larger set of samples does confer further performance improvements for certain language pairs, albeit to a marginal extent, and this phenomenon is not universally applicable. At the time of submission, we select the model trained on the augmented dataset constructed using the respective most effective method to generate predictions for the test set in each language pair, except for the English-German. Despite not being highly competitive, our system consistently surpasses the baseline performance on most language pairs and secures a third-place ranking in the English-Marathi[1].

1 Introduction

Quality Estimation (QE) strives to assess the output of Machine Translation (MT) systems without the availability of a reference translation of known high quality (Blatz et al., 2004; Specia et al., 2009, 2013; Kanojia et al., 2021). This capability serves as a valuable asset for expediting and cost-effectively facilitating the evaluation phases throughout the development cycle of MT systems.

In this paper, we describe our contribution to the QE shared task at the Eighth Conference on Machine Translation (WMT23). We participate in the Task 1 of the shared task and we specifically focus on the sentence-level subtask, which centers on predicting the quality score of neural MT outputs at the sentence level without access to reference translations. Our study encompasses six language pairs: *English-German (En-De)*, *English-Marathi (En-Mr)*, *English-Hindi (En-Hi)*, *English-Tamil (En-Ta)*, *English-Telegu (En-Te)*, *English-Gujarati (En-Gu)*, with annotations derived in two different ways: multi-dimensional quality metrics (MQM) (Freitag et al., 2021) and direct assessments (DA) (Fomicheva et al., 2022). Participating systems are assigned the task of predicting the quality score (MQM or DA) for each source-target sentence pair, and their performance is evaluated using Spearman's rank correlation coefficient as the primary metric, supplemented by the Kendall and Pearson coefficients as secondary metrics for assessment.

Our approach investigates the potential to enhance the performance of QE models by exposing them to a diverse range of training examples. To this end, we identify eleven different data augmentation methods and apply each of them individually to augment the training set for each language pair. Our results reveal that, for most language pairs, these methods result in varying degrees of performance improvement, with the most effective methods being synonym substitution using the PPDB, words insertion guided by contextual word embeddings, back-translation, and direct paraphrasing. We also show that for some language pairs, it

[1]Our code and data are available at https://github.com/Yulong-W/DataAug-QE.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 856–862
December 6–7, 2023.

is feasible to further enhance the model's performance by training it on an augmented set formed through the amalgamation of part or all of the said augmentation methods; however, the extent of improvement remains constrained. For each language pair except English-German, we generate predictions utilising the model trained on the augmented dataset constructed through the respective most effective method. Although our submission may not be considered highly competitive, they consistently achieve significantly improved performance compared to the organisers' baseline for the majority of language pairs. Notably, for the English-Marathi pair, our submission ranks third place with the Spearman score of 0.650. This observation indicates that the training data augmentation approach may hold particular promise and offer advantages when applied to the English-Marathi language pair.

2 Methodology

As mentioned above, we identified a total of eleven distinct data augmentation methods, as detailed in Table 1. For all the given source sentences and their corresponding MT hypothesis in the training dataset for each language pair, each method is independently applied only to the source sentences, leading to the creation of the respective transformed source-target sentence pairs. Our hypothesis posits that training the QE model on the augmented training set, which incorporates these transformed instances, holds the potential to bolster its performance. For each original instance, we generated one augmented sample per method and assigned to the augmented data the same quality score as the original translation hypothesis. However, it is noteworthy that certain methods, such as AS and RD, possess the potential to alter the meaning of the source-side sentence (Kanojia et al., 2021), consequently inducing changes in MT output and, by extension, the assigned quality label. In such instances, there is a likelihood of introducing noises to the augmented training dataset. A systematic exploration of the meaning-preserving capacity of these perturbation methods and the impact of those introduced noises at training time on the performance of the model necessitates further investigation.

3 Experiments

In this section, we describe our experimental settings, present the results achieved on the development and test sets, and perform analysis derived from our experimental findings. We additionally offer insights into the influence of the quantity of augmented training examples on the performance of the QE model.

