Quality Estimation for Machine Translation with Multi-granularity Interaction*

Ke Tian[1,2] and Jiajun Zhang[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{ke.tian,jjzhang}@nlpr.ia.ac.cn

Abstract. Quality estimation (QE) for machine translation is the task of evaluating the translation system quality without reference translations. By using the existing translation quality estimation methods, researchers mostly focus on how to extract better features but ignore the translation oriented interaction. In this paper, we propose a QE model for machine translation that integrates multi-granularity interaction on the word and sentence level. On the word level, each word of the target language sentence interacts with each word of the source language sentence and yields the similarity, and the L_∞ and entropy of the similarity distribution are employed as the word-level interaction score. On the sentence level, the similarity between the source and the target language translation is calculated to indicate the overall translation quality. Finally, we combine the word-level features and the sentence-level features with different weights. We perform thorough experiments with detailed studies and analyses on the English-German dataset in the WMT19 sentence-level QE task, demonstrating the effectiveness of our method.

Keywords: quality estimation · neural machine translation · multi-granularity.

1 Introduction

In recent years, neural machine translation (NMT) [1–4] makes great progress, and quality estimation of machine translation methods has also received much attention. Usually, evaluating system quality is to calculate the BLEU [5] when there are one or more reference translations available. In the model prediction or practical applications, it is costly to collect high-quality reference translations for each translation. Quality estimation of machine translation is the task of evaluating the translation system quality without reference translations. The prediction results can quickly measure the quality of the system translation. It plays an indispensable guiding role in post-translation editing and computer-aided translation. In the QE task, sentence-level QE is a popular research topic. Most

* Supported by organization x.

sentence-level QE tasks predict a score which indicates how much effort is needed to post-edit translations to be acceptable results as measured by the Human-targeted Translation Edit Rate (HTER). In general, sentence-level QE is seen as a supervised regression task. In traditional feature-based QE approaches, which has 17 features that describe the translation quality, such as translation complexity indicators, fluency indicators, and adequacy indicators, it exploits a support vector regression algorithm to score the translation. With the rapid development of deep learning in natural language processing (NLP), many researchers have applied the neural network model to the QE task. With pre-trained language models showing excellent performance in natural language downstream tasks, multilingual pre-trained language models attract researchers' attention, such as Multilingual BERT [6], XLM [7].

Most researchers only focus on how to extract better features but ignore the translation oriented interaction. Although the word vectors fully interact in the neural network model, we believe that more translation characteristics should be added for cross-lingual tasks such as translation quality estimation. Either between word or sentence translation pairs, more translation oriented features can be tapped in.

In this paper, in order to solve the above problems, we propose a translation quality estimation method that incorporates multi-granularity interaction, making full use of the interactive information on the word and sentence level. And this method achieves good results in the WMT19 sentence-level QE task on the English-German dataset. On the word level, each word of the target language sentence interacts with each word of the source language sentence and yields the similarity, and the L_∞ and entropy of the similarity distribution are employed as the word-level interaction score. In terms of sentence level, we calculate the similarity between sentence vectors by cosine similarity. We specifically analyze that the similarity of translated word pairs can effectively measure translation quality.

2 Related Work

Traditional baseline model QuEst++ [8] extracted features based on handcrafted rules and used SVM regression to predict the score. With the great success of deep neural networks on many tasks in natural language processing(NLP), many researchers have applied the neural network model to the QE task. Shah et al. [9] combined neural features that include word embedding features and neural language model features with other features extracted by QuEst++. Kim et al. [10–12] proposed the Predictor-Estimator framework, within which predictor is product quality vectors by a bidirectional RNN encoder-decoder with attention mechanism, and estimator uses quality vectors to predict the score. Li et al. [13] combined the two-stage predictor-estimator framework to extract more abundant features through joint training. Fan et al. [14] proposed "Bilingual expert" model which uses transformer [15] architecture as feature extractor. These models have achieved good results by using the powerful feature extraction ability of neural networks. In the past two years, the pre-training models such as EMLo [16], GPT [17], and BERT have developed rapidly and greatly improved the performance of downstream tasks in natural language processing. Lu et al. [18] proposed a feature extraction method based on the multi-language pre-training language model so that the source and target language sentence can interact more intensively, which is of great help to the cross-language task.

Kepler et al. [19] proposed the model that mainly integrates different sub-models, such as APE-BERT, PREDEST-BERT, and PREDEST-XLM, etc. From their experimental results, the key to the superior performance of their model depends on the PREDEST-XLM sub-model. Zhou et al. [20] used the translation model as a feature extraction module, and mainly improved the "Bilingual Expert" model with a SOURce-Conditional ELMo-style (SOURCE) strategy. Hou et al. [21] employed bi-directional translation knowledge and large-scale monolingual knowledge to the QE task. Kim et al. [22] proposed a "bilingual" BERT using multi-task learning for machine translation quality estimation.

