Target Oriented Data Generation for Quality Estimation of Machine translation

Huanqin Wu, Muyun Yang*, Jiaqi Wang, Junguo Zhu, and Tiejun Zhao

Harbin Institute of Technology, Harbin, China
wuhuanqin@foxmail.com, yangmuyun@hit.edu.cn, 17862702130@163.com,
zhujunguohit@gmail.com, tjzhao@hit.edu.cn

Abstract. Quality estimation (QE) is a non-trivial issue for machine translation (MT) and the neural approach appears a promising solution to this task. Annotating QE training corpora is a costly process but necessary for supervised QE systems. To provide informative large scale training data for the MT quality estimation model, this paper proposes an approach to generate pseudo QE training data. By leveraging the provided labeled corpus in this task, our method generates pseudo training samples with a purpose of similar distribution of translation error of the labeled corpus. It also describes a sentence specific data expansion strategy to incrementally boost the model performance. The experiments on the different open datasets and models confirm the effectiveness of the method, and indicate that our proposed method can significantly improve the QE performance.

Keywords: Machine translation · Quality estimation · Pseudo data.

1 Introduction

Quality Estimation (QE) is a significant task in the study of machine translation (MT). It plays an important role in guiding for better the MT outputs in real application. Different from automatic MT evaluation, QE systems aim at predicting the translation quality of MT system outputs without reference translations [1]. With the popularity of free web MT services, vast users are increasingly demanding the QE system, since the quality of the MT results becoming crucial to web users.

Traditional approaches address QE task as a regression or classification problem via machine learning models, and focused on feature extraction and feature selection. Deep learning relieves the problem of manual feature engineering and there appear several QE methods based on deep learning. Various neural networks are applied for estimating the quality of machine translation output by [2–5, 7, 8]. Experimental results show that these neural based models can achieve state-of-the-art performance.

It is worth noting that, the success of deep neural networks usually relies on large scale of annotated data. But in practice, for quality estimation task

* Corresponding author

in machine translation, there are very limited amount of labeled data and it is too expensive to develop such labeled data. To address this issue, this paper introduces a novel target oriented pseudo training data generation approach to automatically generate large scale training sample for QE task. The motivation is to generate pseudo training samples according to data distribution of target limited labeled data, which is readily available in the task. To best exploit the pseudo data, a sentence specific expansion strategy is also proposed. In order to mitigate the effects of noise in pseudo training data on the QE model, we adopt the framework of two-step training, which means pre-training QE model under pseudo data and fine-tuning it using human labeled data.

We demonstrate the effectiveness of our approach on the WMT sentence-level English-to-Spanish QE task and CWMT sentence-level Chinese-to-English and English-to-Chinese QE task. Experimental results show that our proposed method significantly outperforms baseline QE models on these three QE tasks. The contributions of this paper are as follows:

- We present a method to generate large scale QE training data based a bilingual corpus and a limited human labeled QE data automatically.
- We propose a sentence specific expansion strategy to exploit pseudo data, which are very effective and important as the experiments show.
- We prove that our generated training data can be used for different QE models as an additional corpus to improve the QE performance.

The remainder of this paper is organized as follows. In Section 2, we introduce the related work of this paper. The target oriented pseudo data generation approach is described in Section 3. In Section 4, we report the experiment and results, and conclude our paper in Section 5.

2 Related Work

Quality estimation of machine translation. Quality Estimation (QE), which aims at estimating quality scores or categories for given translations from an unknown MT system without reference translation, has become of growing importance in the field of MT. Previous studies on QE are extensively based on feature engineering work, which investigates useful QE features as input for regression or classification algorithms to estimate translation quality scores or categories.

Recently, neural network methods have been applied to QE task. Kreutzer et al. [2] proposed a window-based FNN architecture for QE called QUality Estimation from scraTCH (QUETCH). Patel and M [4] proposed an RNN-based architecture for QE, they treated QE as a sequential labelling and used the bilingual context window to compose an input layer. Martins et al. [5] proposed extensions of QUETCH to the bilingual context window by using convolutional neural network model, bidirectional RNN model and convolutional RNN model. The bilingual context windows are commonly used to compose the input layer in conventional approaches. To obtain a bilingual context window, a word alignment

component was additionally required in these QE models, while word alignment component may exit error.