	Data Augmentation Method
m_1	**WordNet-based Synonym Substitution (WSS):** Substitute words by WordNet's synonym (Fellbaum, 1998)
m_2	**PPDB-based Synonym Substitution (PSS):** Substitute words with synonyms from English PPDB (Pavlick et al., 2015)
m_3	**Antonym Substitution (AS):** Substitute random words with their antonyms
m_4	**Random Swap (RS):** Swap words in the sentence randomly
m_5	**Random Deletion (RD):** Delete words in the sentence randomly
m_6	**Spelling Mistake Substitution (SMS):** Substitute content words randomly by spelling mistake words dictionary
m_7	**GloVe Similarity-based Substitution (GSS):** Substitute words based on GloVe similarity (Pennington et al., 2014)
m_8	**Contextual Words Insertion (CWI):** Insert words using contextual word embeddings from the RoBERTa-base
m_9	**Contextual Words Substitution (CWS):** Substitute words by contextual word embeddings from the RoBERTa-base (Liu et al., 2019)
m_{10}	**Back Translation (BT)**
m_{11}	**Direct Paraphrasing (DP)**

Table 1: Various data augmentation methods.

3.1 Experimental Settings

Language Pairs (LPs). We conducted experiments on six language pairs. The training, development, and test datasets for each language pair utilised in our study are accessible via the shared task website[2], and we present the dataset statistics in Table 2. We applied each data augmentation method on the source sentences in the training set of each language pair.

Models. Our training methodology adheres to the PyTorch-based COMET framework (Rei et al., 2020), with the foundational pre-trained model be-

[2] https://wmt-qe-task.github.io/subtasks/task1/

LPs	Training	Development	Test
		MQM	
En-De	28909	1005	1897
		DA	
En-Mr	26000	1000	1086
En-Hi	7000	1000	1074
En-Ta	7000	1000	1075
En-Te	7000	1000	1075
En-Gu	7000	1000	1075

Table 2: Number of examples in the training, development and test set, respectively, for each language pair.

ing XLM-RoBERTa-large (Conneau et al., 2020). We fine-tuned the pre-trained XLM-RoBERTa-large model on the original and the augmented training sets for each language pair, respectively and evaluated them on the development set. The best-performing model was chosen from those trained on the corresponding augmented training datasets (in the case of English-German, the chosen model was trained on the augmented dataset created by applying the top four[3] effective data augmentation techniques to each source sentence) to generate the predictions on the test set. All experiments were conducted using 2 16GB Nvidia v100 GPUs.

Data Augmentation Methods. Methods WSS to BT: We utilised the NLPAug library (Ma, 2019) to perform the augmentation. In method PSS, we used the small size English PPDB (Pavlick et al., 2015). In the absence of any synonymous expressions documented for all the words within a source-side sentence in methods WSS and PSS, the augmented sample will persist unaltered in comparison to its original version. For methods WSS to CWS, the percentage of word will be augmented is set to the default value of 0.3, as in the implementation of the NLPAug library (Ma, 2019). In method BT, a sentence is translated from English to German, then back to English to obtain its paraphrased version (Ng et al., 2019). Method DP: Direct paraphrasing was performed by soliciting a Generative Pre-trained Transformer (GPT) (Brown et al., 2020) series model, specifically GPT-3.5-turbo, to generate responses for the prompt: *Generate a similar paraphrase for this sentence: [source sentence]*, using the OpenAI ChatGPT API.

[3]At the time of results submission, this number (i.e., 4) was randomly set. However, as illustrated in Figure 1, augmenting the training dataset for the English-German language pair using the best two methods yielded the most optimal performance.

3.2 Evaluation Results and Discussion

Table 3 illustrates the performance gap of the QE model on the development set before and after training on each augmented dataset created through the respective data augmentation method for each examined language pair. As can be seen from Table 3, in the majority of instances, the training data augmentation approaches demonstrated their effectiveness in enhancing the performance of the QE model. In the following, we discuss the observations for all the studied language pairs.

