BJTU's Submission to CCMT 2020 Quality Estimation Task

Hui Huang, Jin'an Xu(✉), Wenjing Zhu, Yufeng Chen, and Rui Dang

School of Computer and Information Technology, Beijing Jiaotong University,
Beijing, China
{18112023,jaxu,18120461,chenyf,19125167}@bjtu.edu.cn

Abstract. This paper presents the systems developed by Beijing Jiaotong University for the CCMT 2020 quality estimation task. In this paper, we propose an effective method to utilize pretrained language models to improve the performance of QE. Our model combines three popular pretrained models, which are Bert, XLM and XLM-R, to create a very strong baseline for both sentence-level and word-level QE. We tried different strategies, including further pre-training for bilingual input, multi-task learning for multi-granularities and weighted loss for unbalanced word labels. To generate more accurate prediction, we performed model ensemble for both granularities. Experiment results show high accuracy on both directions, and outperform the winning system of last year on sentence level, demonstrating the effectiveness of our proposed method.

Keywords: Machine Translation · Quality Estimation · Pretrained language model

1 Introduction

Machine translation quality estimation (Quality Estimation, QE) aims to evaluate the quality of machine translation automatically without golden reference. QE can be implemented on different granularities, thus to give an estimation for different aspects of machines translation output.

This paper introduces in detail the submission of Beijing Jiaotong University to the quality estimation task in the 16th China Conference on Machine Translation (CCMT2020). This year, the QE task consists of two language directions of Chinese-English and English-Chinese, and two granularities of word-level and sentence-level subtasks, thus four subtasks in total.

We propose an effective method to utilize pretrained language models to improve the performance of QE. Our model combines three popular pretrained models, which are Bert [1], XLM [2] and XLM-R [3], to create a very strong baseline for both sentence-level and word-level QE.

We also tried different strategies to boost the final results, including further pre-training for bilingual input, multi-task learning for multi-granularities and weighted loss for unbalanced word labels. To improve the final accuracy, we ensembled the results generated by different models for both sentence and word level.

J. Li and A. Way (Eds.): CCMT 2020, CCIS 1328, pp. 105–113, 2020.
https://doi.org/10.1007/978-981-33-6162-1_10

Experiment results show that our model achieves high accuracy on both directions, surpassing previous models on sentence-level, and obtaining competitive performance on word-level, demonstrating the effectiveness of our proposed method.

2 Model Description

2.1 Pretrained Models for Quality Estimation

Our method is based on three recent proposed pretrained models, Bert, XLM and XLM-R. All of these three models are based on multi-layer Transformer architecture with different training procedures.

For both word-level and sentence-level QE task, we concatenate source sentences and machine translated sentences following the way pre-trained models treat sentence pairs, and do prediction on the top of them.

For sentence-level prediction, we tried two different strategies. The first one is to directly use the first token according to the special token [CLS] to perform prediction, as we believe this logit integrates sentence-level information. The second one is to add another layer of RNN on the top of pre-trained models, to better leverage the global context information, as shown in Fig. 1.

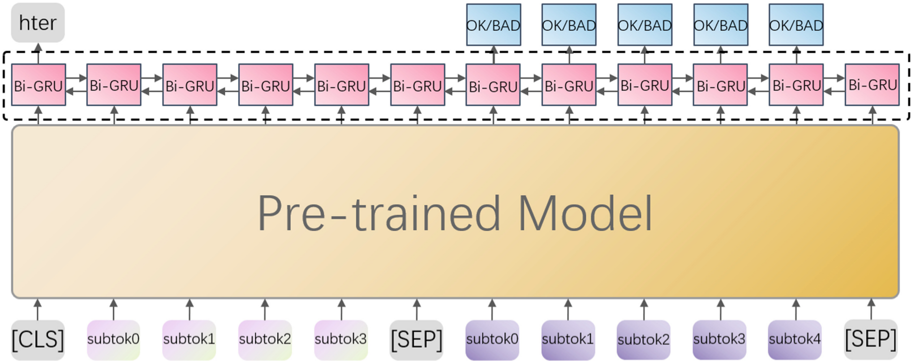

Fig. 1. Pre-trained model for QE with multi-task learning. The component circled with dashed line is alternative.

For word-level prediction, we use each logit according to each token in the sentence to generate word-quality label.