Differently, Kim et al. [3] proposed predictor-estimator architecture, which firstly trained a word predictor model based on RNN and used it to extract feature for QE task. Fan et al. [8] present a novel QE model based on transformer and achieved state-of-the-art performance. They introduced the neural "bilingual expert" model based on self-attention as the prior knowledge model. Then, they use a simple Bi-LSTM as the QE model with the extracted model derived and manually designed mis-matching features.

Because of these two architectures use complex architecture and requires resource-intensive pre-training. In addition, Ive et al. [6] and Zhu et al. [7] proposed light-weight neural approach, which employ only two bi-directional RNNs (bi-RNN) as encoders to learn the representation of the (source, MT) sentence pair.

Pseudo data for quality estimation. For the QE task in machine translation, available labeled training data is limited to train a neural model. To avoid this problem, bilingual data or additional MT systems is employed for training QE model in various ways.

Kim et al. [3] and Fan et al. [8] used large-scale bilingual corpus to pre-train a neural word predictor model or neural bilingual expert model. Then they use the pre-trained model to make quality vectors for training QE model by small amount of labeled data. Zhu et al. [7] also used bilingual corpus but directly for QE model, in their work, parallel bilingual sentence pairs are used as positive cases while random bilingual sentence pairs are used as negative cases, the goal is to maximize the QE score of the positive and negative cases.

The above efforts well addressed the issue of insufficient data by using bilingual corpus. However, in their works, minor mistakes in the translation process are ignored. Actually, for the QE model, minor mistakes should be paid more attention. In order to model these minor mistake, there are some methods that using additional MT systems for generating pseudo training data. Liu et al. [9] proposed the approach under the framework of maximum marginal likelihood estimation to build QE systems, they firstly used a bilingual corpus for optimizing an additional translation model, then running n-best decoding on the source side of another bilingual corpus using translation model. At last, they used the MT results as training data to get QE model. Using addition MT systems can provide lots of training data with minor mistakes. But training of MT system consume a lot of resources and time. In addition, MT systems are usually system-specific.

In this paper, instead of directly using bilingual data or generating negative data based on additional MT systems, we introduce a target oriented method to automatic generated pseudo training samples for QE model. In our method, we don't need to pre-train a neural model on larger scale bilingual data or train additional MT systems. Larger scale of effect QE training data can be obtained just by a small amount of resources according to our approach.

3 Target Oriented Data Generation and Specific Expansion

This section will describe the target oriented pseudo data generation approach for QE task. The process of our approach has two steps: target oriented data generation and sentence specific expansion. The overview of our method as shown in Fig. 1.

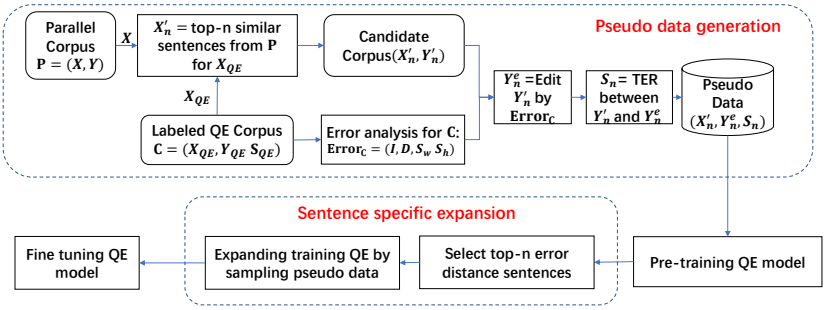

Fig. 1. Overview of our method.

