QE BERT: Bilingual BERT using Multi-task Learning for Neural Quality Estimation

Hyun Kim and **Joon-Ho Lim** and **Hyun-Ki Kim**
SW & Contents Research Laboratory,
Electronics and Telecommunications Research Institute (ETRI), Republic of Korea
{h.kim, joonho.lim, hkk}@etri.re.kr

Seung-Hoon Na
Computer Science and Engineering,
Chonbuk National University, Republic of Korea
nash@jbnu.ac.kr

Abstract

For translation quality estimation at word and sentence levels, this paper presents a novel approach based on BERT that recently has achieved impressive results on various natural language processing tasks. Our proposed model is re-purposed BERT for the translation quality estimation and uses *multi-task learning* for the sentence-level task and word-level subtasks (i.e., source word, target word, and target gap). Experimental results on Quality Estimation shared task of WMT19 show that our systems show competitive results and provide significant improvements over the baseline.

1 Introduction

Translation quality estimation (QE) has become an important research topic in the field of machine translation (MT), which is used to estimate quality scores and categories for a machine-translated sentence without reference translations at various levels (Specia et al., 2013).

Recent Predictor-Estimator architecture-based approaches (Kim and Lee, 2016a,b; Kim et al., 2017a,b, 2019; LI et al., 2018; Wang et al., 2018) have significantly improved QE performance. The Predictor-Estimator (Kim and Lee, 2016a,b; Kim et al., 2017a,b, 2019) is based on a modified neural encoder architecture that consists of two subsequent neural models: 1) a word prediction model, which predicts each target word given the source sentence and the left and right context of the target word, and 2) a quality estimation model, which estimates sentence-level scores and word-level labels from features produced by the predictor. The word prediction model is trained from additional large-scale parallel data and the quality estimation model is trained from small-scale QE data.

Recently, BERT (Devlin et al., 2018) has led to impressive improvements on various natural language processing tasks. BERT is a bidirectionally trained language model from large-scale "monolingual" data to learn the "monolingual" context of a word based on all of its surroundings (left and right of the word).

Both BERT that is based on the Transformer architecture (Vaswani et al., 2017) and the word prediction model in the Predictor-Estimator that is based on the attention-based recurrent neural network (RNN) encoder-decoder architecture (Bahdanau et al., 2015; Cho et al., 2014) have some common ground utilizing generative pretraining of sentence encoder.

In this paper, we propose a "bilingual" BERT using multi-task learning for translation quality estimation (called the QE BERT). We describe how we have applied BERT (Devlin et al., 2018) to the QE task to make much improvements. In addition, for recent QE task, which consists of one sentence-level subtask to predict HTER scores and three word-level subtasks to detect errors for each source word, target (mt) word, and target (mt) gap, we also have applied multi-task learning (Kim et al., 2019, 2017b) to enhance the training data from other QE subtasks[1]. The results of experiments conducted on the WMT19 QE datasets show that our proposed QE BERT using multi-task learning provides significant improvements over the baseline system.

2 QE BERT

In this section, we describe two training steps for QE BERT: pre-training and fine-tuning. Figure 1 shows QE BERT architecture to predict HTER scores in sentence-level subtask and to detect errors in word-level source word, mt word, and mt gap subtasks. The sentences are tokenized using

[1] Kim et al. (2019, 2017b) use multi-task learning to take into account the training data of other QE subtasks as alternative route of handling the insufficiency of target training data.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 87–91
Florence, Italy, August 1-2, 2019.