The loss functions for word and sentence-level are as follows:

$$L_{word} = \sum_{s \in D} \sum_{x \in s} -(p_{OK} \log p_{OK} + p_{BAD} \log p_{BAD}) \tag{1}$$

$$L_{sent} = \sum_{s \in D} \| sigmoid(W_s h(s)) - hter_s \| \tag{2}$$

where s and x denote each sentence and word in the dataset, p_{OK} and p_{BAD} denote the probability for each word to be classified as OK/BAD, $h(s)$ denotes the hidden representation for each sentence, and W_s denote the transformation matrices for sentence and word level prediction, and $hter_s$ denote the HTER[1] measure for each sentence.

2.2 Further Pretraining for Bilingual Input

Despite the shared multilingual vocabulary, Bert is originally a monolingual model, treating the input as either being from one language or another. To help Bert adapts to sentence pairs from different languages, we implement a further pretraining step, training Bert model with massive parallel machine translation data [4].

For our task of QE, we combine bilingual sentence pairs from large-scale parallel dataset, and randomly mask sub-word units with a special token, and then train Bert model to predict masked tokens. Since our input are two parallel sentences, during the predicting of masked words given its context and translation reference, Bert can capture the lexical alignment and semantic relevance between two languages.

After this further pretraining step, Bert model is familiar with bilingual inputs, and acquires the ability to capture translation errors between different languages. This method is similar to the pretraining strategy mask-language-model in [1], while its original implementation is based on only sentences from monolingual data.

In contrast, XLM and XLM-R are multilingual models which receive two sentences from different languages as input, which means further pretraining is likely to be redundant. This is verified by our experiment results demonstrated in the following section.

2.3 Multi-task Learning for Multi-granularities

The QE subtasks at sentence and word-level are highly related because their quality annotations are commonly based on the HTER measure. Quality annotated data of other subtasks could be helpful in training a QE model specific to a target task [5].

We also implemented multi-task learning on our pretrained models. Since the linear transformation for predictions according to different granularities are implemented on different positions, we can perform multi-task training and inference naturally without any structure adjustment. Since we tried two different models, with or without bidirectional RNN, our model can be illustrated as the following figure:

During training, predictions for different granularities are generated at the same time on different positions, and losses are combined and back-propagated simultaneously. The loss function is as follows:

$$L_{join} = \sum_{s \in D} \sum_{x \in s} cross\_entropy(W_w h(x), y_x) + \|sigmoid(W_s h(s)) - hter_s\| \quad (3)$$

where $h(x)$ and $h(s)$ denote the hidden representations for each word and sentence, and W_w and W_s denote the transformation matrices for sentence and word prediction.

[1] https://github.com/jhclark/tercom.

Most model components are common across sentence-level and word-level tasks except for the output matrices of each task, which is especially beneficial for sentence-level prediction, since the training objective for sentence QE only consists of one single logit containing limited information.

2.4 Weighted Loss for Unbalanced Word Labels

The quality of machine translated sentences in QE data is very high [6], which means that a huge proportion of the sentences do not need post-editing at all. This leads to an unbalanced label distribution where most of the word labels are BAD, which makes it very likely to give a skewed prediction with a very low F1 score for BAD words.

To improve the overall performance, we add up to the weight for BAD words when calculating cross-entropy loss, enabling the model emphasize more on the incorrectly translated words. The word-level loss function is as follows:

$$L_{word} = \sum_{s \in D} \sum_{x \in s} -(p_{OK} \log p_{OK} + \lambda p_{BAD} \log p_{BAD}) \tag{4}$$

where λ is a hyper-parameter larger than 1.

We also tried other data augmentation skills to balance word labels, which is demonstrated in the next section.

2.5 Multi-model Ensemble

Until now, we have built three different QE models trained with different architectures, which can capture different information from the same text. Considering the variation of different strategies and initialized parameters, we can have multiple models for each subtask, which can be integrated to generate stronger performance [7].

For word level QE, to ensemble multiple predictions for each token, we tried two different strategies. The first one is to take the average of logit generated by softmax layer for each token, and then argmax it to get OK/BAD label. The second one is to vote based on different labels generated by different models. For an instance, if there are two Oks and one BAD out of three predictions for a token, then the ensembled result for this token would be OK.

For sentence level QE, we simply take the average of predicted HTER scores from different models.

