HW-TSC Submission for CCMT 2022 Translation Quality Estimation Task

Chang Su[1], Miaomiao Ma[1], Hao Yang[1], Shimin Tao[1], Jiaxin Guo[1], Minghan Wang[1], Min Zhang[1], and Xiaosong Qiao[1]

2012 Labs, Huawei Technologies CO., LTD, Beijing, China
{suchang8, mamiaomiao, yanghao30, taoshimin, guojiaxin1, wangminghan, zhangmin186, qiaoxiaosong}@huawei.com

Abstract. This paper presents the method used by Huawei Translation Services Center (HW-TSC) in the quality estimation (QE) task — sentence-level post-editing effort estimation — in the 18th China Conference on Machine Translation (CCMT) 2022. This method is based on a predictor-estimator model. The predictor is an XLM-RoBERTa model pre-trained on a large-scale parallel corpus and extracts features from the source language text and machine-translated text. The estimator is a fully connected layer that is used to regress the post-editing distance scores using the extracted features. In the experiment, it is found that pre-training the predictor with the semantic textual similarity (STS) task in the parallel corpus and using augmented training data constructed by different machine translation (MT) engines can improve the prediction effect of the Human-targeted Translation Edit Rate (HTER) in both Chinese-English and English-Chinese tasks.

Keywords: translation quality estimation (QE), Transformer, pre-training

1 Introduction

The off-line technical estimation task of the 18th China Conference on Machine Translation (CCMT) includes a sentence-level Chinese-English and English-Chinese machine translation (MT) quality estimation (QE) task, which aims to measure the MT quality by estimating the Human-targeted Translation Edit Rate (HTER) of the translation without reference translations. This paper describes in detail the data processing strategies, technical methods, and model structure used by HW-TSC's Text Machine Translation Laboratory in this estimation task, as well as the performance of the used models in the Chinese-English and English-Chinese MT QE tasks.

2 Estimation System

In this sentence-level QE task, HW-TSC uses the predictor-estimator structure proposed in the early research[1]. As shown in Figure 1, the language model

XLM-RoBERTaBase[2] (XLM-RB) is used as the predictor (L = 12, H = 768, A = 12; Total Parameters = 288M) to extract source features from the source text and target features from the target text. After that, average pooling is applied to the extracted features of each sentence to obtain the source sentence features and target sentence features. The source sentence feature (SF), target sentence feature (TF), difference between the SF and TF (diff), and dot product of the source and target text features (prob) are concatenated to obtain a global feature. The global feature is sent to an estimator constructed by two fully connected layers (FFNs), which maps the feature to sample label space and performs regression prediction on the HTER score.

The final system submitted uses the ensemble model policy that uses the model with Dropout to average multiple predicted results, thereby improving the model robustness and significantly improving accuracy of the system in the test set. The ensemble models are:

1) Models that achieve the best perform in the development set during multiple training processes;
2) Best models selected from step 1) based on the development set, with random Dropout enabled.

3 Data

Training Data

1) In the English-Chinese task, the CCMT 2022 sentence-level translation QE task provides 3043 source sentences and 14,789 translations and corresponding editing results.
2) In the Chinese-English task, the CCMT 2022 sentence-level translation QE task provides 2503 source sentences and 10,070 translations and corresponding edited translations.
3) Google, Baidu, Youdao, and Huawei translation engines are used separately to translate the source sentences provided by the CCMT 2022 sentence-level translation QE task. The obtained translations generate additional training data together with the provided edited translations.
4) In addition to the data provided in the QE task, HW-TSC also uses the Chinese corpora provided in the English-Chinese, Chinese-English, Mongolian-Chinese, Uyghur-Chinese, and Tibetan-Chinese tasks of the CCMT 2022 bilingual translation task, as well as the English-Chinese and Chinese-English parallel corpora.

Development Data

1) In the English-Chinese task, the CCMT 2022 sentence-level translation QE task provides 2826 (1381 + 1445) source sentences, translations, and corresponding edited translations.

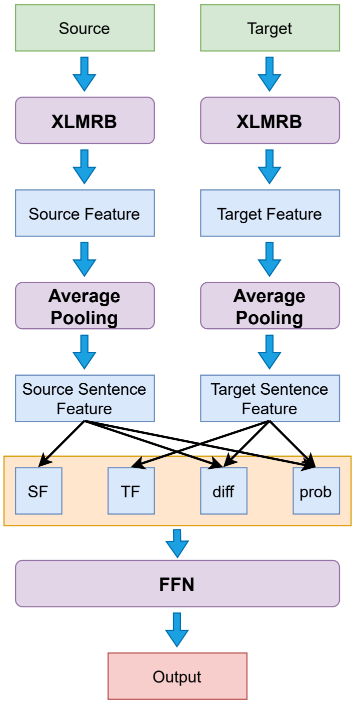

Fig. 1: Predictor-estimator based QE model for estimating sentence-level HTER score

2) In the Chinese-English task, the CCMT 2022 sentence-level translation QE task provides 2528 (1143 + 1385) source sentences, translations, and corresponding edited translations.