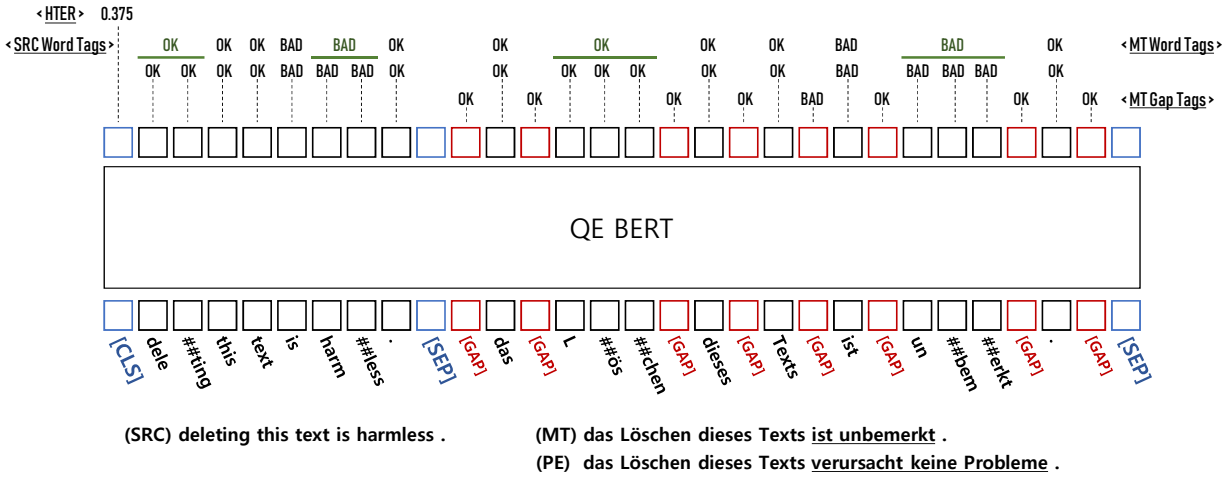

Figure 1: QE BERT architecture.

WordPiece tokenization.

2.1 Pre-training

The original BERT (Devlin et al., 2018) is focused on "monolingual" natural language understanding using generative pretraining of sentence encoder. QE BERT, which is focused on "bilingual" natural language understanding[2], is pre-trained from parallel data to learn the bilingual context of a word based on all of its left and right surroundings.

In pre-training, a default [SEP] token is used to separate source sentence and target sentence of parallel data. In addition, [GAP] tokens, which are newly introduced in this paper for word-level target gap, are inserted between target words.

As a pre-training task of QE BERT, only the masked LM task between parallel sentences is conducted where 15% of the words are replaced with a [MASK] token and then original values of the masked words are predicted[3]. The pre-training enables to make a large-scale parallel data helpful to QE task. As an initial checkpoint of pre-training, we used the released multilingual model[4].

2.2 Fine-tuning

QE BERT is fine-tuned from QE data with the above pre-trained model for a target-specific QE task.

Similar to the pre-training step, a [SEP] token is used to separate source sentence and machine translation sentence of QE data. [GAP] tokens are inserted between words of the machine translation sentence.

2.2.1 Word-level QE

To compute a word-level QE, the final hidden state (h_t) corresponds to each token embedding is used as follows:

$$P = \text{softmax}(W_\bullet\, h_t) \quad (1)$$

where P is the label probabilities and W is the weight matrix used for word-level fine-tuning. Because word-level QE task consists of source word, mt word, and mt gap subtasks, three different types of weight matrix are used for each task: $W_{src.word}$, $W_{mt.word}$, and $W_{mt.gap}$.

Because each word of sentences could be tokenized to several tokens, we primarily compute the token-level labels as follows:

$$\text{QE}_{\text{token}} = \begin{cases} \text{OK} \;, & \text{if } \text{argmax}(P) = 1 \\ \text{BAD}, & \text{if } \text{argmax}(P) = 0. \end{cases} \quad (2)$$

And then, we compute word-level labels from the token-level labels. In training, if a word is labeled as 'BAD', all of tokens in the word boundary have 'BAD' labels. In inference, if any token in the word boundary is labeled as 'BAD', the output of the word-level QE has a 'BAD' label.

2.2.2 Sentence-level QE

To compute a sentence-level QE, the final hidden state (h_s) corresponds to the [CLS] token embed-

[2]In Lample and Conneau (2019), translation language model (TLM) pretraining is used for cross-lingual understanding by concatenating parallel sentences.

[3]In Devlin et al. (2018), two pre-training tasks – masked LM and next sentence prediction – are conducted simultaneously.

[4]"BERT-Base Multilingual Cased" model, released in https://github.com/google-research/bert.

ding, which is a fixed-dimensional pooled representation of the input sequence, is used as follows:

$$\text{QE}_{\text{sent}} = \text{sigmoid}(W_s\ h_s) \quad (3)$$

where W_s is the weight matrix used for sentence-level fine-tuning.

2.2.3 Multi-task learning

The QE subtasks at word and sentence levels are highly related because their quality annotations are commonly based on the HTER measure. Quality annotated data of other QE subtasks could be helpful in training a QE model specific to a target QE task (Kim et al., 2019). To take into account the training data of other QE subtasks as a route of supplementation of target training data, we apply multi-task learning (Kim et al., 2019, 2017b).