3.1 Target Oriented Data Generation

Candidate corpus selection. In order to generate pseudo training samples that have similar translation error distribution with human labeled QE corpus, a large scale candidate corpus is collected firstly, it should be noted that candidate corpus is expected to be similar with the labeled QE corpus. In this paper, we select the top-n similar sentences from larger scale bilingual corpus for each sentence in labeled QE corpus. The result of similar sentences selection will be used as candidate corpus. Noted that we use TF-IDF to measure the sentence similarity between two sentences.

Specifically, given a labeled QE corpus $\{< X_{QE}, Y_{QE}, S_{QE} >_j\}_{j=1}^{M}$ and a bilingual corpus $\{< X, Y >_i\}_{i=1}^{N}$, for each source language sentence X_{QE} in $< X_{QE}, Y_{QE}, S_{QE} >$, we can get top-n similar sentences $< \{X_i^{'}, Y_i^{'}\}_{i=1}^{n} >$ from bilingual corpus by the similarity of source language sentences. Finally, we can get candidate corpus $\{< \{X_i^{'}, Y_i^{'}\}_{i=1}^{n} >_j\}_{j=1}^{M}$

Translation error distribution analysis. As shown the Fig. 1, the translation error distribution of target labeled data is pre-analyzed. In this paper, translation error distribution is defined as the minimum number of edits for human post-edition on translation, including: insertion(I), deletion(D), substitution(S_u) and shift(S_h).

In order to describe the translation error in target labeled QE data, for each labeled QE sentence pair, we define a quadruple like $< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >$ to record the type of translation error (include: $I, D, S_u$and S_h) and number of each error type(we use n_i, n_d, n_u and n_h to record the number for each error type). At last, the translation error distribution of target human labeled QE data can be defined as $\{< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >_j\}_{j=1}^{M}$.

Pseudo training samples generation. Given a human labeled QE data translation error distribution$\{< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >_j\}_{j=1}^{M}$ and a candidate data $\{< \{X_i^{'}, Y_i^{'}\}_{i=1}^{n} >_j\}_{j=1}^{M}$, pseudo translations will be obtained by editing $< \{Y_i^{'}\}_{i=1}^{n} >_j$ according to $< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >_j$.

During this process, the type of translation error(include: $I, D, S_u$and S_h) to be edited and the number of the each error type(n_i, n_d, n_u and n_h) are considered, which are deemed as the property of pseudo data. To investigate the key factors in fitting the target translation error distribution, the effects of different properties to QE model is empirically examined in sub-section 4.3 of this paper.

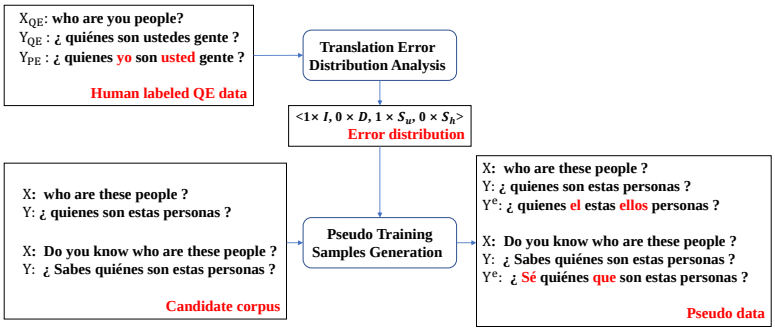

Fig. 2. Example of pseudo training sample generation.

Algorithm 1 presents the detailed procedure of generating pseudo data for QE. Specifically, when candidate sentence need to be substituted or inserted, we randomly select a word from the vocabulary to substitute or insert the original one. In addition, we randomly select a word in candidate sentence to delete when it need to be deleted. For the shift operation of chunk, we also randomly select a chunk in the sentence and shift its. Then, the generated pseudo data not only bears a similar TER score to the target human labeled QE data, but also obeys the similar distribution of translation errors.

