# NJU Submissions for the WMT19 Quality Estimation Shared Task

**Qi Hou, Shujian Huang*, Tianhao Ning, Xinyu Dai** and **Jiajun Chen**
National Key Laboratory for Novel Software Technology, Nanjing, China
Nanjing University, Nanjing, China
{houq, huangsj, ningth, daixy, chenjj}@nlp.nju.edu.cn

## Abstract

In this paper, we describe the submissions of the team from Nanjing University for the WMT19 sentence-level Quality Estimation (QE) shared task on English-German language pair. We develop two approaches based on a two-stage neural QE model consisting of a feature extractor and a quality estimator. More specifically, one of the proposed approaches employs the translation knowledge between the two languages from two different translation directions; while the other one employs extra monolingual knowledge from both source and target sides, obtained by pre-training deep self-attention networks. To efficiently train these two-stage models, a joint learning training method is applied. Experiments show that the ensemble model of the above two models achieves the best results on the benchmark dataset of the WMT17 sentence-level QE shared task and obtains competitive results in WMT19, ranking 3rd out of 10 submissions.

## 1 Introduction

Sentence-level Quality Estimation (QE) of Machine Translation (MT) is a task to predict the quality scores for unseen machine translation outputs at run-time, without relying on reference translations. There are some interesting applications of sentence-level QE, such as deciding whether a given translation is good enough for publishing, informing readers of the target language only whether or not they can rely on a translation, filtering out sentences that are not good enough for post-editing by professional translators, selecting the best translation among multiple MT systems and so on.

The common methods formalize the sentence-level QE as a supervised regression task. Traditional QE models (Specia et al., 2013, 2015) have two independent modules: feature extractor module and machine learning module. The feature extractor module is used to extract human-crafted features, which describe the translation quality, such as source fluency indicators, translation complexity indicators, and adequacy indicators. And the machine learning module serves for predicting how much effort is needed to post-edit translations to acceptable results as measured by the Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006) based on extracted features above.

With the great success of deep neural networks in a number of tasks in natural language processing (NLP), some researches have begun to apply neural networks to QE task and these neural approaches have shown promising results. Shah et al. (2015, 2016) combine neural features, such as word embedding features and neural network language model (NNLM) features with other features produced by QuEst++ (Specia et al., 2015). Kim and Lee (2016); Kim et al. (2017a,b) apply modified recurrent neural network (RNN) based neural machine translation (NMT) model (Bahdanau et al., 2014) to the sentence-level QE task, which does not require manual effort for finding the best relevant features. Wang et al. (2018) replace the above NMT model with modified self-attention mechanism based transformer model (Vaswani et al., 2017). This approach achieves the best result we know so far in the WMT17 sentence-level QE task on English-German language pair.

In this paper, we present two different approaches for the sentence-level QE task, which employ bi-directional translation knowledge and large-scale monolingual knowledge to the QE task, respectively. Also, a simple ensemble of them can help to achieve better quality estimation performance in the sentence-level QE task. The remainder of this paper is organized as follows. In Section 2 and Section 3, we separately describe

* Corresponding author.



*Proceedings of the Fourth Conference on Machine Translation (WMT)*, Volume 3: Shared Task Papers (Day 2) pages 97–102

the two proposed QE models above. In Section 4, we report experimental results and conclude our paper in Section 5.

## 2 Employing Bi-directional Translation Knowledge

Sennrich et al. (2015) apply the idea of back-translation to improve the performance of NMT model by extending the parallel corpus with monolingual data. Kozlova et al. (2016) propose two types of features including pseudo-references features for source sentence and back-translations features for machine translation to enrich the baseline features in sentence-level QE task. Inspired by these successful practices, we present a Bi-directional QE model, as depicted in Figure 1.

### 2.1 Model Architecture

The Bi-directional QE model contains a neural feature extractor and a neural quality estimator. The feature extractor relies on two symmetric word predictors to extract quality estimation feature vectors (QEFVs) of the source sentence and target sentence (i.e., machine translation output). The quality estimator is based on two identical Bi-directional RNN (BiRNN) (Schuster and Paliwal, 1997) for predicting quality scores using QEFVs as inputs.

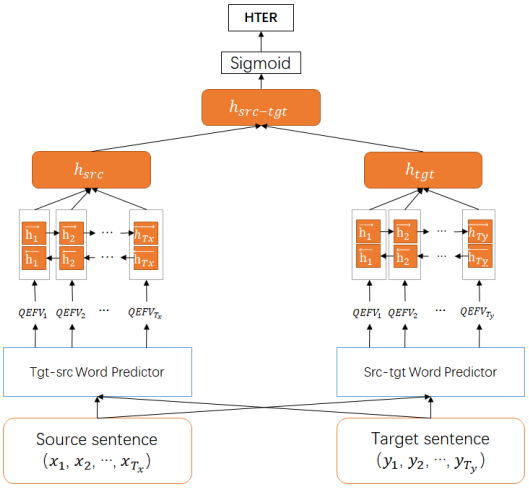


Figure 1: An illustration of the architecture of the proposed Bi-directional QE model.