For multi-task learning of word-level QE, we use a linear summation of word-level objective losses as follows:

$\mathcal{L}_{WORD} = \mathcal{L}_{src.word} + \mathcal{L}_{mt.word} + \mathcal{L}_{mt.gap}$

where most QE BERT components are common across word-level source word, mt word, and mt gap subtasks except for the output matrices $W_{src.word}$, $W_{mt.word}$, and $W_{mt.gap}$.

Kim et al. (2019) showed that it is helpful to use word-level training examples for training a sentence-level QE model. For multi-task learning of sentence-level QE, we combine sentence-level objective loss and word-level objective losses by simply performing a linear summation of the losses for each task as follows:

$\mathcal{L}_{SENT} = \mathcal{L}_{hter} + \mathcal{L}_{src.word} + \mathcal{L}_{mt.word} + \mathcal{L}_{mt.gap}$

where most QE BERT components are common across sentence-level and word-level tasks except for the output matrices of each task.

3 Experimentation

3.1 Experimental settings

The proposed learning methods were evaluated on the WMT19 QE Shared Task[5] of word-level and sentence-level English-Russian and English-German.

We used parallel data provided for the WMT19 news machine translation task[6] to pre-train QE BERT. The English-Russian parallel data set consisted of the ParaCrawl corpus, Common Crawl corpus, News Commentary corpus, and Yandex Corpus. The English-German parallel data set consisted of the Europarl corpus, ParaCrawl corpus, Common Crawl corpus, News Commentary corpus, and Document-split Rapid corpus.

In pre-training, we used the default hyperparameter setting of the released multilingual model. In fine-turing, a sequence length of 512 was used to cover the length of QE data.

To make ensembles, we combined five instances having different hyperparameter weight for 'BAD' label (i.e., 1:10, 1:15, 1:20, 1:25, and 1:30). For word-level ensemble results, we voted the predicted labels from each instance. For sentence-level ensemble results, we averaged the predicted HTER scores from each instance.

3.2 Comparison of learning methods

Tables 1 and 2 show the experimental results obtained from the QE BERT using the different learning methods for the WMT19 word-level and sentence-level QE tasks. For both language pairs, using multi-task learning consistently improves the scores.

We made ensembles by combining five instances of QE BERT models. The word-level results of ensemble A are based on mixtures of the best performance systems on each subtasks (i.e., source word, mt word, and mt gap tasks). On the other hand, the word-level results of ensemble B are based on an all-in-one system using a unified criterion[7] with same model parameters for all word-level subtasks.

Finally, Tables 3 and 4 show the results obtained in the WMT19 test set for our submitted systems and official baseline systems.

4 Conclusion

In this paper, we explored an adaptation of BERT for translation quality estimation. Because the quality estimation task consists of one sentence-level subtask to predict HTER scores and three word-level subtasks to detect errors for each source word, target word, and target gap, we also applied multi-task learning to enhance the training data from other subtasks. The results of experiments conducted on WMT19 quality estimation datasets strongly confirmed that our proposed bilingual BERT using multi-task learning

[5] http://www.statmt.org/wmt19/qe-task.html

[6] http://www.statmt.org/wmt19/translation-task.html

[7] The averaged performance on source word, mt word, and mt gap tasks is used as the unified criterion to select model parameters of the all-in-one system.

Word level	Source Word (F_1-**Mult** ↑	F_1-BAD ↑	F_1-OK ↑)	MT (All) (F_1-**Mult** ↑	F_1-BAD ↑	F_1-OK ↑)
<**English-Russian**>						
QE-BERT Word	0.3344	0.3663	0.9128	0.3895	0.4051	0.9617
QE-BERT Multitask-Word	0.3513	0.3780	0.9294	0.3943	0.4076	0.9673
QE-BERT Multitask-Word Ensemble A*	0.3600	0.3861	0.9326	0.4128	0.4275	0.9657
QE-BERT Multitask-Word Ensemble B*	0.3452	0.3700	0.9331	0.3934	0.4071	0.9665
<**English-German**>						
QE-BERT Word	0.3755	0.4113	0.9130	0.4028	0.4198	0.9595
QE-BERT Multitask-Word	0.3918	0.4288	0.9138	0.4074	0.4258	0.9567
QE-BERT Multitask-Word Ensemble A*	0.4044	0.4391	0.9210	0.4318	0.4501	0.9593
QE-BERT Multitask-Word Ensemble B*	0.3916	0.4262	0.9189	0.4288	0.4466	0.9602