English-German. Method PSS exhibited the most significant performance improvement across all three evaluation metrics. Augmenting the training set with method CWI yielded the same improvements in terms of Spearman and Kendall correlations compared to augmenting it with PSS, albeit resulted a lower Pearson score. However, it was observed that presenting the model with modified training examples generated using method WSS and RS did not contribute to the enhancement of Spearman correlation. In fact, it even had an adverse effect, causing a slight reduction (0.3%) in the Kendall score.

English-{Marathi, Gujarati}. Training the model on the augmented set incorporating examples generated by substituting words with synonyms from PPDB (method PSS) proved to be the most effective approach in enhancing the correlation between the predictions of the model and human judgments of quality, with Spearman correlation increased by 6.8% and 7.1%, respectively. Other types of approaches also resulted in varying degrees of performance improvement.

English-{Hindi, Tamil}. For the English-Hindi language pair, augmenting the training set with both CWI and DP has been observed to yield identical improvements in terms of Spearman and Kendall correlations, emerging as the most effective approach. In the case of English-Tamil, the most notable enhancement was achieved by paraphrasing the source sentences in the original training dataset using the GPT-3.5-turbo model (method DP), as measured by Spearman and Kendall correlations. However, concerning the Pearson metric, method BT (back-translation) led to the most substantial improvements for both language pairs, amounting to 12.6% and 10.8%, respectively.