The source-to-target word predictor modifies self-attention mechanism based transformer model (Vaswani et al., 2017) to i) apply additional backward decoder for the target sentence with the right to left masked self-attention and ii) generate QEFVs for target words as outputs, which is similar with QEBrain model as described in Wang et al. (2018). It is a conditional probabilistic model that generates a target word $y$ at $j$-th position via the source context $\mathbf{x} = (x_1, ..., x_{T_x})$ and target context $y_{-j} = (y_1, ..., y_{j-1}, y_{j+1}, ..., y_{T_y})$ as follows:

$$P(y_j \mid y_{-j}, \mathbf{x}; \theta) = \text{softmax}([\overrightarrow{s_j}; \overleftarrow{s_j}]) = \frac{\exp(w_j^{\mathrm{T}} W s_j)}{\sum_{k=1}^{K_y} \exp(w_k^{\mathrm{T}} W s_j)} \quad (1)$$

where $T_x$ and $T_y$ are the length of the source and target sentences. $s_j = [\overrightarrow{s_j}; \overleftarrow{s_j}]$ is the concatenation of $\overrightarrow{s_j}$ and $\overleftarrow{s_j}$, $\overrightarrow{s_j}$ is the hidden state at the last layer of forward decoder and $\overleftarrow{s_j}$ is the hidden state of backward decoder. $w_j \in \mathbb{R}^{K_y}$ is the one-hot representation of the target word, and $K_y$ is the vocabulary size of the target language. $W \in \mathbb{R}^{K_y \times 2d}$ is the weight matrix, and $d$ is the size of a unidirectional hidden layer.

To describe how well a target word $y_j$ in a target sentence is translated from a source sentence, the $\text{QEFV}_j$ is defined as follows:

$$\text{QEFV}_j = [(w_j^{\mathrm{T}} W) \odot s_j^{\mathrm{T}}]^{\mathrm{T}} \quad (2)$$

where $\odot$ is an element-wise multiplication.

Similarly, the target-to-source word predictor encodes a target sentence as input and decodes every word for source sentence step by step. We use the identical modified transformer model to generate $\text{QEFV}_i$ for every source word $x_i$ as output.

The quality estimator firstly uses the Bi-directional Long Short-term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997) model to encode given QEFVs of the source and target sentences such that

$$\overrightarrow{h_{1:T_x}}, \overleftarrow{h_{1:T_x}} = \text{BiLSTM}(\{\text{QEFV}_i\}_{i=1}^{i=T_x}) \quad (3)$$

$$\overrightarrow{h_{1:T_y}}, \overleftarrow{h_{1:T_y}} = \text{BiLSTM}(\{\text{QEFV}_j\}_{j=1}^{j=T_y}) \quad (4)$$

Secondly, the quality estimator compresses the concatenation of two sequential hidden states along the depth direction to a single one by averaging them respectively as follows:

$$h_{src} = \frac{1}{T_x} \sum_{i=1}^{i=T_x} ([\overrightarrow{h_i}; \overleftarrow{h_i}]) \quad (5)$$

$$h_{tgt} = \frac{1}{T_y} \sum_{j=1}^{j=T_y} ([\overrightarrow{h_j}; \overleftarrow{h_j}]) \quad (6)$$

Finally, sentence-level quality score of a translation sentence is calculated as follows:

$$\mathrm{QE}_{\mathrm{sentence}}(y, x) = \sigma(\mathbf{v}^{\mathrm{T}}[h_{src}; h_{tgt}]) \quad (7)$$

where $\mathbf{v}$ is a vector, $\sigma$ denotes the logistic sigmoid function.

In general, the word predictors in both directions can supervise each other and jointly complete the goal of feature extractor, which enhances the representation ability of the whole QE model. At the same time, bi-directional translation knowledge is transferred from feature extractor to quality estimator, which can be deemed to data augmentation of the original parallel corpus. Therefore, this approach can increase the diversity of training samples and improve the robustness of QE model.

### 2.2 Model Training

The training objective of Bi-directional QE model is to minimize the Mean Average Error (MAE) between the gold standard labels and predicted quality scores over the QE training samples. Because the training set for QE task is not sufficient for training the entire QE model, we need to use large-scale parallel corpus in source-to-target direction and reverse (target-to-source) direction to pre-train two word predictors respectively. Then, the parameters of the whole Bi-directional QE model are trained jointly with the training samples of sentence-level QE task.