In the above methods of QE, researchers mostly focus on using different model to extract better features, such as neural networks using recurrence, convolution and self-attention. But they ignored the translation oriented interaction. For the disadvantages of the above model, we will propose a QE model for machine translation that integrates multi-granularity interaction on the word and sentence level.

3 Methodology

In this section, Fig. 1 shows the model architecture. Following the recent trend in the NLP task exploiting large-scale language model pre-training for a series of different downstream tasks, we used multilingual BERT as feature extractors. The features fuse the translation oriented interaction on the word and sentence level and they are used to predict HTER score.

3.1 Model Architecture

The model consists of a feature extractor that produces contextual token representations, and an estimator that turns these representations into predictions for sentence-level scores. Although the multilingual pre-trained language model is well suited to handle cross-language tasks, it is still a single language followed by the same language is used as input for pre-training. In order to adjust the model and make it compatible with the input combination of the source and target language sentences, we adopt a cross-language joint encoding method that uses bilingual parallel corpus to pre-training multilingual BERT. This significantly improves the performance of the model.

As a multilingual model, source and target language sentences need to be input into the model together, so that the words in and between sentences can fully interact and get better vector representation. We combine the two sentences as input according to the template: [CLS] target [SEP] source [SEP], where

[CLS] and [SEP] are special symbols from BERT, denoting the beginning of the sentence and sentence separators, respectively. The pre-training sub-task is to predict whether the second sentence is the translation of the first sentence.

Instead of just using contextual token representations to predict the score, we allow the contextual token representations of source and target language sentences to further interact. In other words, interaction is to explicitly model translation oriented features on the word and sentence level. The details of the interaction will be described in the next section. Finally, the word-level and sentence-level translation oriented features fuse with contextual token representation to predict scores.

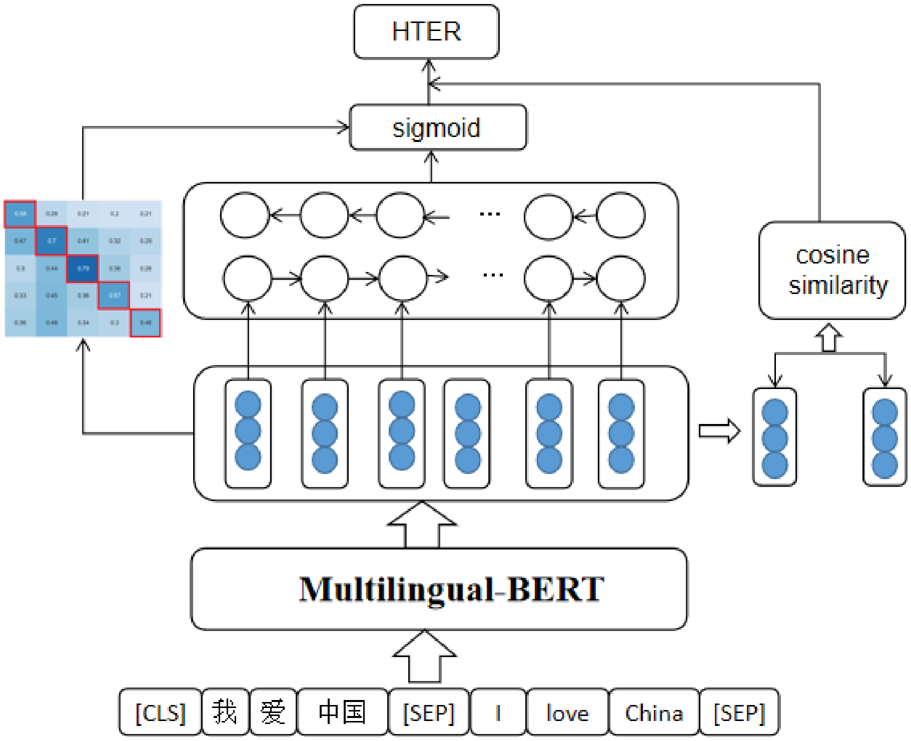


Fig. 1. The architecture of the proposed QE model that fuses multi-granularity interaction.