Compared with the proposed pseudo data generation, MT seems to be another alternative at hand to generate the pseudo data for QE training. Actually, MT has been used to generate pseudo-reference translations for QE task [10] [11]. The reason we do not apply MT outputs for QE task come from two major concerns. First, MT is too "heavy", since MT (either NMT or SMT) usually requires

Algorithm 1 Target-oriented Data Generation

Input: Labeled MT error distribution $\{< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >_j\}_{j=1}^{M}$; Candidate corpus $\{< \{X_i^{'}, Y_i^{'}\}_{i=1}^{n} >_j\}_{j=1}^{M}$;

Output: Pseudo training data $\{< \{X_i^{'}, Y_i^{e}, S_i\}_{i=1}^{n} >_j\}_{j=1}^{M}$

$j = 1$
Pseudo data $P = \{\}$
while $j <= M$ **do**
  $< \{Y_i^e\}_{i=1}^{n} >_j \Leftarrow$ Editing $< \{Y_i^{'}\}_{i=1}^{n} >_j$ in $< \{X_i^{'}, Y_i^{'}\}_{i=1}^{n} >$ according to $< n_i \times I, n_d \times D, n_u \times S_u, n_h \times S_h >_j$.
  $< \{S_i\}_{i=1}^{n} >_j \Leftarrow$ TER score between $< \{Y_i^e\}_{i=1}^{n} >_j$ and $< \{Y_i^{'}\}_{i=1}^{n} >_j$
  $P = P \cup < \{X_i^{'}, Y_i^{e}, S_i\}_{i=1}^{n} >_j$
  $j = j + 1$
end while
return P

large-scale training corpus and a substantial time of training. Second, the MT translations are system-specific, differing from numbers or even types of errors from the target data. In other words, MT generated training data may not be informative enough, which is the focus on the proposed approach(empirical results is provided in sub-section 4.2).

3.2 Sentence Specific Expansion of Pseudo Data.

After the pre-training under pseudo data, the QE model will converge, but not necessarily at the global optimum because of the noise in pseudo data. To deal with this issue, an approach to sentence specific expansion of pseudo data is proposed. The motivation is to provide more pseudo data only for those target samples not well trained.

Leveraging the fact that there are error between QE model predicted score and gold score, we define error distance (ED) for modeling the difference between the score given by QE model and the score assigned to the manual labeled data as follow:

$$ED = (QEScore - GoldScore)^2 \tag{1}$$

We hope to provide new information for the samples unsuccessfully learned for the model by oversampling pseudo training sample. Therefore, we use error distance to measure whether translation errors in the target human labeled QE corpus have been learned well.

Specifically, we firstly use the pre-trained QE model to predict QE score for the sentences pair in labeled QE corpus. On the basis of that, we compute the error distance for predicted scores and gold scores. Then we simply choose top half error distance samples in the labeled QE corpus and apply oversampling to expand more pseudo data for these samples. In this paper, oversampling means

re-feed the pseudo data already generated. All these oversampling samples will be used to continue training the pre-trained QE model.

4 Experiments and Results

4.1 Experiments Setting

Dataset. In our experiments, we use the benchmark data from WMT2015 and CWMT2018 QE task, which contain 3 translation pairs: English-to-Spanish(en-es), Chinese-to-English(zh-en) and English-to-Chinese(en-zh), to evaluate our proposed method. For the WMT2015 QE task, we choose the development data provided by official as labeled QE corpus to generating pseudo data and test our method on official test set. Also for CWMT2018 QE task, we use the development set as labeled QE corpus for generating pseudo data and test data provided by official are used for test set. In addition, we set different size of pseudo data by controlling the number of top-n similar sentences in candidate corpus selection.

In order to generate pseudo data, we need to collect large scale candidate corpus from bilingual corpus. In this process, Bilingual data is employed. For the WMT en-es QE task, we use Europarl v7 [12] as bilingual corpus. For the CWMT en-zh and zh-en QE task, we use the bilingual data provided by CWMT2018 MT task.

Model and Training. In order to verify our method, we choose two different but typical neural QE models for the experiment.