Test Data
The off-line test set of the CCMT 2022 provides 10,000 parallel sentence pairs for the English-Chinese and Chinese-English sentence-level translation QE tasks separately.

4 Method

4.1 System Training

The model system used by HW-TSC is trained in three steps:

1) Chinese language model training. Referring to the previous research[3], in this paper, a masked language model (MLM) is trained on a large-scale Chinese corpus. This generates a model for extracting Chinese text features, which is used as a center language encoder (CLE) for the next-step training.

From the word tokens of the Chinese sentences, one token is randomly selected and masked and then sent to the Transformer Encode. The obtained word feature vector is sent to a fully connected classification model, and the model predicts the masked word token, as shown in Figure 2a.

2) Predictor pre-training. According to an early work[4], in this paper, the XLMRB model proposed in Section 1 is trained with the semantic textual similarity (STS) task on English-Chinese and Chinese-English parallel corpora. On the parallel corpora, the XLM-RB obtains feature vectors of the Chinese and English sentences separately, and the CLE model obtains the Chinese sentence feature vector. The mean squared error (MSE) loss function is used for separate supervised training of these vectors, making the sentence feature vectors obtained by the XLM-RB highly similar, as shown in Figure 2b.
3) Translation QE model training. The XLM-RB trained in step 2 is used as the predictor to train the translation QE model on the translation QE training set.

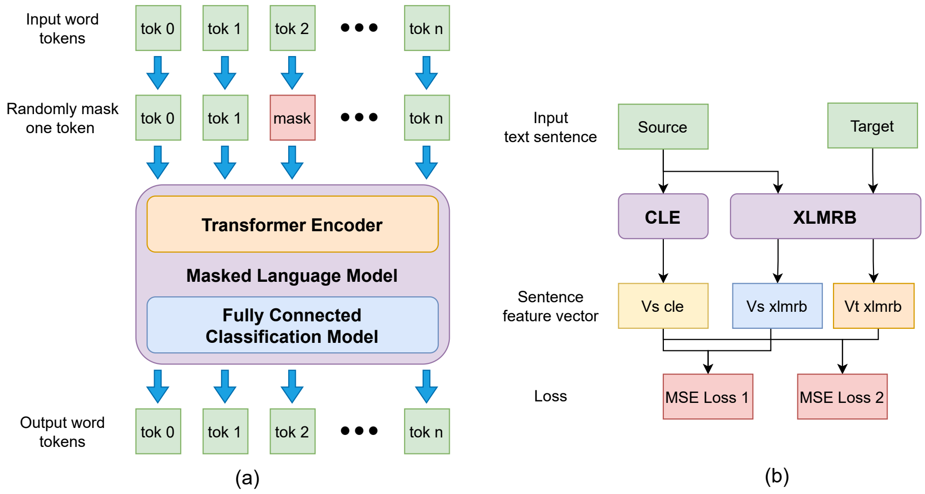

Fig. 2: (a):Masked language model, (b): Schematic diagram of the parallel corpus semantic textual similarity training task

4.2 System Test

As described in Section 1, the ensemble model policy is used in the final system submitted. In this policy, multiple models are used to separately predict the HTER scores of the sentences in the test set, and an average value of the HTER scores of each sentence is used as a score of the ensemble policy.

5 Experiment

5.1 System Environment

OS: Ubuntu 18.04.5 LTS
Deep learning framework: Pytorch 1.8.0
CPU: Intel(R) Xeon(R) Gold 6278C CPU @ 2.60GHz
Memory: 128 GB
GPU: Nvidia Tesla T4
GPU Memory: 16 GB

5.2 Experiment Settings

The system used by HW-TSC is an English-Chinese and Chinese-English multi-task system, and the same system trained is used to obtain the experiment results.

Training Process

Step-1 training 1 described in Section 4: In this paper, the sbert-chinese-general-v2[5] model provided by Hugging Face is used as the pre-trained model to train the MLM on the corpus of 18 million Chinese sentences provided in the English-Chinese, Chinese-English, Mongolian-Chinese, Uyghur-Chinese, and Tibetan-Chinese tasks of the CCMT 2022 bilingual translation task. The pre-trained model sbert-chinese-general-v2 is obtained by training the BERT model of the bert-base-chinese[6] version provided by Hugging Face on SimCLUE, a dataset with millions of semantically similar texts.

Step-2 training described in Section 4: In this paper, the xlm-roberta-base[7] provided by Hugging Face is used as the pre-trained model for STS task training on the bilingual parallel corpus of 9 million of English-Chinese and Chinese-English sentences provided in the CCMT 2022 translation QE task under the sentence-transformers[8] framework.

Step-3 training described in Section 4: In this paper, the English-Chinese and Chinese-English training sets of the CCMT 2022 sentence-level translation QE task are used for training based on the system structure described in Section 2.

Training parameters used in the three steps are shown in Table 1.

