An improved Multi-task Approach to Pre-trained Model Based MT Quality Estimation

Binhuan Yuan*, Yueyang Li*, Kehai Chen†, Hao Lu, Muyun Yang and Hailong Cao

Department of Computer Science and Technology, Harbin Institute of Technology
20s003024@stu.hit.edu.cn, yuanbinhuan@126.com, chenkehai@hit.edu.cn,
925445386@qq.com, yangmuyun@hit.edu.cn

Abstract. Machine translation (MT) quality estimation (QE) aims to automatically predict the quality of MT outputs without any references. State-of-the-art solutions are mostly fine-tuned with a pre-trained model in a multi-task framework (i.e., joint training sentence-level QE and word-level QE). In this paper, we propose an alternative multi-task framework in which post-editing results are utilized for sentence-level QE over an mBART-based encoder-decoder model. We show that the post-editing sub-task is much more informative and the mBART is superior to other pre-trained models. Experiments on WMT2021 English-German and English-Chinese QE datasets showed that the proposed method achieves 1.2%-2.1% improvements in the strong sentence-level QE baseline.

Keywords: quality estimation, multitask learning, mBART.

1 Introduction

Machine translation (MT) quality estimation (QE) is used as an automatic evaluation for selecting the most suitable machine translation without golden reference. QE is usually implemented either in sentence-level or word-level. Sentence-level QE subtask takes HTER [3] Metric to represent the quality of MT, and the word-level QE task measures the translation quality by generating a quality tag for each word in the output of MT.

The sentence-level and word-level QE subtasks both rely on the triplets of *src* (source sentence), *mt* (machine translated sentence) and *pe* (post-edited sentence). Therefore, sentence-level task is usually training jointly with word-level task so as to improves model performance. It should be noted that, for sentence-level task, *pe* is only used for calculating the label HTER, it is not integrated into the training phase.

In contrast to existing practice, we propose to integrate *pe* into the sentence-level QE model, which is named as *pe* based multi-task learning QE. Following recent employment of pre-trained model, we adopt a multi-task translation QE model based on mBART [4][5]. Evaluated on the WMT2021 English-German/English-Chinese QE dataset and CCMT2021 English-Chinese/Chinese-English QE datasets, the proposed

[*] Equal contribution. Listing order is random.

[†] Corresponding author.

method is revealed a substantial improvement in sentence-level QE compared with jointly training by word-level task. We also reveal that compared to other pre-trained models like BERT [1] and XLM-R [2], mBART achieved better performance.

This paper is organized as follows. In Section 2, we introduce the related work of QE. The proposed multi-task QE method based on mBART is described in Section 3., we report the experiment and results in Section 4, and conclude our paper in Section 5.

2 Related works

With the purpose of estimating machine translations without reference translation, the early research on QE tasks adopted traditional feature extraction and feature selection methods to train the models. Commonly used features included the length of the translation, the matching degree of special symbols, punctuation, and capital letters, etc. Gaussian process [9], heuristic [12] and principal component analysis [16] were commonly used feature selection methods.

With the development of deep learning, QE tasks had gradually shifted into neural network-based framework. The simple network of QE is based on context window [6], and it could be improved by CNN and RNN [15]. In order to integrate large-scale parallel corpus into RNN model, the model could be implemented by Predictor-Estimator structure [7]. With the rise of transformer, transformer-based QE models was implemented for its abilities of using large-scale parallel corpus and learning lexical and syntactic information [8].

With the emergence of pre-trained model, researchers attempted to use pre-trained models (e.g., XLM [13] and XLM-R [14]) to implement machine translation quality estimation, which obtained fairly good results compared with previous research based on barely transformer. Those researches are both based on encoder framework, which consider QE as a regression task for matching HTER. However, As QE tasks and MT are highly related, QE models can also be implemented based on encoder-decoder framework. The QE model with encoder-decoder framework achieved the state-of-the-art performance in WMT 2017/2018 QE task [17] and mBART [4] based model achieved good results on DA (Direct Assessment) QE task [11]. It should be noted that previous methods usually neglected *pe* data in sentence-level QE task. In other words, information in *pe* data is unexploited. The only exception is in word level QE, which relies on *pe* to derive the quality label for each word.