- Bilingual sentence representation QE model(**BSR-QE**) [7]: BSR-QE used Bi-LSTM to get two context vectors and computed the weighted cosine distance of the two vectors to estimate the QE score .
- Bilingual expert QE model(**BE-QE**) [8]: BE-QE firstly pre-trained a transformer based bilingual expert model under bilingual corpus, and then extracting QE features for Bi-LSTM QE model based on the result of bilingual expert model.

In order to mitigate the effects of noise in pseudo data, we adopt the two-step training strategy for training QE model, and all the pseudo data are actually employed only in the stage of pre-training QE model. The best parameters achieved are kept and updated by the provided labeled data in the stage of fine-tuning.

Baselines. We set up a variety of baseline pseudo approaches include:

- Random bilingual data: parallel bilingual sentence pairs are used as positive cases while random bilingual sentence pairs are used as negative case for pre-training QE model.
- MT data: a natural idea is directly using MT results as pseudo data for pre-training QE model. We first train an NMT systems[13] by larger scale bilingual corpus, then generating translation for sources sentences in bilingual corpus. Based on that, TER score between MT translation and target sentence in bilingual corpus will be used as QE score of MT translation.

Evaluation. Following the practices in WMT2015 and CWMT2018, The primary metrics of sentence level QE task are Pearson's correlation(for CWMT QE task) and Spearman's rank correlation(for WMT QE task) of the entire testing data. Alternatively, mean average error (MAE) and root mean squared error (RMSE) is used to measure the performance of overall predictions.

4.2 Experiments Result

Result on BSR-QE model. In this part, we will analyze the performance of our approach to different language QE tasks. For comparison, we list the results of baseline system in QE task WMT2015. And we also list the results of another two approaches of using pseudo data: one is generated by MT, and the other one is generated randomly from bilingual data [7]. In addition, we list the results on CWMT2018 en-zh and zh-en QE task. All these results are shown in Table 1 and 2. Noted that we generate 200K pseudo data both for target oriented data generation and MT results in pre-training QE model at Table 1 and 2.

Table 1. BSR-QE results of sentence level QE on WMT2015

Task	Pre-training data	MAE ↓	RMSE ↓	Spearman's ↑
en-es	WMT2015 Official baseline system	14.821	19.132	0.133
	Random bilingual data	14.481	18.862	0.271
	MT data	14.943	20.611	0.226
	Target oriented pseudo data	14.232	18.663	0.291
	Sentence specific expansion	14.152	18.123	0.306

Table 2. BSR-QE results of sentence level QE on CWMT2018

Task	Pre-training data	MAE ↓	RMSE ↓	Pearson's↑
zh-en	Random bilingual data	0.157	0.227	0.340
	Target oriented pseudo data	0.155	0.221	0.387
	Sentence specific expansion	0.156	0.220	0.405
en-zh	Random bilingual data	0.188	0.238	0.223
	Target oriented pseudo data	0.174	0.226	0.274
	Sentence specific expansion	0.168	0.223	0.302

From the result, we can find that our proposed method obtains significant improvements over two baselines. Also in en-zh and zh-en QE task, our proposal can significantly improve the QE performance.

Result on BE-QE model. Different from BSR-QE model, BE-QE model needs firstly pre-trained neural bilingual expert model under larger scale of bilingual

corpus. In order to test our method on this framework, we use the pseudo data just for pre-training QE model instead of bilingual expert model.

Specifically, we firstly use bilingual data to train the neural bilingual expert model. Then we extract QE features for pseudo data by the neural bilingual expert model to pre-train QE model. At last, the QE model will be fine-tuned by the QE features extracted from real QE data. Noted that we also use 200K pseudo data for experiment. The result on sentence level zh-en QE task can be seen in Table 3.

Table 3. BE-QE results of sentence level zh-en QE on CWMT2018

Method	Pearson's↑
BE-QE baseline	0.465
BE-QE baseline + pseudo data	0.482

From Table 3, we can find that our method can outperform the BE-QE baseline method. Although BE-QE model used neural bilingual expert model, which is pre-trained under larger scale bilingual data, our target oriented pseudo data generation also can get effective improvement. The result verifies our approach also can be useful for the two-step QE framework, which contains feature extractor model and QE model.

