Ensemble Fine-tuned mBERT for Translation Quality Estimation

Shaika Chowdhury *
University of Illinois at Chicago, US
schowd21@uic.edu

Naouel Baili
IQVIA, US
naouel.baili@iqvia.com

Brian Vannah
IQVIA, US
brian.vannah@iqvia.com

Abstract

Quality Estimation (QE) is an important component of the machine translation workflow as it assesses the quality of the translated output without consulting reference translations. In this paper, we discuss our submission to the WMT 2021 QE Shared Task. We participate in Task 2 sentence-level sub-task that challenge participants to predict the HTER score for sentence-level post-editing effort. Our proposed system is an ensemble of multilingual BERT (mBERT)-based regression models, which are generated by fine-tuning on different input settings. It demonstrates comparable performance with respect to the Pearson's correlation and beats the baseline system in MAE/ RMSE for several language pairs. In addition, we adapt our system for the zero-shot setting by exploiting target language-relevant language pairs and pseudo-reference translations.

1 Introduction

Progress in machine translation (MT) has accelerated due to the introduction of deep learning based approaches, dubbed as neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2014). Several metrics (e.g., BLEU (Papineni et al., 2002), METEOR (Agarwal and Lavie, 2008)) are used to automatically evaluate the quality of the translations outputted by the NMT systems. However, these evaluation metrics require comparing the NMT outputs against human-prepared reference translations, which cannot be readily obtained. To tackle this predicament, recently quality estimation (QE) (Blatz et al., 2004; Specia et al., 2018) has emerged as an alternative evaluation approach for NMT systems. QE obviates the need for human judgements and hence can be efficiently integrated into the dynamic translation pipeline in the industry setting.

QE is performed at different granularity (e.g., word, sentence, document) (Kepler et al., 2019a); in this work we focus on the sentence-level post-editing effort task, which predicts the quality of the translated sentence as a whole in terms of the number of edit operations that need to be made to yield a post-edited translation, termed as HTER (Snover et al., 2006).

Sentence-level QE using neural approaches is generally treated as a supervised regression problem involving mainly two steps. In the first step, an encoder is used to learn vector representation/s of the source and translation sentences. While in the second step, the learned representations are passed through a sigmoid output layer to estimate the HTER score. These two steps can be performed either with a single model in an end-to-end fashion (e.g., Bi-RNN (Ive et al., 2018)), or using two separate models (e.g., POSTECH (Kim et al., 2017)). The different QE systems vary in their choice of the encoder, which range from RNN-based to Transformer-based models.

In this work, we leverage the fine-tuning capability of a Transformer-based encoder, namely the mBERT (Devlin et al., 2018) pre-trained model. Alongside the standard practice of feeding both the source and target (i.e., translation) sentences as the input sequence (Kepler et al., 2019a; Kim et al., 2019), we also explore other input settings based on only the target-side sentences (i.e., monolingual context). To this end, our final QE system is an ensemble of several mBERT models [1], each generated by fine-tuning on a different input combination comprising the source and/or target sentences. We experiment with the following three input settings: (1) both source and target, (2) just target and (3) both target and a randomly-sampled target sentence in the data forming the input se-

*work done during internship at IQVIA

[1] we also experimented with XLM-RoBERTa (Conneau et al., 2019) as the component model in our preliminary run; however, the results were worse compared to mBERT

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 897–903
November 10–11, 2021.

quence. Empirical analysis on 6 language pairs shows that the ensemble model is able to perform better than the individual fine-tuned models. Moreover, we provide experimental results for zero-shot QE, where training data for the test language pair is not available. This we tackle by improvising on the available training/dev data that match the target language of the test language pair and also by generating the pseudo-reference translations in that language.

2 Data

We use the WMT21 QE Shared Task 2 sentence-level data (Specia et al., 2021; Fomicheva et al., 2020a,b) for the following 7 language pairs: English-German (En-De), Romanian-English (Ro-En), Estonian-English (Et-En), Nepalese-English (Ne-En), Sinhala-English (Si-En), Russian-English (Ru-En) and Khmer-English (Km-En). Source-side data for each language pair includes sentences from Wikipedia articles, with part of the data gathered from Reddit articles for Ru-En. To obtain the translations, state-of-the-art MT models (Vaswani et al., 2017) built using fairseq toolkit (Ott et al., 2019) were used. The label for this task is the HTER score for the source-translation pair. Annotation was performed first at the word-level with the help of TER [2] tool. The word-level tags were then aggregated deterministically to obtain the sentence-level HTER score. The training, development, test and blind test data sizes for each language pair (except Km-En) are 7K, 1K, 1K and 1K instances respectively. As Km-En language pair was introduced for zero-shot prediction, only the test data containing 990 source and translation sentences was provided.