3 PE Based Multi-Task Learning for Sentence Level QE

3.1 Multi-task Learning Framework for QE

Given that QE tasks is highly correlated with machine translation which is implemented by encoder-decoder architecture, we choose mBART [4] as our base model. mBART is based on multi-layers transformer architecture and utilizes the bidirectional modeling capability of the encoder while retaining the autoregressive feature. We feed the source text (*src*) into the encoder and the machine translation (MT) into the decoder, and the output of the decoder is used to implement the sentence level task and word level task, respectively.

The multi-task learning QE based on mBART is shown in Fig.1. For sentence-level task, we take the last token which is a special token <*eos*> to calculate the sentence-level loss, which we believe that the logit contains adequate information. We use sigmoid as the activation function. The loss function for sentence-level is as follows:

$$L_{sentence\_level} = MSE(HTER, sigmoid(FC(u))) \tag{1}$$

where *u* denotes the hidden representation for the special token <*eos*>. MSE represent Mean Square Error function, $L_{\text{sentence\_level}}$ denotes the sentence-level loss, *FC* denotes a fully connected layer.

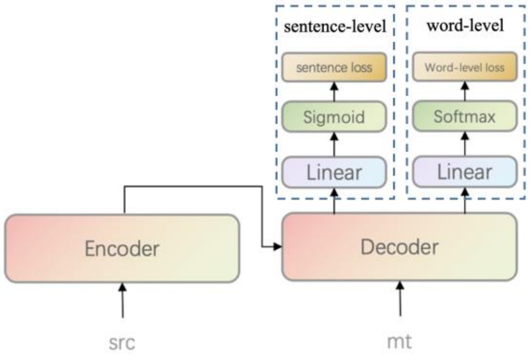

Fig. 1. Multi-task Learning Framework for MT QE

For word-level task (used as the baseline in this paper), we utilize each token's correlated logits to generate word-quality label. The loss function for word-level is as follows:

$$L_{\text{word\_level}} = \sum_{i=1}^{k}(-I(label = OK)\log(logit_i[0]) - I(label = BAD)\log(logit_i[1])) \tag{2}$$

The final overall loss is the sum of sentence-level loss and word-level loss，α is a constant weight.

$$L = L_{\text{sentence\_level}} + \alpha \times L_{\text{word\_level}} \tag{3}$$

3.2 *PE* based Multi-task Learning QE

Under the encoder-decoder structure of mBART, we design a translation task from *src* to *pe* as an auxiliary task for sentence-level QE. The model is shown in Fig.2.

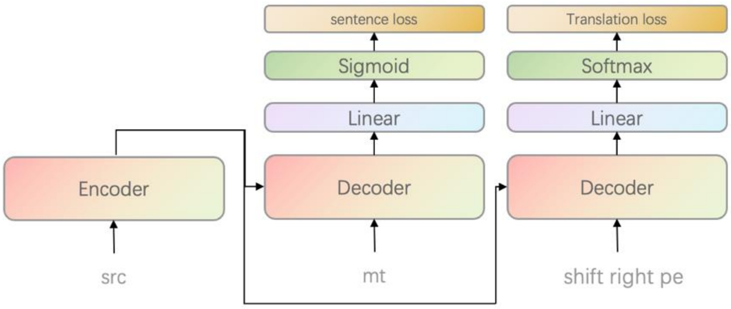

Fig. 2. sentence-level joint translation task

For the translation part, we feed the right-shifted *pe* $x = [x_1, \ldots . x_{\mathrm{k+1}}]$ into the decoder which share parameter with the sentence-level part.

The translation loss $L_{\mathrm{translation}}$ is calculated by the cross-entropy loss function:

$$L_{\mathrm{translation}} = \sum_{i=1}^{k} -\log\left(logit_i[x_{i+1}]\right) \tag{4}$$

where $x_{\mathrm{i+1}}$ denotes each token in the input sentence.

The final overall loss is the sum of sentence-level loss and translation loss， β is a constant weight.

$$L = L_{\mathrm{sentence\_level}} + \beta \times L_{\mathrm{translation}} \tag{5}$$

Compared with word-level task, translation task can evaluate not only the translation quality of each single word, but also the translation quality at the sentence-level by using the context information in the *pe* data. Meanwhile, compared with encoder-based QE structures, mBART can utilize *pe* data more directly and avoid additional label cost in word level quality annotation.