3.2 Multi-granularity Interaction

Word-level feature Each word of the input bilingual sentences pair (source language sentence S, target language sentence T) is represented by pre-trained multilingual BERT $s = (s_1, s_2, ..., s_i)$ and $t = (t_1, t_2, ..., t_j)$. Compared to word embeddings, contextual embeddings provide different vector representations of the same word in different contexts. Since BERT uses subword encoding, we average the vectors of all subwords to represent a complete word. Next, we calculate the similarity matrix between each word in the target translation and each word in the source language sentence. The L_∞ and entropy of the similarity distribution are employed as that our method interaction core. Finally, the similarity scores of the translation word pairs are selected from the similarity matrix. There similarity scores are concatenated with entropy as features in the translation oriented interaction. As is shown in Fig. 2.

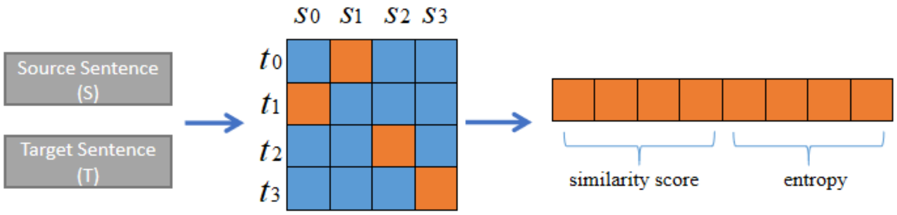


Fig. 2. An illustration of the word-level interaction.

we use the cosine metric to compute the similarity of each word pair between source and target language sentences. s_i denotes the i^{th} token embedding of the source language sentence. t_j denotes the j^{th} word embedding of the target language sentence.

$$sim_w = \frac{s_i \cdot t_j}{\|s_i\| * \|t_j\|} \tag{1}$$

The similarity distribution is the similarity score of a word in the target language sentence and each word in the source language sentence. The entropy of similarity distribution measures the confidence level of translation. H_j denotes entropy of the target language sentence's j^{th} word. p_i denotes probability after softmax.

$$H_j = -\sum_{i=1}^{n} p_i \cdot logp_i \tag{2}$$

Finally, the similarity score of word level and the entropy of similarity distribution are combined to form word-level interact feature. T denotes the number of words in the target language sentence.

$$E_i = Concat(sim_{w_1}, sim_{w_2}, \cdots, sim_{w_T}, H_1, H_2, \cdots, H_T) \tag{3}$$

Sentence-level feature The vector representations of all the words in the target language sentence are averaged as the vector representation of the current sentence. And the calculation is shown in Formula 4. T denotes the number of words in the target language sentence. h_{tgt} denotes target language sentence vector. It it the same for the source language sentence.

$$h_{tgt} = \frac{1}{T}\sum_{j=1}^{T} t_j \tag{4}$$

The similarity calculation of sentence level is same as word level. h_{src} denotes source language sentence vector. And the calculation is shown in Formula 5.

$$sim_s = \frac{h_{src} \cdot h_{tgt}}{\|h_{src}\| * \|h_{tgt}\|} \tag{5}$$

Ensemble feature We concatenate the word vector and the feature vector of the translation oriented interaction, and use the sigmoid activation function to map the value between 0 and 1. Then we subtract the cosine similarity from 1, since there is a negative correlation between cosine similarity and HTER value. Finally, we predict HTER score by linearly interpolating the word and sentence-level features.

$$hter = \lambda_1 \cdot sigmoid\left((E_w \bigoplus E_i) \cdot W\right) + \lambda_2 \cdot (1 - sim_s) \tag{6}$$

λ_1 and λ_2 denote word-level and sentence-level feature weights that are hyper-parameter. E_w denotes all word embedding. E_i denotes the translation oriented interaction feature embedding. sim_s denotes the similarity score of sentence vector between the source and target language sentence. $\bigoplus$ denotes vector concatenation operation. W denotes the learnable parameter matrix.

3.3 Model Training

Because the size of the training set for the QE task is too small to train the model, we use about 5 million bilingual parallel corpora to pre-train multilingual BERT. This also makes it more familiar with the input of the combination of source and target language sentences.

Assume that the training set for the QE task includes N source language sentences $x^{(n)}$, the target language sentences $y^{(n)}$, and the corresponding gold standard labels $HTER^{(n)}(n = 1, ..., N)$. The training objective is to minimize the mean square error over the training data:

$$R_{MSE} = \frac{1}{N}\sum_{i=1}^{n}(QE_{score}(x^{(n)}, y^{(n)}) - HTER^{(n)})^2 \tag{7}$$

4 Experiments

4.1 Dataset

The bilingual parallel corpus that we use for pre-trained multilingual BERT is officially released by the WMT17 Shared Task: Machine Translation of News1, including Europarl v7, Common Crawl corpus, News Commentary v12, and Rapid corpus of EU press releases. In the pre-training stage, we construct positive and negative samples from bilingual data. The positive sample is the parallel data correctly translated, while the negative sample is the source language sentence and the randomly sampled translation. The positive and negative samples are randomly shuffled to construct pre-train data.