Method	En-De	En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Average
				Spearman/Kendall/Pearson			
orig.	0.433/0.328/0.393	0.499/0.349/0.593	0.479/0.336/0.476	0.541/0.379/0.604	0.449/0.302/0.365	0.524/0.373/0.523	
WSS	0.433/0.327/0.404	0.518/0.363/0.607	0.492/0.346/0.512	0.548/0.386/0.652	0.430/0.291/0.362	0.536/0.384/0.579	
	0.0/-0.3/+2.8	+3.8/+4.0/+2.4	+2.7/+3.0/+7.6	+1.3/+1.8/+7.9	-4.2/-3.6/-0.8	+2.3/+2.9/+10.7	1.0
PSS	0.451/0.342/0.438	0.533/0.376/0.624	0.501/0.352/0.522	0.558/0.395/0.659	0.435/0.293/0.367	0.561/0.401/0.596	
	+4.2/+4.3/+11.5	**+6.8/+7.7/+5.2**	+4.6/+4.8/+9.7	+3.1/+4.2/+9.1	-3.1/-3.0/+0.5	**+7.1/+7.5/+14.0**	3.8
AS	0.442/0.335/0.408	0.516/0.364/0.616	0.503/0.354/0.528	0.542/0.382/0.662	0.431/0.294/0.360	0.547/0.392/0.586	
	+2.1/+2.1/+3.8	+3.4/+4.3/+3.9	+5.0/+5.4/+10.9	+0.2/+0.8/+9.6	-4.0/-2.6/-1.4	+4.4/+5.1/+12.0	1.8
RS	0.433/0.327/0.400	0.517/0.364/0.617	0.496/0.349/0.527	0.549/0.388/0.654	0.430/0.293/0.365	0.550/0.394/0.588	
	0.0/-0.3/+1.8	+3.6/+4.3/+4.0	+3.5/+3.9/+10.7	+1.5/+2.4/+8.3	-4.2/-3.0/0.0	+5.0/+5.6/+12.4	1.6
RD	0.442/0.335/0.424	0.507/0.356/0.601	0.494/0.347/0.526	0.551/0.389/0.648	0.437/0.296/0.373	0.552/0.397/0.591	
	+2.1/+2.1/+7.9	+1.6/+2.0/+1.3	+3.1/+3.3/+10.5	+1.8/+2.6/+7.3	-2.7/-2.0/+2.2	+5.3/+6.4/+13.0	1.9
SMS	0.435/0.329/0.404	0.517/0.363/0.604	0.500/0.351/0.525	0.547/0.386/0.656	0.439/0.300/0.369	0.552/0.396/0.590	
	+0.5/+0.3/+2.8	+3.6/+4.0/+1.9	+4.4/+4.5/+10.3	+1.1/+1.8/+8.6	-2.2/-0.7/+1.1	+5.3/+6.2/+12.8	2.1
GSS	0.440/0.333/0.417	0.521/0.366/0.610	0.500/0.352/0.522	0.555/0.392/0.653	0.435/0.298/0.369	0.547/0.392/0.578	
	+1.6/+1.5/+6.1	+4.4/+4.9/+2.9	+4.4/+4.8/+9.7	+2.6/+3.4/+8.1	-3.1/-1.3/+1.1	+4.4/+5.1/+10.5	2.4
CWI	0.451/0.342/0.430	0.518/0.364/0.619	0.509/0.358/0.535	0.546/0.385/0.661	0.450/0.308/0.384	0.554/0.397/0.590	
	+4.2/+4.3/+9.4	+3.8/+4.3/+4.4	**+6.3/+6.5**/+12.4	+0.9/+1.6/+9.4	**+0.2/+2.0/+5.2**	+5.7/+6.4/+12.8	3.5
CWS	0.444/0.337/0.412	0.513/0.359/0.609	0.506/0.355/0.525	0.554/0.392/0.656	0.442/0.302/0.377	0.543/0.387/0.588	
	+2.5/+2.7/+4.8	+2.8/+2.9/+2.7	+5.6/+5.7/+10.3	+2.4/+3.4/+8.6	-1.6/0.0/+3.3	+3.6/+3.8/+12.4	2.6
BT	0.441/0.334/0.423	0.522/0.366/0.612	0.504/0.354/0.536	0.559/0.397/0.669	0.435/0.295/0.373	0.552/0.395/0.593	
	+1.8/+1.8/+7.6	+4.6/+4.9/+3.2	+5.2/+5.4/**+12.6**	+3.3/+4.7/**+10.8**	-3.1/-2.3/+2.2	+5.3/+5.9/+13.4	2.8
DP	0.440/0.333/0.418	0.514/0.361/0.601	0.509/0.358/0.534	0.568/0.400/0.607	0.439/0.296/0.360	0.539/0.384/0.562	
	+1.6/+1.5/+6.4	+3.0/+3.4/+1.3	**+6.3/+6.5**/+12.2	**+5.0/+5.5**/+0.5	-2.2/-2.0/-1.4	+2.9/+2.9/+7.5	2.8

Table 3: The performance (%) of the QE model trained on the original, and the augmented training sets generated through applying the data augmentation methods, when evaluated on the development set for the examined language pairs. Values shown in the shaded areas are changes (%) relative to the original performance of the model, with the rightmost column shows their averages in terms of Spearman correlation. We highlight the values that denote the most substantial performance improvements across the Spearman, Kendall, and Pearson metrics.

English-Telegu. Our experimental training data augmentation approach was found to be notably ineffective when applied to the language pair English-Telegu. As shown in Table 3, in regard to Spearman and Kendall correlations, only method CWI yielded slight performance improvements, while the other approaches predominantly resulted in a decrease in the performance of the model. Indeed, these alternative approaches led to varying degrees of performance decline, with the most significant decrease being 4.2% in Spearman and 3.6% in Kendall, respectively. This may be attributed to the heightened sensitivity of English to Telegu translation concerning modifications applied to the source sentences. Consequently, noises might be introduced during the process of augmenting the training set, thereby contributing to a decline in the performance of the QE model.