4.3 Discussion

The scale of pseudo data. In this part, we will compare the performance of pseudo data at different corpus size on BSR-QE model. For comparison, we list the results of different pseudo data corpus size in WMT2015 and CWMT2018 QE task. All these results are shown in Fig. 3.

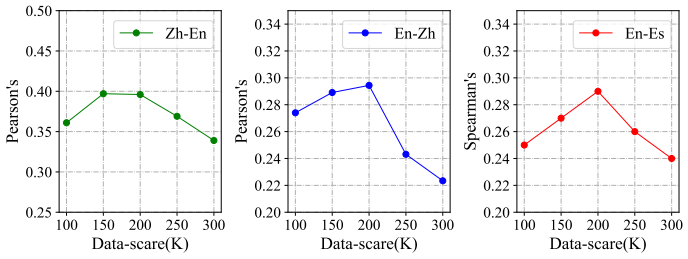

Fig. 3. Performance of our approach when changing the scale of pseudo data

From Fig. 3, we find that the performance of our approach rises firstly when increasing the scale of pseudo data, then drops. This situation reflected target oriented data generation dependent on the similar translation error distribution

between labeled QE corpus. As the increases of data scale, more sentences which is not very similar to the labeled QE corpus are collected, then more error will produce.

Effects of different properties for pseudo data. We choose pseudo data with a data size of 200K on BSR-QE model as a baseline to explore the effect of different pseudo data property for the performance of the model. In our work, the pseudo data property means the editing method for candidate data. pseudo data property includes the number of editing words and type of editing. To explore the effectiveness of different property in pseudo data, we generated pseudo data with different property.

For the pseudo data property values, we use random generation or artificially control. In detail, for the number of editing word, we set random number or same as labeled QE data. As for the type of editing, we also set random type or same as labeled QE data. The result can be seen in Table 4.

Table 4. Effects of different properties for pseudo data on En-Es QE task

Error word number		Error type		Spearman's↑
Random	As labeled	Random	As labeled	
✓		✓		0.205
✓			✓	0.222
	✓	✓		0.214
	✓		✓	0.291

From the Table 4, we can know that the best result is controlling the number of error words and error type as human labeled QE data. We can conclude that the number of error word and the error type play an important role in pseudo data generation and it needs to be artificially controlled according to human labeled data distribution.

Effects of two-step training for QE model. We also test out whether two-step training method is effective. In this experiments, we used three different types of training data: only pseudo training data, only QE data, and two-step training method, which means using pseudo training data in the pre-training step and QE data for fine-tuning step. Noted that we choose pseudo data with a data size of 200K on BSR-QE model for experiments. The results are given in Table 5.

From Table 5 we can find that using either pseudo training data or QE data alone can not bring inspiring result. By using two-step training method, the model could give significant improvements, which demonstrate the effectiveness of two-step training approach. An intuition behind this phenomenon is that though pseudo training data is fairly big enough to train a reliable model parameters, there is still a gap to the real QE tasks.

Table 5. Effects of two step training on WMT2015 QE task

Training data	Spearman's↑
Only Pseudo Training Data	0.136
Only QE Data	0.232
Pre-training by pseudo data + fine-tuning by QE data	0.291

4.4 Case Study

To further understand our method, we select some test results from English-Chinese QE data, and compare the scores predicted from QE model pre-trained by random bilingual data and our proposal. As illustrated in Fig. 4, for each of the machine translations , we show their respective actual HTER scores, as well as the predicted QE scores from QE model trained by random bilingual data and target oriented pseudo data. At the same time, we sort the quality of the three translations according to their respective scores.