In the QE experiment, to test the performance of the proposed QE model, we conduct experiments on the WMT19 sentence-level QE task for English-German (en-de) direction. The details of the dataset are shown in Tables 1.

Table 1. Details of the en-de dataset of the WMT19 sentence-level QE task.

	Train	dev	test
sentences	13442	1000	1023

4.2 Experimental Setup

In the experiment, we use the multilingual BERT after pre-training with bilingual parallel corpus, which has 12 Bi-transformer [13], and the total number of parameters is 1.1×10^8. During the training, we limit the number of training epoch as 3, learning rate 2×10^{-5}, batch size 32, max sequence length 128.

In the pre-training, we keep the default hyperparameter settings of the multilingual model. For the quality estimator module, the number of hidden units for forward and backward LSTM is 1000. We use a minibatch stochastic gradient descent algorithm and Adam to train the QE model.

4.3 Experimental Result

In this section, we will report the experimental results of our proposed model on the WMT19 sentence-level QE task for the English-German direction. And we list the results of other models in the WMT19 sentence-level QE task and the baseline model are listed in the table 2.

Table 2. Results of the different models on the WMT19 sentence-level QE task.

system	pearson	spearman
UNBABEL	0.5718	0.6221
PREDEST-BERT	0.5190	-
CMULTIMLT	0.5474	0.5947
NJUNLP	0.5433	0.5694
ETRI	0.5260	0.5745
baseline	0.4001	0.4607
Our model	0.5496	0.5980

From the results in Table 2, we can see that our proposed model outperforms most of the baseline models. We also observe that our model underperforms the model UNBABEL. The reason is that UNBABEL is an ensembled model which integrates seven models. When we compare our model to their best single model PREDEST-BERT, we find that our model performs much better.

Then, we also compare the results of experiments that fuse different features, as shown in Table 3. We find that good performance can be achieved when all features are ensembled.

Table 3. Results of the models that fuse different features on the WMT19 sentence-level QE task.

system	pearson	spearman
BERT+LSTM	0.5057	0.5345
BERT+LSTM+word-level	0.5120	0.5606
BERT+LSTM+sentence-level	0.5332	0.5571
BERT+LSTM+sentence-leve+word-level	0.5496	0.5980

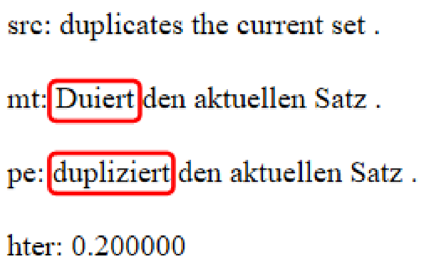
From the results in Table 3, both word-level features and sentence-level features are helpful to our tasks. It can be seen from the comparison that the features of sentence-level improve more based on the original model.

4.4 Word-level Feature Analysis

We select an example from the dataset to explain the word level feature. As shown in Fig. 3 and Fig.4. It can be seen that 'dupliziert' is wrongly translated as 'Duiert'. According to the similarity matrix, the similarity score is slightly lower between the two words. However, other words are correctly translated, the corresponding similarity scores are much higher. As shown in Fig. 4, the similarity score corresponding to the punctuation is also low, and the reason is probably that the punctuation by itself does not take any meaning, unlike the nouns or verbs. Thus, we believe that the similarity score between words across languages can measure the quality of translation.

5 Conclusion

In this paper, we propose a translation quality evaluation method that fuses multi-granularity interaction. On the word level, each word of the target language sentence interacts with each word of the source language sentence, and the L_∞ and entropy of the similarity distribution are employed as the word-level interaction score. There similarity scores are concatenated with entropy as features in the translation oriented interaction. On the sentence level, the source and target language sentence vector similarity are used to measure the quality



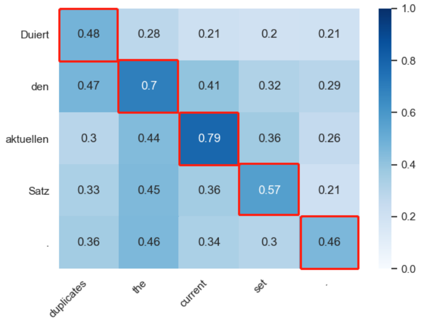
src: duplicates the current set .

mt: Duiert den aktuellen Satz .

pe: dupliziert den aktuellen Satz .

hter: 0.200000

Fig. 3. An illustration of example.