## 3 Employing Monolingual Knowledge

In fact, most language pairs do not have a large amount of parallel corpus to train the modified NMT model. But finding monolingual data for any language is relatively easy. Therefore, we propose a QE model to integrate monolingual knowledge, as depicted in Figure 2.

### 3.1 Model Architecture

The BERT-based QE model also consists of a neural feature extractor and a neural quality estimator. The feature extractor is implemented by a pre-training representation learning model for language understanding called Multilingual-BERT (Devlin et al., 2018), which extracts hidden states corresponding to the last attention block as QEFVs for the sentence pair of source sentence and target sentence. Further, we can use a self-attention based transformer model (Vaswani et al., 2017) to translate the source sentence to pseudo-reference, which is the same language as the target sentence. Then, the input of feature extractor is replaced with the sentence pair of pseudo-reference and target sentence.

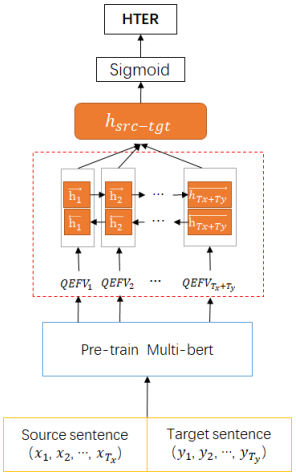


Figure 2: An illustration of the architecture of the proposed BERT-based QE model.

The quality estimator applies BiLSTM based model to predict quality scores using QEFVs as inputs such that

$$\overrightarrow{h_{1:T_x+T_y}}, \overleftarrow{h_{1:T_x+T_y}} = \mathrm{BiLSTM}(\{\mathrm{QEFV}_i\}_{i=1}^{i=T_x+T_y}) \quad (8)$$

$$h_{src-tgt} = \frac{1}{T_x + T_y} \sum_{i=1}^{i=T_x+T_y} ([\overrightarrow{h_i}; \overleftarrow{h_i}]) \quad (9)$$

$$\mathrm{QE}_{\mathrm{sentence}}(x, y) = \sigma(\mathbf{v}_1^{\mathrm{T}} h_{src-tgt}) \quad (10)$$

where $\mathbf{v}_1$ is a vector.

### 3.2 Model Training

Consistently, the pre-trained feature extractor and initialized quality estimator of BERT-based QE model are trained jointly over the training samples of sentence-level QE task by minimizing the MAE loss function.

## 4 Experiments

### 4.1 Dataset and Metrics

The bilingual parallel corpus that we used for training word predictors is officially released by the WMT17 Shared Task: Machine Translation of News[1], including Europarl v7, Common Crawl corpus, News Commentary v12, and Rapid corpus of EU press releases. The newstest2016 was used

[1]http://www.statmt.org/wmt17/translation-task.html

| | Train | Dev | Test 2017 |
|---|---|---|---|
| Sentences | 23,000 | 1,000 | 2,000 |

Table 1: Statistics of the en-de dataset of the WMT17 sentence-level QE task.

| | Train | Dev | Test 2019 |
|---|---|---|---|
| Sentences | 13,442 | 1,000 | 1,023 |

Table 2: Statistics of the en-de dataset of the WMT19 sentence-level QE task.

as development dataset. Pre-processing script can be found at github[2].

To test the performance of the proposed QE models, we conducted experiments on the WMT17 and WMT19 sentence-level QE task for English-to-German (en-de) direction. Because the gold standard labels of testing data on the WMT18 sentence-level QE task are unobtainable. The statistics of the dataset are shown in Tables 1 and 2.

Pearson's correlation coefficient (Pearson) (as primary metric), Mean Average Error (MAE) and Root Mean Squared Error (RMSE) are used to evaluate the correlation between the predicted quality scores and the true HTER scores.

### 4.2 Experimental Setting

Both of the word predictors of Bi-directional QE Model hold the same parameters. The number of layers for the self-attention encoder and forward/backward self-attention decoder are all set as 6, where we use 8-head self-attention in practice. The dimensionality of word embedding and self-attention layers are all 512 except the feed-forward sub-layer is 2048. The dropout rate is set as 0.1. Worth mentioning, the normal transformer model introduced in BERT-based QE model is trained using the same parallel corpus and parameter settings as word predictors.

For quality estimator module, the number of hidden units for forward and backward LSTM is 512. And we uniformly use a minibatch stochastic gradient descent (SGD) algorithm together with Adam (Kingma and Ba, 2014) to train all models described.