Sentence type	Sentence	Gold Score	Random data	Our proposal
Source Language	This results in language models that are too large to easily fit into memory.	-	-	-
Reference	这导致语言模式过于庞大而不能轻易地放入存储器中。	-	-	-
MT result1	语言模型太大以至于无法很好地适应内存容量。	0.47 (3)	0.24 (1)	0.23 (3)
MT result2	语言模型的结果太多以致于很难融入记忆。	0.2 (2)	0.23 (2)	0.20 (2)
MT result3	这个结果在语言的模型太大容易地装入内存。	0.15 (1)	0.31 (3)	0.18 (1)
Source Language	Mobile advertising revenue represented roughly 73% of advertising revenue .	-	-	-
Reference	移动广告收入约为广告总收入的 73 %。	-	-	-
MT result1	手机广告的收入约是广告收入的百分之 73 。	0.32 (3)	0.21 (2)	0.35 (3)
MT result2	手机广告收入约为广告总收入的 73 %	0.24 (2)	0.23 (3)	0.32 (2)
MT result3	移动广告收益在广告业总收益中约占 73 %。	0.05 (1)	0.19 (1)	0.21 (1)

Fig. 4. Case of test result from QE model pre-trained by random bilingual data and our proposal

From the result, we find that the translation quality ranking given by the QE model trained on our method is consistent with the quality ranking of the gold score, so that the quality of the translation can be predicted more accurately.

5 Conclusion

To alleviate the data shortage in training of neural QE model, we present a target-oriented approach to automatically generating labeled samples. The key idea is that generating pseudo training samples with a purpose of similar distribution of translation error of the target is helpful to train the neural model. Furthermore, we propose a sentence specific expansion method, to maximally mining the utility of pseudo data. The experimental results on the English-Spanish, English-Chinese and Chinese-English sentence-level quality estimation task shows a significant improvement of our approach.

In the future, we plan to use reinforcement learning to learn a policy for generating the most informative pseudo data for QE task. In addition, we will expand the target-oriented pseudo data generating method for other NLP tasks.

Acknowledgments . This paper is supported by the National Key R&D Program of China (No. 2018YFC0830700).

References

1. Specia L., Turchi M., Cancedda N., Dymetman, M., Cristianini, N. Estimating the sentence-level quality of machine translation systems. In 13th Conference of the European Association for Machine Translation. 2009: 28-37.
2. Kreutzer J., Schamoni S., Riezler S. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015: 316-322.
3. Kim H., Lee J H. A recurrent neural networks approach for estimating the quality of machine translation output. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 494-498.
4. Patel R N., Sasikumar M. Translation Quality Estimation using Recurrent Neural Network. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016, 2: 819-824.
5. Martins, A. F., Astudillo, R., Hokamp, C., Kepler, F. Unbabel's participation in the WMT16 word-level translation quality estimation shared task. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016, 2: 806-811.
6. Ive J., Blain F., Specia L. DeepQuest: a framework for neural-based quality estimation. Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3146-3157.
7. Zhu, J., Yang, M., Li, S., Zhao, T. Learning bilingual sentence representations for quality estimation of machine translation. China Workshop on Machine Translation. Springer, Singapore, 2016: 35-42.
8. Fan, K., Wang, J., Li, B., Zhou, F., Chen, B., Si, L. (2019, July). "Bilingual Expert" Can Find Translation Errors. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 6367-6374).
9. Liu, L., Fujita, A., Utiyama, M., Finch, A., Sumita, E., Liu, L., Sumita, E. Translation quality estimation using only bilingual corpora[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2017, 25(9): 1762-1772.
10. Duma, M., Menzel, W. (2018, October). The Benefit of Pseudo-Reference Translations in Quality Estimation of MT Output. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers (pp. 776-781).
11. Albrecht, J., Hwa, R. (2007, June). Regression for sentence-level MT evaluation with pseudo references. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 296-303).
12. Koehn P. A parallel corpus for statistical machine translation[J]. Proceedings of the Third Workshop on Statisti-cal Machine Translation, 2005(1):3-4.
13. Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Nădejde, M. (2017). Nematus: a toolkit for neural machine translation. arXiv preprint arXiv:1703.04357.