3.3 Multi-Model Ensemble

Given that various models with different initialized parameters, we can utilize multiple models to construct our system. Following existing practices in this aspect, we further implemented three other different QE models, mBERT, XLM-RoBERTa-base and XLM-RoBERTa-large to obtain different information from the same data. We average the HTER obtained by these three models and our system to generate stronger performance.

mBERT and XLM-RoBERTa are both encoder-based multilingual pre-trained models. The framework of QE is shown in the Fig.3. *src* and *mt* are concatenated as encoder input. The output of the encoder passes through the linear layer, which utilizes sigmoid as the activation function. For CCMT does not provide word level QE data, we didn’t apply multi-task learning for encoder-based framework.

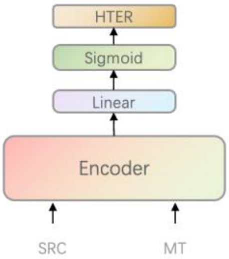

Fig. 3. sentence-level joint translation task

4 Experiments

4.1 Dataset

To compare with recent public results, we use the QE data from WMT2021 Machine Translation Quality Estimation tasks for English-German, and CCMT2021 Machine Translation Quality Estimation tasks for English-Chinese. Each dataset contains both sentence-level and word-level tasks. The dataset of WMT2021 provided 7k samples for training in both directions, and CCMT2021 provided more than ten thousand samples, slightly more data than WMT2021. The dataset statistics are shown in Table 1.

Table 1. The statistics of quality estimation datasets.

Dataset	Train	Dev	Test
WMT2021 EN-DE	7000	1000	1000
WMT2021 DE-EN	7000	1000	1000
CCMT2021 EN-ZH	10070	1385	1412
CCMT2021 ZH-EN	14789	1445	1528

4.2 Model Training and Evaluation Metric

In the training process, AdamW is selected as the optimizer. We set the batch-size as 8 and the learning rate is set to 1e-5, and the warmup steps are 1000 steps. The training adopts the early stop strategy, that is, if the model does not improve on the validation set in 2000 steps, stop training. The proposed approach is trained over a single Nvidia 3090. In the sentence-level translation quality estimation task, three evaluation metrics are used: Spearman's Rank Correlation Coefficient (Spearman), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The Spearman correlation coefficient is used as the main metric, in which the higher value indicates better performance of the QE model. The mean absolute error and the root mean square error are also provided for reference, in which the lower value indicates better performance of the QE model.

4.3 Experimental Results and Analysis

We first compare mBART with other pre-trained models on the WMT2021 Dataset. We choose monolingual BERT, XLM-Roberta, and mBERT as baselines. As shown in Table 2, the mBART model surpasses all the other pre-trained models and achieves the highest Pearson correlation in both DE-De and EN-ZH tasks.

Table 2. Experiment results with different pretrain models

	Model	Pearson↑	MAE↓	RMSE↓
EN-DE	BERT	0.544	0.122	0.172
	bert-base-multilingual	0.544	0.123	0.176
	XLM-RoBERTa-base	0.505	0.125	0.175
	XLM-RoBERTa-large	0.548	**0.116**	0.176
	mBART	**0.554**	0.125	**0.166**
EN-ZH	BERT	0.27	0.234	0.312
	bert-base-multilingual	0.265	0.278	0.314
	XLM-RoBERTa-base	0.256	**0.232**	0.282
	XLM-RoBERTa-large	0.30	0.233	**0.270**
	mBART	**0.327**	0.253	0.304

The experiment results of our system on WMT2021 are shown in Table 3. It shows that the multi-task learning method can achieve better results compared with using mBART only. For sentence-level QE, jointly trained with translation task obtained better performance than the single word-level task. However, combining word-level task and translation task will lead to a performance decline. We also compare the proposed QE model with the best results of WMT2021. HW-TSC [9] utilizes the auxiliary data for training which is obtained by a mature translation system. IST-Unbabel [10] uses the ADAPT strategy and a more complicated feature extraction classifier to enhance its performance. As a result, there is still a gap between our method and the best results.

The experiment results of our system on CCMT2021 are shown in Table 4. The proposed approach outperforms all the other pre-trained models in the CCMT2021 dataset. Jointly training with translation task boost the performance of our mBART-based system, and the ensemble of multiple models can also make improvement in both directions.