3 Our Approach

A key innovation in recent neural models lies in learning the contextualized representations by pre-training on a language modeling task. One such model, the multilingual BERT (mBERT) [3], is a transformer-based masked language model that is pre-trained on monolingual Wikipedia corpora of 104 languages with a shared word-piece vocabulary. Training the pre-trained mBERT model for a supervised downstream task, aka *finetuning*, has dominated performance across a wide spectrum of NLP tasks (Devlin et al., 2018). Our proposed approach leverages this fine-tuning capability of mBERT so as to form the component models in the ensemble QE system (Section 3.3). That is, each component model is a re-purposed mBERT that is fine-tuned for the sentence-level HTER score prediction task on one of the three input settings discussed in Section 3.2.

3.1 Fine-tuning mBERT for Regression

mBERT's model architecture is similar to BERT [4] and contains the following parameter settings: 12 layers, 12 attention heads and 768 hidden dimension per token. However, the only difference is that mBERT is trained on corpora of multiple languages instead of just on English. This enables mBERT to share representations across the different languages and hence can be conveniently used for all language pairs in the WMT21 data.

We first load the pre-trained mBERT model [5] and use its weights as the starting point of fine-tuning. The pre-trained mBERT is then trained on QE-specific input sequences (Section 3.2) for a few epochs such that the constructed sequence X is consumed by mBERT to output the contextualized representation $\mathbf{h} = (h_{CLS}, h_{x_1}, h_{x_2}, ..h_{x_T}, h_{SEP})$. Here, $[CLS]$ is a special symbol that denotes downstream classification and $[SEP]$ is for separating non-consecutive token sequences. Considering the final hidden vector of the $[CLS]$ token as the aggregate representation, it is then passed into the output layer with sigmoid activation to predict the HTER score:

$$y = sigmoid(\mathbf{W} \cdot \mathbf{h}_{CLS} + \mathbf{b}) \qquad (1)$$

$\mathbf{W}$ is a weight matrix for sentence-level QE fine-tuning that is trained along with all the parameters of mBERT end-to-end.

3.2 Input Settings

We construct the input sequence for each language pair in the following three ways:

SRC-MT: Given a source sentence $\mathbf{s} = (s_1, s_2, ...s_N)$ from a source language (e.g., English) and its translation $\mathbf{t} = (t_1, t_2, ...t_M)$ from a target language (e.g., German), we concatenate them together as $X = ([CLS], t_1, t_2, ...t_M, [SEP], s_1, s_2, ...s_N, [SEP])$ to form the input sequence.

[2] http://www.cs.umd.edu/ snover/tercom/

[3] https://github.com/google-research/bert/blob/master/multilingual.md

[4] https://huggingface.co/bert-base-uncased

[5] https://huggingface.co/bert-base-multilingual-uncased

MT: The target sentence is only used to form the input sequence, $X = ([CLS], t_1, t_2, ...t_M, [SEP])$.

MT-MT':

Given the translation **t** for a source sentence **s**, we randomly sample another translation **t'** $= (t'_1, t'_2, ...t'_K)$ from the training data having HTER label close to **t** [6]. Although the source sentences for **t** and **t'** are different, we assume the additional monolingual context would help mBERT learn the correlating QE-specific features between **t** and **t'** for the target-side language. The resultant input sequence is $X = ([CLS], t_1, t_2, ...t_M, [SEP], t'_1, t'_2, ...t'_K, [SEP])$.

We fine-tune each of these mBERT models using AdamW optimizer (Kingma and Ba, 2014; Loshchilov and Hutter, 2017) for two epochs with a batch size of 32 and a learning rate of $2e^{-5}$.

3.3 Ensemble Model

To take advantage of the individual strengths of the three mBERT component models fine-tuned on the aforementioned input settings, we combine their HTER score predictions by training an ensemble model. In particular, we experiment with three different ensemble models - Gradient Boosting (Friedman, 2001), AdaBoost (Freund and Schapire, 1997) and Average. For Gradient Boosting and AdaBoost we use the implementation in scikit-learn [7] with 10-fold cross validation. The settings for Gradient Boosting are: number of estimators 600, learning rate 0.01, minimum number of samples 3 and other default settings. We use the default settings for AdaBoost. In Average ensembling, we average the HTER score predictions by the three mBERT models. Our system submission to WMT21 is based on Gradient Boosting as it gave the best performance on the test data, as shown in Table 1.