Word level	MT Word (F_1-**Mult** ↑	F_1-BAD ↑	F_1-OK ↑)	MT Gap (F_1-**Mult** ↑	F_1-BAD ↑	F_1-OK ↑)
<**English-Russian**>						
QE-BERT Word	0.4215	0.4561	0.9240	0.1609	0.1631	0.9863
QE-BERT Multitask-Word	0.4313	0.4616	0.9344	0.1734	0.1758	0.9866
QE-BERT Multitask-Word Ensemble A*	0.4354	0.4642	0.9381	0.1791	0.1812	0.9884
QE-BERT Multitask-Word Ensemble B*	0.4180	0.4446	0.9403	0.1710	0.1730	0.9882
<**English-German**>						
QE-BERT Word	0.4307	0.4640	0.9283	0.2729	0.2765	0.9871
QE-BERT Multitask-Word	0.4365	0.4724	0.9241	0.2936	0.2983	0.9840
QE-BERT Multitask-Word Ensemble A*	0.4429	0.4766	0.9293	0.3060	0.3107	0.9849
QE-BERT Multitask-Word Ensemble B*	0.4443	0.4767	0.9320	0.2884	0.2930	0.9845

* Our submissions at the WMT19 QE task

Table 1: Results of the QE BERT model on the *development* set of the WMT19 *word-level* QE task.

Sentence level	**Pearson's** r ↑	Spearman's ρ ↑	MAE ↓	RMSE ↓
<**English-Russian**>				
QE-BERT Sent	0.4683	0.4524	0.1151	0.2072
QE-BERT Multitask-Sent-Word	0.4948	0.4908	0.1106	0.2056
QE-BERT Multitask-Sent-Word Ensemble*	0.5229	0.5102	0.1080	0.2016
<**English-German**>				
QE-BERT Sent	0.4849	0.5401	0.1072	0.1698
QE-BERT Multitask-Sent-Word	0.5199	0.5859	0.1026	0.1670
QE-BERT Multitask-Sent-Word Ensemble*	0.5450	0.6229	0.0978	0.1665

* Our submissions at the WMT19 QE task

Table 2: Results of the QE BERT model on the *development* set of the WMT19 *sentence-level* QE task.

Word level	**Source Word** F_1-**Mult** ↑	**MT (All)** F_1-**Mult** ↑
<**English-Russian**>		
Baseline	0.2647	0.2412
QE-BERT Multitask-Word Ensemble A*	0.4202	0.4515
QE-BERT Multitask-Word Ensemble B*	0.4114	0.4300
<**English-German**>		
Baseline	0.2908	0.2974
QE-BERT Multitask-Word Ensemble A*	0.3946	0.4061
QE-BERT Multitask-Word Ensemble B*	0.3960	0.4047

* Our submissions at the WMT19 QE task

Table 3: Results of the QE BERT model on the *test* set of the WMT19 *word-level* QE task.

Sentence level	**Pearson's** r ↑	Spearman's ρ ↑
<**English-Russian**>		
Baseline	0.2601	0.2339
QE-BERT Multitask-Sent-Word Ensemble*	0.5327	0.5222
<**English-German**>		
Baseline	0.4001	0.4607
QE-BERT Multitask-Sent-Word Ensemble*	0.5260	0.5745

* Our submissions at the WMT19 QE task

Table 4: Results of the QE BERT model on the *test* set of the WMT19 *sentence-level* QE task.

achieved significant improvements. Given this promising approach, we believe that BERT-based quality estimation models can be further advanced with more investigation.

Acknowledgments

This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1):3:1–3:22.

Hyun Kim and Jong-Hyeok Lee. 2016a. Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation*, pages 787–792, Berlin, Germany. Association for Computational Linguistics.

Hyun Kim and Jong-Hyeok Lee. 2016b. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498, San Diego, California. Association for Computational Linguistics.

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 Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2019. Multi-task stack propagation for neural quality estimation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4):48:1–48:18.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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 Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 822–828, Belgium, Brussels. Association for Computational Linguistics.