	duplicates	the	current	set	.
Duiert	0.48	0.28	0.21	0.2	0.21
den	0.47	0.7	0.41	0.32	0.29
aktuellen	0.3	0.44	0.79	0.36	0.26
Satz	0.33	0.45	0.36	0.57	0.21
.	0.36	0.46	0.34	0.3	0.46

Fig. 4. An illustration of word-level similarity matrix.

of the overall translation. Then, we predict the HTER score by linearly interpolating the word and sentence-level feature. Experimental results demonstrate that the method obtains more improvements over most models. And experimental results illustrate the validity of word-level interaction. In the future, we will conduct relevant experiments to verify the effect of the model in other language directions, and explore how to apply our approaches for word-level and document-level QE tasks.

6 Acknowledgments

The research work has been funded by the Natural Science Foundation of China under Grant No. U1836221 and 61673380. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources and Beijing Academy of Artificial Intelligence (BAAI2019QN0504).

References

1. Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua.: Neural machine translation by jointly learning to align and translate. In: arXiv preprint arXiv:1409.0473.(2014)
2. Zhang, Jiajun and Liu, Shujie and Li, Mu and Zhou, Ming and Zong, Chengqing.: Bilingually-constrained phrase embeddings for machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistic. pp.111-121(2014)
3. Zhang, Jiajun and Zong, Chengqing.: Deep neural networks in machine translation: An overview. In: .IEEE Intelligent Systems. pp.16-25(2015)
4. Zhou, Long and Zhang, Jiajun and Zong, Chengqing.: Synchronous bidirectional neural machine translation. In: Transactions of the Association for Computational Linguistics. pp.91-105(2019)

5. Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp.311-318(2002)
6. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805. 2018
7. Lample, Guillaume and Conneau, Alexis.: Cross-lingual language model pretraining. In: arXiv preprint arXiv:1901.07291. 2019
8. Specia, Lucia and Paetzold, Gustavo and Scarton, Carolina.: Multi-level Translation Quality Prediction with QuEst++. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, pp.115-120(2015)
9. Shah, Kashif and Ng, Raymond WM and Bougares, Fethi and Specia, Lucia.: Investigating continuous space language models for machine translation quality estimation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp.1073-1078(2015)
10. Kim, Hyun and Jung, Hun-Young and Kwon, Hongseok and Lee, Jong-Hyeok and Na, Seung-Hoon,: Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. In: ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). pp.1-22(2017)
11. Kim, Hyun and Lee, Jong-Hyeok.: Recurrent neural network based translation quality estimation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp.787-792(2016)
12. Kim, Hyun and Lee, Jong-Hyeok and Na, Seung-Hoon.: Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In: Proceedings of the Second Conference on Machine Translation. pp.562-568(2017)
13. Li, Maoxi and Xiang, Qingyu and Chen, Zhiming and Wang, Mingwen.: A unified neural network for quality estimation of machine translation. In: IEICE TRANSACTIONS on Information and Systems. pp.2417-2421(2018)
14. Fan, Kai and Wang, Jiayi and Li, Bo and Zhou, Fengming and Chen, Boxing and Si, Luo.: "Bilingual Expert" Can Find Translation Errors. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp.6367-6374(2019)
15. Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia.: Attention is all you need. In: Advances in neural information processing systems. pp.5998-6008(2017)
16. Peters, Matthew E and Neumann, Mark and Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke.: Deep contextualized word representations, In: arXiv preprint arXiv:1802.05365. 2018
17. Radford, Alec and Narasimhan, Karthik and Salimans, Tim and Sutskever, Ilya.: Improving language understanding by generative pre-training. In: URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf 2018
18. Lu, Jinliang and Zhang, Jiajun.: Quality estimation based on multilingual pre-trained language model[J].J Xiamen Univ NatSci, 2020,59(2).(in Chiese)
19. Kepler, Fabio and Trénous, Jonay and Treviso, Marcos and Vera, Miguel and Góis, António and Farajian, M. Amin and Lopes, António V. and Martins, André F. T.: Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp.78-84(2019)

20. Zhou, Junpei and Zhang, Zhisong and Hu, Zecong.: SOURCE: SOURce-Conditional Elmo-style Model for Machine Translation Quality Estimation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp.106-111(2019)
21. Hou, Qi and Huang, Shujian and Ning, Tianhao and Dai, Xinyu and Chen, Jiajun.: NJU Submissions for the WMT19 Quality Estimation Shared Task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp.95-100(2019)
22. Kim, Hyun and Lim, Joon-Ho and Kim, Hyun-Ki and Na, Seung-Hoon.: 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). 85-89(2019)