Overall, our investigation revealed that, for the examined language pairs, method PSS yielded a relative performance increase of 3.8% on average, establishing itself as the most effective, with the second-best being CWI (3.5%). Interestingly, both method BT and method DP, designed for paraphrasing purposes, exhibited an identical average performance improvement of 2.8%. Conversely, the average increase was only 1.0% for method WSS, despite sharing the same objective of synonym substitution with method PSS. This suggests that employing synonym substitution via the English PPDB confers greater benefits to enhancing the performance of the QE model compared to performing it via WordNet. Furthermore, potential meaning alternation methods, such as AS and RD (Kanojia et al., 2021), yielded a lower average enhancement compared to some meaning-preservation methods like BT and DP. However, additional experimental confirmation is requisite.

3.3 Official Test Results

Based on the insights derived from Table 3, we systematically selected the most efficacious approach to augment the training set for each language pair and trained the respective model. Subsequently, we utilised each resulting model to generate predictions on the corresponding test dataset. For English-German language pair, it was observed that the performance of the QE model (0.303 Spearman),

when trained on the augmented dataset generated by applying PSS, was inferior to the baseline score determined during our initial test phase. Therefore, we took the initiative to curate a new training set wherein four augmented examples were generated for each original sample, employing the top four data augmentation methods identified as correspondingly effective. We then employed the re-trained model to generate quality predictions for English-German pair. The performance of our submitted models is presented in Table 4[4].

LPs	Spearman	Kendall	Pearson
	MQM		
En-De	0.316	0.237	0.221
	DA		
En-Mr	0.650	0.466	0.663
En-Hi	0.494	0.345	0.570
En-Ta	0.547	0.384	0.531
En-Te	0.337	0.228	0.281
En-Gu	0.540	0.386	0.581

Table 4: Official results of our submission to the WMT sentence-level QE shared task 2023.

Our most promising results were observed in language pair English-Marathi, where our submission ranked third among the six participating teams. This highlights the effectiveness of the training data augmentation approach in improving the capability of the QE model to precisely predict the quality score of English-Marathi translation pairs in the absence of a reference. However, when considering English-German, despite training the model on an augmented dataset with larger and more diverse samples, its performance still falls below the baseline score (0.340 Spearman). This discrepancy suggests that data augmentation approach may not be as efficient in enhancing the QE performance for this specific language pair. Nevertheless, we observed that this performance (0.316 Spearman) remains slightly superior to that achieved with the training set containing fewer augmented samples (0.303 Spearman), which indicates that increasing the number of augmented training examples might contribute to enhancing the performance of the model, and we provide further elaboration in Section 3.4 below. In contrast, for the remaining four language pairs we investigated, the performance of our submitted models consistently outperformed the baseline score. Specifically, our submission demonstrated a notable enhancement over the baseline score in Spearman correlation for English-Hindi (+0.213), English-Telugu (+0.144), and English-Gujarati (+0.203), while the improvement for English-Tamil was comparatively less pronounced. Despite the above-baseline performance achieved, our submission is presently ranked fifth in these language pairs, signifying the necessity for additional investigation and refinement of our approach to attain elevated performance levels.

[4]A comparison of our results with the organiser's baseline and submissions from other participating teams is available at http://www2.statmt.org/wmt23/quality-estimation-task_results.html.

3.4 Impact of Training Example Quantity

Thus far, a singular augmented example has been generated corresponding to each defined augmentation method for every original training sample in our studied language pairs, with the exception of English-German. To examine the impact of the number of augmented samples on the performance of the QE model and to explore potential complementarity among these augmentation techniques, we trained the models for each language pair on augmented training sets of varying sizes, generated by employing the respective top N effective augmentation methods (where N ranges from 1 to 11), and then assessed their performance, as shown in Figure 1.