Table 3. Experiment results with multitask on WMT2021

	Model	Pearson↑	MAE↓	RMSE↓
EN-DE	WMT2021 baseline	0.529	0.129	0.183
	HW-TSC	**0.653**	**0.108**	**0.151**
	IST-Unbabel	0.617	0.116	0.172
	mBART	0.554	0.125	0.166
	mBART + word level	0.585	0.123	0.169

	mBART + translation	0.606	0.119	0.167
	mBART + translation + word	0.596	0.127	0.162
EN-ZH	WMT2021 baseline	0.282	0.246	0.287
	HW-TSC	**0.368**	0.248	0.297
	IST-Unbabel	0.290	**0.220**	0.266
	mBART	0.327	0.253	0.304
	mBART + word level	0.335	0.235	0.280
	mBART + translation	0.347	0.221	**0.265**
	+ translation + word	0.338	0.230	0.272

Table 4. Experiment results on CCMT2021

	Model	Pearson↑	MAE↓	RMSE↓
EN-ZH	mBART	0.348	0.085	0.125
	mBART + translation	0.375	0.089	0.118
	bert-base-multilingual	0.261	0.094	0.129
	XLM-RoBERTa-base	0.306	0.083	0.12
	XLM-RoBERTa-large	0.331	0.087	0.12
	ensemble	**0.419**	**0.079**	**0.114**
ZH-EN	mBART	0.483	0.078	0.113
	mBART + translation	0.498	0.0745	0.116
	bert-base-multilingual	0.422	0.091	0.119
	XLM-RoBERTa-base	0.414	0.077	0.117
	XLM-RoBERTa-large	0.463	0.076	0.117
	ensemble	**0.541**	**0.072**	**0.106**

4.4 Ablation Study

In this section, we will investigate the effect of translation task. We use *pe (post editing)* to correct the error of *mt (machine translation)* in different proportions, then the corrected *mt* is used as the input of decoder for the translation task. The result is shown in table 4. We observe that with the increase of the correction ratio, the performance of the model improves significantly. This means that when introducing *pe* into sentence-level evaluation system, the proposed approach can obtain more useful information from *pe* data.

Table 5. Effect of PE translation tasks

	Model	Pearson↑	MAE↓	RMSE↓
EN-DE	*Mt*	0.570	0.123	0.168
	20%	0.585	0.120	0.176

	40%	0.593	0.131	0.190
	60%	0.594	0.119	0.169
	80%	0.597	0.121	0.169
	100%	**0.606**	**0.119**	**0.167**
EN-ZH	*Mt*	0.332	0.254	0.304
	20%	0.339	0.255	0.305
	40%	0.337	0.238	0.282
	60%	0.343	0.240	0.289
	80%	0.346	0.234	0.276
	100%	**0.347**	**0.221**	**0.265**

We also test the influence of weight on multi-task learning as shown in Fig. 3 and 4. Generally speaking, the performance of the translation multi-task method is better than the word-level multi-task method.