These proposed models were compared with the traditional QE framework QuEst++ (Specia et al., 2015), the neural network features based QE model SHEF/QUEST-EMB (Shah et al., 2016) and the QE model combined with NMT model, including POSTECH (Kim et al., 2017b), QE-Brain (Wang et al., 2018), and UNQE (Li et al., 2018).

[2]https://github.com/zhaocq-nlp/MT-data-processing

### 4.3 Experimental Results

In this section, we will report the experimental results of our approaches for WMT17 and WMT19 sentence-level QE task in English-German direction. For WMT17 QE task, we tried to verify our proposed models and chose the best two models to participate in WMT19 QE task. In Table 3 and Table 4, results of WMT17 and WMT19 QE tasks are listed respectively.

| Method | test 2017 en-de | | |
|---|---|---|---|
| | **Pearson** ↑ | MAE ↓ | RMSE ↓ |
| Baseline | 0.397 | 0.136 | 0.175 |
| SHEF/QUEST-EMB | 0.496 | 0.126 | 0.166 |
| POSTECH Single | 0.6599 | 0.1057 | 0.1450 |
| QEBrain Single | 0.6837 | 0.1001 | 0.1441 |
| UNQE Single | 0.700 | - | - |
| Bi-directional QE | **0.7097** | 0.1028 | 0.1352 |
| BERT-based QE | 0.6827 | 0.1081 | 0.1456 |
| +NMT | **0.703** | 0.1007 | 0.1377 |
| POSTECH Ensemble | 0.6954 | 0.1019 | 0.1371 |
| QEBrain Ensemble | 0.7159 | 0.0965 | 0.1384 |
| UNQE Ensemble | 0.710 | - | - |
| Ours Ensemble | **0.7337** | 0.0964 | 0.1294 |

Table 3: Results of the models on the WMT17 sentence-level QE. "BERT-based QE model" represents the original model with the sentence pair of source sentence and target sentence as inputs. "+NMT" represents that we use the sentence pair of pseudo-reference and target sentence as inputs of BERT-based QE model. And the rest of these two models remain the same.

| Method | test 2019 en-de | |
|---|---|---|
| | **Pearson** ↑ | Rank |
| Baseline | 0.4001 | |
| Bi-directional QE | 0.5412 | 4 |
| Ours Ensemble | 0.5433 | 3 |

Table 4: Results of submitted models on the WMT19 sentence-level QE.

From the results listed in Table 3, our proposed single models, Bi-directional QE and BERT-based QE (+NMT) can outperform all the other compared single models for the primary metric. Then,

we ensemble the two best single models above, where corresponding weights are tuned according to Pearson's correlation coefficient on the development dataset. The ensemble model can be comparable or better than the state-of-the-art (SOTA) ensemble models of WMT17 sentence-level QE task.

Considering the experimental results obtained from WMT17 QE task, we submitted the ensemble model and Bi-directional QE model to WMT19 sentence-level QE task, and ranked 3rd and 4th respectively according to WMT19 QE website.

## 5 Conclusion

This paper introduces two proposed QE models, Bi-directional QE model and BERT-based QE model, for the WMT19 sentence-level Quality Estimation shared task on English-German language pair. They can be used selectively in situations where parallel corpus and/or monolingual corpus are available. Experimental results showed that our ensemble model outperformed the SOTA results on WMT17 sentence-level QE task in English-German direction and ranked 3rd in WMT19 QE task. In future work, we would like to explore how to apply our approaches for finer-grained QE task, such as phrase-level and word-level.

## References


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

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*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017a. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-resource Language Information Processing*, 17(1):3.

Hyun Kim and Jong-Hyeok Lee. 2016. Recurrent neural network based translation quality estimation. In *Proceedings of the 1st Conference on Machine Translation*, pages 787–792.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017b. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 562–568.

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

Anna Kozlova, Mariya Shmatova, and Anton Frolov. 2016. Ysda participation in the wmt16 quality estimation shared task. In *Proceedings of the 1st Conference on Machine Translation*, pages 793–799.

Maoxi Li, Qingyu Xiang, Zhiming Chen, and Mingwen Wang. 2018. A unified neural network for quality estimation of machine translation. *IEICE Transactions on Information and Systems*, E101.D(9):2417–2421.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Kashif Shah, Fethi Bougares, Loïc Barrault, and Lucia Specia. 2016. Shef-lium-nn: Sentence level quality estimation with neural network features. In *Proceedings of the 1st Conference on Machine Translation*, volume 2, pages 838–842.

Kashif Shah, Raymond WM Ng, Fethi Bougares, and Lucia Specia. 2015. Investigating continuous space language models for machine translation quality estimation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1073–1078.

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, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 115–120.

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.

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.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for wmt18 quality estimation task. In *Proceedings of the 3rd Conference on Machine Translation*, pages 809–815.