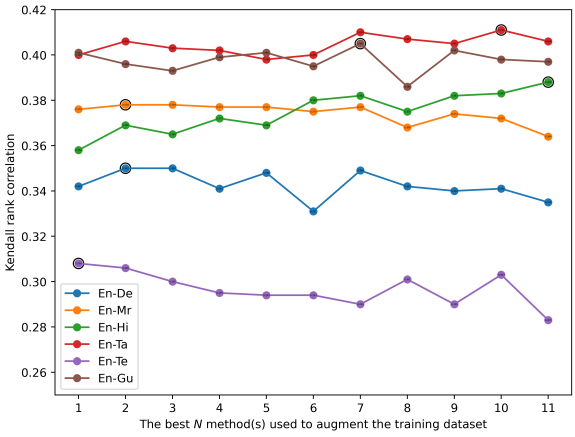

Figure 1: The performance of the QE models on the development set across six language pairs, trained on augmented datasets generated utilising the respective best N data augmentation methods. The optimal performance was denoted by encircling the respective data point with a black circle.

It can be seen from Figure 1 that for five lan-

guage pairs English-{Tamil, Gujarati, Marathi, Hindi, German}, increasing the number of augmented training samples can enhance the model's performance, although this phenomenon is not universal for certain language pairs, such as English-Tamil. However, we observed that there was a negligible degree of performance improvement across these five language pairs, with the most notable enhancement being merely 0.03 (from 0.358 to 0.388), as demonstrated in the case of English-Hindi. Even worse, for the language pair English-Telegu, exposing the model to a more diverse set of training examples resulted in a decline in performance. Notably, training the model on an augmented set comprising eleven augmented samples per original instance led to the nadir in performance, recording a value of 0.283. This underscores the constraints of current data augmentation methods in boosting the efficacy of the QE model, emphasizing the imperative to devise more effective approaches. Nonetheless, a positive insight has been discerned; the language pair English-Hindi appears to derive particular benefits from the augmentation of training examples. As the number of applied top N augmentation methods increased, the performance of the model consistently surpassed that of the model with only the best one applied, notwithstanding fluctuations in performance. Finally, based on the empirical findings depicted in Figure 1, definitive conclusion regarding the complementarity of specific data augmentation approaches cannot be drawn, as it is inherently specific to each language pair. For instance, the efficacy of combining the best two augmentation methods was observed in the English-{Marathi, German} pairs, whereas for English-{Tamil, Gujarati, Hindi}, optimal performance was attained through the amalgamation of the top 10, 7, and 11 training data augmentation methods, respectively.

4 Conclusion

In this paper, we proposed a training data augmentation approach to the WMT 2023 sentence-level QE shared task. We systematically identified eleven various data augmentation methods and applied each of them individually on the source-side sentences to generate augmented training samples for the six studied language pairs. The experimental results demonstrated that in most cases, these methods can enhance the correlation between the predictions of the QE model and human-provided quality scores to varying degrees, albeit not to a significant extent. In addition, we show that training the model on the augmented set, generated through the combination of these methods, contributed further to performance enhancement, although this phenomenon was not universally observed and the degree of improvement was at a negligible level. Our methodology yielded a third-ranking outcome for English-Marathi and a fifth-place ranking for other DA annotated language pairs, among the submissions from the six teams. In terms of future work, we intend to explore other more effective augmentation approaches and extend our study to encompass a more diverse array of language pairs and QE models.

Limitations

The work presented in this paper should be considered preliminary, given that we exclusively conducted experiments employing a training data augmentation approach and assessed its impact solely on the original development set. There is ample room for further exploration into the robustness of the QE model without any augmentation interventions on the studied perturbations and the impact of these proposed perturbations, when applied during training, on the capability of the QE systems to identify critical errors in translation resulting from modifications to the source sentences. Moreover, the extent to which the introduced perturbations may alter the meaning of the source-side sentences remains unclear, necessitating further investigation.

Acknowledgements

The authors would like to thank the University of Manchester Department of Computer Science Kilburn Scholarship, the Manchester-Melbourne-Toronto Research Fund 2022, and the Turing Scheme for supporting this work. We also express our sincere gratitude for the invaluable comments and suggestions provided by the anonymous reviewers and acknowledge the support of the Computational Shared Facility at The University of Manchester in facilitating the execution of our experiments.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation

for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. 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, Sofia, Bulgaria. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.