3.4 Zero-Shot QE

Performing sentence-level QE in the zero-shot setting presents a unique challenge as the QE system is expected to predict HTER scores for sentences in a test language pair (e.g., Km-En) without having been trained on any instances from that test language pair. We address this by training on language pairs in the WMT21 QE data that match the target-side language (i.e., En) in the test language pair. The reason we focus on the target-side language is because the component mBERT models in the proposed ensemble QE system are fine-tuned on monolingual input sequences in the target-side language, which could potentially help the QE system generalize on the unseen test language pair. We consider the training and development data for the following language pairs in WMT21 QE data: Ro-En, Si-En, Et-En. Additionally, we augment this data by generating pseudo-references in the target language. A *pseudo-reference* (Scarton and Specia, 2014) is a translation for a source sentence that is outputted by a different NMT system than the one that produced the actual translations (e.g., transformer-based translation system proposed in (Vaswani et al., 2017)) and has shown to improve sentence-level QE performance (Soricut and Narsale, 2012). We use Google Translate [8] to get the pseudo-references in En for the Ro, Si and Et source sentences. The HTER scores for the translation and pseudo-reference pairs are then obtained using the TER tool. We train the ensemble QE system on the combined WMT21 QE data and the pseudo-reference parallel data, and test on the unseen test language pair.

Table 1: Performance of ENSBRT with different ensemble methods on the En-De test set.

	Avg	**AdaBoost**	**GradBoost**
Pearson's	0.266	0.458	**0.473**
Spearman's	0.249	0.436	**0.443**

4 Baseline

The baseline QE system (BASELINE) set by the WMT21 organizers this year is the Transformer-based Predictor-Estimator model (Kepler et al., 2019b; Moura et al., 2020). XLM-RoBERTa is used as the Predictor for feature generation. The baseline system is fine-tuned on the HTER scores and word-level tags jointly.

5 Results

Table 2 presents the experimental results of mBERT fine-tuned on the SRC-MT, MT and MT-MT'

[6] to ensure that **t'** is similar to **t**, we check that the difference between their HTER scores is within 0.1

[7] https://scikit-learn.org

[8] https://github.com/lushan88a/google_trans_new

Table 2: Performance in Pearson's correlation of mBERT fine-tuned with different input settings on the test set. ENSBRT is the proposed ensemble mBERT QE system.

	En-De	**Ro-En**	**Ru-En**	**Si-En**	**Et-En**	**Ne-En**
SRC-MT	0.389	0.793	0.400	0.526	0.601	0.489
MT	0.469	0.762	0.374	0.552	0.580	0.491
MT-MT'	0.431	0.761	0.350	0.492	0.556	0.454
ENSBRT	**0.473**	**0.802**	**0.418**	**0.576**	**0.632**	**0.525**

Table 3: Performance of BASELINE and ENSBRT on the WMT21 blind test set for different language pairs. Bold indicates ENSBRT beats BASELINE in that metric.

		En-De	**Ro-En**	**Ru-En**	**Si-En**	**Et-En**	**Ne-En**	**Km-En**
BASELINE	Pearson's ↑	0.529	0.831	0.448	0.607	0.714	0.626	0.576
	MAE ↓	0.183	0.142	0.255	0.204	0.195	0.205	0.241
	RMSE ↓	0.129	0.115	0.188	0.159	0.149	0.160	0.196
ENSBRT	Pearson's ↑	0.519	0.795	0.376	0.522	0.666	0.572	0.529
	MAE ↓	**0.171**	0.171	**0.251**	0.206	**0.171**	**0.176**	0.262
	RMSE ↓	0.129	0.141	0.189	0.162	**0.132**	**0.139**	0.197

input settings, as well as the performance of the ensemble of the three mBERT models, which we call *ENSBRT*. First, comparing among the three input settings, it seems that mBERT exhibits competitive results even when it does not have knowledge of the source-side text in the MT and MT-MT' settings, in particular for the following language pairs - En-DE, Si-En, Ne-En. While the ensemble mBERT model, ENSBRT, outperforms the independent counterparts for all the language pairs. This shows that the ensemble method can help to balance out the weakness of any component model, thereby benefiting the sentence-level QE task overall. We also visualize ENSBRT's predictions against the ground truth HTER scores in Figure 1.

Table 3 compares the QE performance between the BASELINE and ENSBRT in terms of Pearson's correlation, RMSE and MAE on the WMT21 blind test set, for which the ground truth HTER scores were not available at the time. We submitted results for 6 language pairs (En-De, Ro-En, Ru-En, Si-En, Et-En, Ne-En) in the normal QE setting and one language pair (Km-En) for zero-shot prediction. ENSBRT demonstrates comparable performance to the BASELINE for Pearson's and outperforms it in either MAE or RMSE for the following language pairs: En-De, Ru-En, Et-En and Ne-En.

6 Conclusion

In this work, we describe the *ENSBRT* system submission to the WMT21 QE Shared Task. ENSBRT is based on fine-tuning the multilingual BERT pretrained model for sentence-level translation quality score prediction. We explore three different input settings for fine-tuning which include either bilingual or monolingual context, and combine the predictions of the three models using ensemble methods as our final system. Furthermore, zero-shot QE is facilitated by using labeled data for existing language pairs and pseudo-references that align with the target language of the unseen test data.