Test Process

As described in this section, the model system used by HW-TSC is trained for 9 times. Top 2 models are selected based on the development set, and Dropout 0.1 is applied to the Top 1 model for three test tasks. A total of 6 results are obtained, and the average value of the 6 results is used as the result of the ensemble policy. Due to the limited amount of training data, to prevent overfitting, a model with a small Dropout value is used to predict the test set results, and then an average value is used. In this way, system robustness and accuracy can both be significantly improved. During the training, the maximum

Table 1: Training parameter settings.

Step	Batch size	optimizer	learning rate(lr)	lr scheduler
1	16	Adam[1]	$1.0e^{-4}$	-
2	8	Adam	$5.0e^{-5}$	-
3	8	Adam	$2.5e^{-5}$	Cosine Annealing Warm[2]

Note: 1) Adam: reference [9]. 2) Cosine Annealing Warm: reference [10].

epoch is set to 10. In addition, early stopping of training is enabled: During the training, if the Pearson's correlation coefficient of the validation set is not among the Top 3 for 5 consecutive times, the training is halted immediately.

Comparison training is also performed in the experiment:

1) The XLM-RB model is used to train the model system directly following Step 3 by using the pre-trained model provided by Hugging Face without Step 1 and Step 2 and without using the augmented data produced by Google, Baidu, Youdao, and Huawei translation engines.
2) The Step-3 model training does not use the augmented data (AD) produced by Google, Baidu, Youdao, and Huawei translation engines.

5.3 Experiment Result

In this estimation task, the estimation metrics, mainly the Pearson's correlation coefficient, are automatically measured. Table 2 show the model system performance on the development set.

After comparison, the experiment results of the English-Chinese MT QE task show that:

1) The model pre-trained with the STS task can improve the Pearson's correlation efficient by 8.5% on the development set and by 0.5% on the test set.
2) The model pre-trained using the augmented training data generated by multiple translation engines can improve the Pearson's correlation efficient by 0.7% on the development set and by 0.1% on the test set.
3) The ensemble model policy that uses the model with Dropout to obtain the average value of multiple predicted results can improve the Pearson's correlation efficient by 7.3% on the development set and by 1% on the test set.

After comparison, the experiment results of the Chinese-English MT QE task show that:

1) The model pre-trained with the STS task can improve the Pearson's correlation efficient by 9% on the development set and by 2% on the test set.
2) The model pre-trained using the augmented training data generated by multiple translation engines can improve the Pearson's correlation efficient by 0.5% on the development set and by 1% on the test set.

3) The ensemble model policy that uses the model with Dropout to obtain the average value of multiple predicted results can improve the Pearson's correlation efficient by 5% on the development set and by 1.5% on the test set.

Table 2: Pearson's correlation between prediction of our different system and labels on development and test data.

Language	Model	Dev set	Test set
en-zh	w/o STS[1] & w/o AD	0.4561	0.3549
	w/o AD[2]	0.5413	0.3597
	Top 1[3]	0.5487	0.3607
	Ensemble[4]	0.6211	0.3704
zh-en	w/o STS & w/o AD	0.4663	0.4527
	w/o AD	0.5527	0.4741
	Top 1	0.5574	0.4850
	Ensemble	0.6008	0.5002

Note: 1) w/o STS: The model is trained directly following Step 3 without Step 1 and Step 2. 2) w/o AD: The Step-3 training does not use augmented data. 3) Top 1: The best single model with STS and AD. 4) Ensemble: The policy used by HW-TSC's system.

6 Conclusion

This paper presents HW-TSC's participation in the MT QE task in the 18th China Conference on Machine Translation. In the experiment, the pre-trained language model XLM-RoBERTa is used as the predictor to extract features from the source text and target text. The estimator concatenates the sentence features of the source text and target text after the minus and dot product operations, and performs regression fitting on the HTER scores through the fully connected layer. About the QE training data, the system used by HW-TSC uses the augmented data produced by Google, Baidu, Youdao, and Huawei MT engines. The experiment results show that in the MT QE task, pre-training the predictor with the STS task, using the augmented data produced by multiple translation engines, and adopting the ensemble model policy that uses a model with dropout to average the values of multiple predicted results can improve the accuracy of MT QE results on both the development set and test set. In the future experiment, the model structure of the estimator can be designed and tested in a more refined and effective manner. In addition, the future research and experiment will focus on how to better use the source text and target text for data augmentation on the limited QE data set, so as to generate QE data more similar to the real-world data, as proposed in an early work[11], to further enhance the QE result.

References

1. Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, 2017.
2. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
4. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
5. HuggingFace. sbert-chinese-general-v2 https://huggingface.co/dmetasoul/sbert-chinese-general-v2, 2022.
6. HuggingFace. bert-base-chinese https://huggingface.co/bert-base-chinese, 2022.
7. NilsReimers. Sentencetransformers documentation https://www.sbert.net/, 2022.
8. HuggingFace. xlm-roberta-base https://huggingface.co/xlm-roberta-base, 2022.
9. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
10. Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
11. Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727, 2021.