Due to time limitation, we did not explore more complex ensemble techniques illustrated in [8], which introduced conspicuous improvement in their work.

3 Experiment

3.1 Dataset

We use the QE data from CCMT2020 Machine Translation Quality Estimation tasks. CCMT QE tasks contain two different language directions (Chinese-English and English-Chinese) on both sentence-level and word-level. The amount of data provided on both language pairs and levels are very small (no more than 15 k triples on all directions), which makes QE a highly data-sparse task.

To further pretrain the Bert model, we use the parallel dataset for Chinese-English Translation task in CCMT2020, which contains nearly 7 million sentence pairs.

3.2 Experiment Results

The experiment results on both directions and granularities are shown in Table 1 and Table 2, where *transformer-dlcl* [9] and *CCNN* were the top2 systems in CCMT 2019 QE task.

For sentence-level QE, we surpass all baselines on both directions with limited computation resource. For word-level QE, we do not manage to surpass the top 1 system of last year. But we have to mention that on word-level task, we do not apply further pretraining step on both models before finetuning, so the computation overhead is very low with just a few hours fine-tuning on one single GPU.

Table 1. Experiment results on CCMT2020 sentence-level QE dev set

Language Direction	System	Pearonr	Spearman	MSE
Chinese- English	CCNN	0.56	0.49	–
	transformer-dlcl	0.6164	–	–
	Bert	0.6069	0.5182	0.5626
	XLM	0.5744	0.5467	0.5606
	XLM-R	0.5657	0.5057	0.5357
	Ensemble Model	**0.6277**	0.5701	–
English-Chinese	CCNN	0.55	0.42	–
	transformer-dlcl	0.5861	–	–
	Bert	0.5172	0.3907	0.4540
	XLM	0.5540	0.4110	0.4825
	XLM-R	0.5365	0.4001	0.4683
	Ensemble Model	**0.5907**	0.5521	–

Table 2. Experiment results on CCMT2020 word-level QE dev set

Language Direction	System	F1-Multi	F1-BAD	F1-OK
Chinese- English	transformer-dlcl	**0.5393**	0.6152	0.8767
	Bert	0.4846	0.5634	0.8602
	XLM	0.4844	0.5635	0.8597
	XLM-R	0.5061	0.5902	0.8575
	Ensemble Model	0.5141	0.5913	0.8649
English-Chinese	transformer-dlcl	**0.4385**	0.8974	0.4886
	Bert	0.3947	0.4508	0.8757
	XLM	0.4073	0.4625	0.8808
	XLM-R	0.4173	0.4669	0.8973
	Ensemble Model	0.4336	0.4841	0.8958

Besides, we do not introduce any pseudo data during the training of our QE system, while transformer-dlcl introduced pseudo data in all subtasks, which led to the improvement of 2-4 points.

In a word, the pretrained language model can be a very strong baseline for QE at both sentence-level and word-level. It requires no complicated architecture engineering and massive training data, and can provide reliable performance.

3.3 Ablation Study

In this section, we will discuss the influence of different strategies on our model. Notice although we described a lot of strategies to boost QE system in former sections, their influence on different granularities are different. Besides, due to the update of codes during the evaluation period, there may be some discrepancy between the following results and the results we released in Sect. 3.2.

Extra Bi-RNN. It is alternative to add an extra layer of bidirectional RNN before the softmax layer. Extra layer may introduce more globalized prediction, but may also introduce noise since we have to random-initialize it.

As shown in Table 3, an Extra layer of Bi-RNN does not necessarily introduce improvement. Sometimes it can and sometime it cannot. But If there is no multi-task learning when doing sentence-level QE, then an extra layer is compulsory for XLM and XLM-R, since these two models are not pretrained with sentence-level task.

Further Pre-training for Bilingual Input. As we have mentioned before, Bert is only trained with monolingual input, so it is reasonable to believe further pre-training could help Bert adapted to multilingual input. But astonishingly, we find further pre-training can only improve the sentence-level QE, and is harmful for word-level QE on Bert, as shown in Table 4, which needs our future investigation.