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

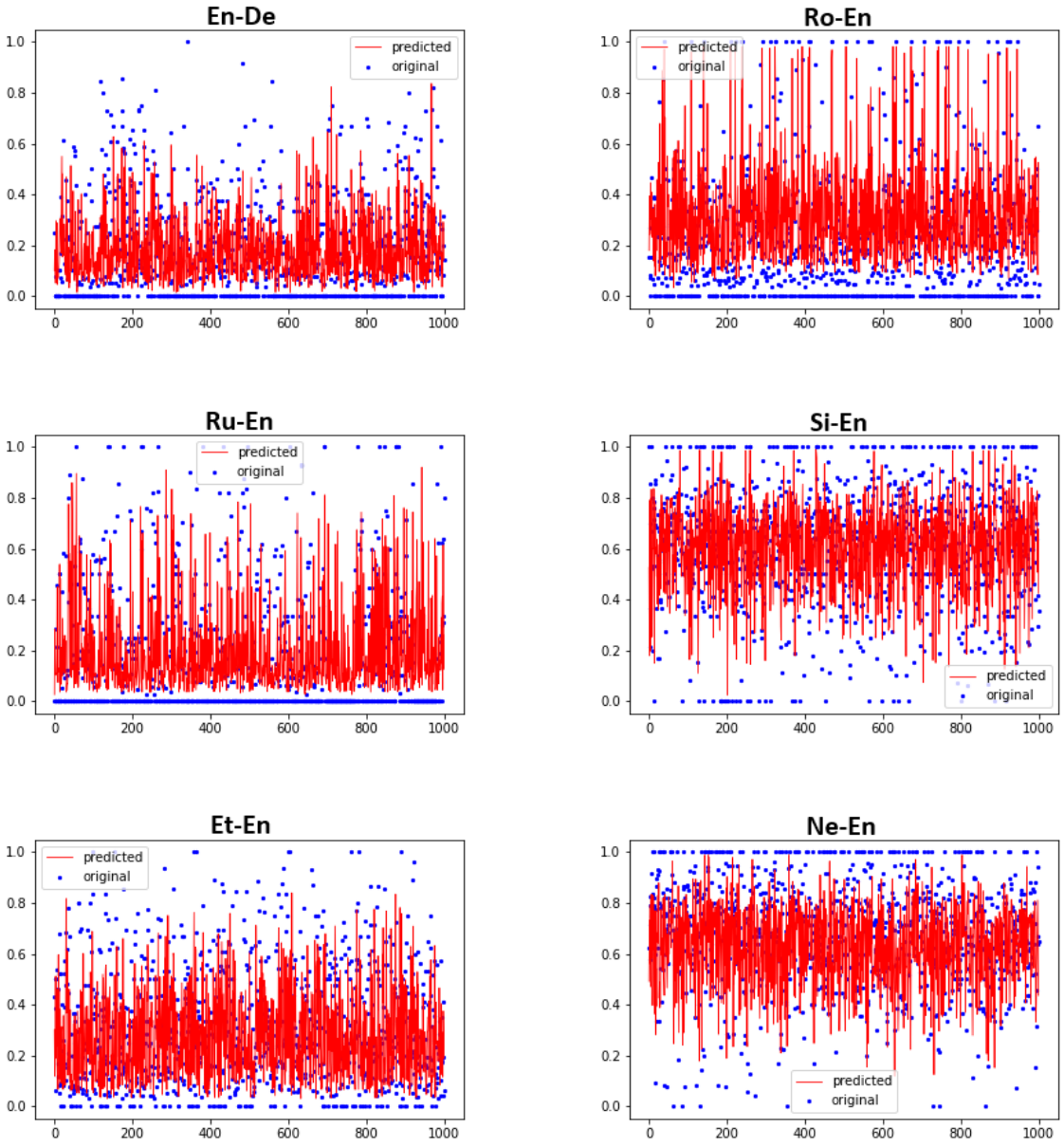

Figure 1: Visualization comparing HTER score predictions by ENSBRT (i.e., predicted (red)) against the gold labels (i.e., original (blue)) for 6 language pairs on the test set. X-axis represents each data point and Y-axis is the HTER score. The closer the corresponding red line and blue dot are to each other the better, as we expect the HTER prediction to be same as or close to the ground truth.

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.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.10352*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019b. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. 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):1–22.

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Joao Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André FT Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036.

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. *arXiv preprint arXiv:1904.01038*.

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

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 101–108.

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 the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Radu Soricut and Sushant Narsale. 2012. Combining quality prediction and system selection for improved automatic translation output. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 163–170.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.