Table 3. Extra Bi-RNN on the top of pre-trained model

Language Direction	System	Level	Extra Bi-RNN	F1-multi
Chinese-English	XLM-R	sentence	No	0.5386
			Yes	0.5657
		word	No	0.5057
			Yes	0.4993
	XLM	sentence	No (w/o muti-task)	0.0975
			No	0.5744
			Yes	0.5666

Multi-task Learning for Multi-granularities. As shown in Table 5, after joint trained with different granularities, the results of sentence-level QE increase a lot, which verifies our conjecture that word-level labels can help the training of sentence-level QE. For word-level QE, the avail of multi-task learning seems limited.

Table 4. Further pre-training for bilingual input

Language direction	System	Level	Further pretrain	Pearsonr/F1-multi
English-Chinese	Bert	sentence	No	0.4230
			Yes	0.5169
		word	No	0.3902
			Yes	0.3837

Label Balancing for Word-level QE. We try three different strategies including up-sampling sentence-pairs with high HTER values and down-sampling sentence-pairs with low HTER values, and find that weight balancing when calculating loss is a simple yet effective strategy, as shown in Table 6.

Table 5. Multi-task learning for multi-granularities

Language direction	Level	Model	Multi-task	Pearsonr/F1-multi
English-Chinese	sentence	Bert	No	0.4893
			Yes	0.5169
	word	Bert	No	0.3962
			Yes	0.3902

Although data sampling can also help the model to emphasize more on the bad words when training, but it will damage the natural distribution of sentence-pairs, and thus harmful to final performance. We try different values for λ ranging from 5 to 20, and finally set λ as 10 in Eq. (4).

Table 6. Label balancing for word QE

Language direction	Level	Model	Balancing strategy	F1-multi
English-Chinese	word	Bert	No	0.3227
			up sampling	0.3847
			down sampling	0.3357
			weight balancing	0.3962

Word-level Multi-model Ensemble. As we have mentioned before, there are two strategies to do word-level ensemble, namely averaging logits and voting. Intuitively, averaging logits should be more effective, since more information is integrated. But experiment defies our hypothesis, as show in Table 7.

As shown in Table 7, we did not see significant outperformance of logit averaging over voting mechanism. This may be caused by the unbalanced word-label, which leading to a biased logit distribution (where most tokens are assigned with a logit close to 1). Even there is one prediction under 0.5, it would not change the result since the other predictions are likely to be almost 1 produced by the softmax layer.

Table 7. Word-level Multi-model Ensemble

Language direction	Level	Model	Balancing strategy	F1-multi
English-Chinese	word	Ensembled	voting	0.4321
			logit averaging	0.4336
Chinese-English	word	Ensembled	voting	0.5116
			logit averaging	0.5141

4 Conclusion

In this paper, we described our submission in quality estimation task, consisting of two language directions and two granularities. We implement the QE system based on three popular pretrained models, namely Bert, XLM and XLM-R, and study different applicable strategies on QE task, i.e. further pretraining on bilingual input, multi-task training on multi-granularities and weighted loss for word labels. We ensembled multiple models to generate more accurate prediction. Our model achieves strong performance on both sentence-level and word-level QE tasks with limited computation resource, and outperforms the previous SOTA models on sentence-level development set, verifying the validity of our proposed strategies.

Massive linguistic knowledge contained in pretrained models is very helpful for the QE task even when there is limited training data. In the future, we will continue our research on the application of pretrained models on different QE tasks.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Contract 61976015, 61976016, 61876198 and 61370130), and the Beijing Municipal Natural Science Foundation (Contract 4172047), and the International Science and Technology Cooperation Program of the Ministry of Science and Technology (K11F100010).

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
4. Kim, H., Lim, J.H., Kim, H.K., Na, S.H.: 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) (2019)

5. Hyun, K., Jong-Hyeok, L., Seung-Hoon, N.: 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, pp. 562–568 (2017)
6. Specia, L., Blain, F., Logacheva, V., Astudillo, R., Martins, A.F.: Findings of the wmt 2018 shared task on quality estimation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 689–709 (2018)
7. Zhi-Hua, Z., Jianxin, W., Wei, T.: Ensembling neural networks: many could be better than all. Artif. Intell. **137**(1–2), 239–263 (2002)
8. Kepler, F., et al.: Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. arXiv preprint arXiv:1907.10352 (2019)
9. Wang, Z., et al.: NiuTrans Submission for CCMT19 Quality Estimation Task. In: Huang, S., Knight, K. (eds.) CCMT 2019. CCIS, vol. 1104, pp. 82–92. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1721-1_9