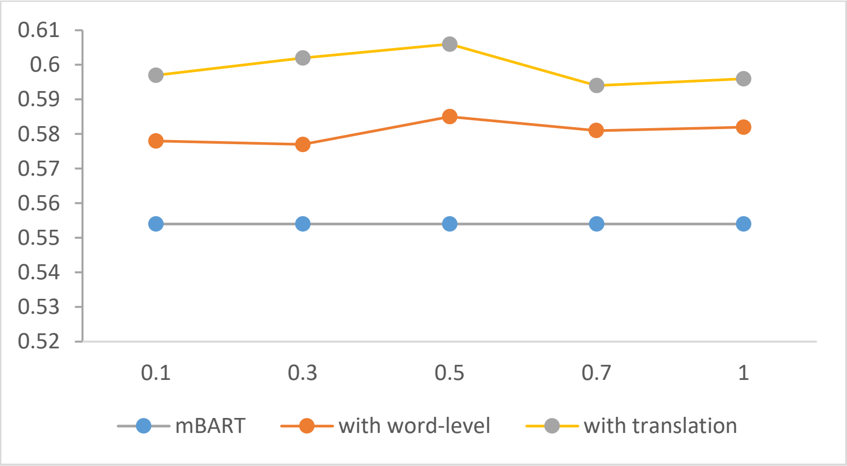

Fig. 3. Influence of joint training task weight on multi-task learning in EN-DE

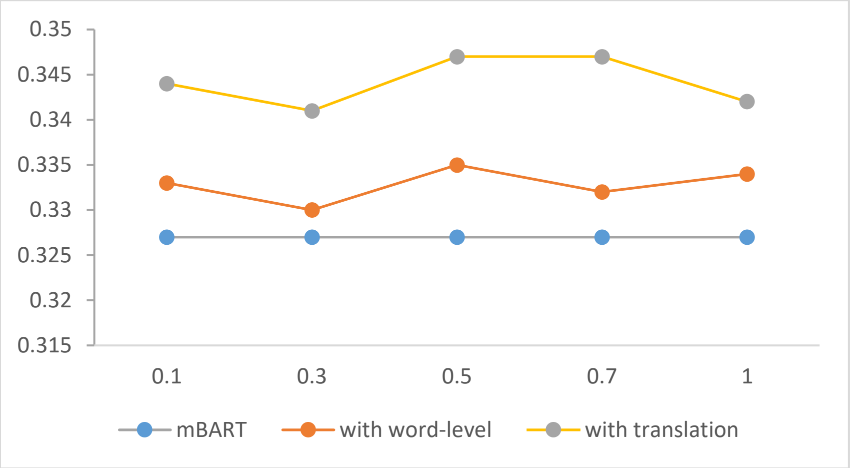

Fig. 4. Influence of joint training task weight on multi-task learning in EN-ZH

Moreover, we test different ways of input to train mBART like feed *mt* into the encoder and put *src* into the decoder or put *src* and *mt* into the encoder together, as shown in Table 6. Compared to other ways of input, our framework achieves significant improvements in EN-DE and EN-ZH tasks.

Table 6. Experiment results with different ways of input

	Method	Pearson↑	MAE↓	RMSE↓
EN-DE	Encoder: *src* Decoder: *mt*	**0.554**	0.125	0.166
	Encoder: *mt* Decoder: *src*	0.438	0.137	0.193
	Encoder: *src mt*	0.417	0.146	0.205
EN-ZH	Encoder: *src* Decoder: *mt*	**0.327**	0.253	0.304
	Encoder: *mt* Decoder: *src*	0.241	0.261	0.281
	Encoder: *src mt*	0.201	0.272	0.295

5 Conclusion

In this paper, we describe our submission in the QE task, which consists of English-Chinese and Chinese-English tasks. Our system is implemented based on the mBART and multi-task QE learning strategies. We propose a sentence-level translation quality estimation model based on the mBART, which achieves better results than other cross-language pre-training models. We also present a training method to introduce translation task into multi-task QE learning which successfully integrates post-edited sentences into sentence-level QE task and greatly improve the system performance with a simple model architecture design.

Acknowledgement

This work is partially funded by the National Key Research and Development Program of China (No. 2020AAA0108000), and by the Key Project of National Natural Science Foundation China (No. U1908216).

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. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019).
3. Specia L, Farzindar A. Estimating machine translation post-editing effort with HTER[C]//Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry. 2010: 33-43.
4. Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
5. Tang Y, Tran C, Li X, et al. Multilingual translation with extensible multilingual pretraining and finetuning[J]. arXiv preprint arXiv:2008.00401, 2020.
6. Kreutzer J, Schamoni S, Riezler S. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015: 316-322.
7. Kim H, Lee J H. A recurrent neural network approach for estimating the quality of machine translation output[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Techologies. Strouds-Burg, PA: ACL ,2016:494-498.
8. Fan K, Wang J, Li B, et al. “Bilingual Expert” Can Find Translation Errors[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6367-6374.
9. Shah K, Cohn T, Specia L. A bayesian non-linear method for feature selection in machine translation quality estimation[J]. Machine Translation, 2015, 29(2): 101-125.
10. Moura J, Vera M, van Stigt D, et al. Ist-unbabel participation in the wmt20 quality estimation shared task[C]//Proceedings of the Fifth Conference on Machine Translation. 2020: 1029-1036.
11. Zerva C, van Stigt D, Rei R, et al. Ist-unbabel 2021 submission for the quality estimation shared task[C]//Proceedings of the Sixth Conference on Machine Translation. 2021: 961-972.
12. González-Rubio J, Navarro-Cerdán J R, Casacuberta F. Dimensionality reduction methods for machine translation quality estimation[J]. Machine translation, 2013, 27(3-4): 281-301.
13. Kepler F, Trénous J, Treviso M, et al. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task[J]. arXiv preprint arXiv:1907.10352, 2019
14. Ranasinghe T, Orasan C, Mitkov R. TransQuest: Translation quality estimation with cross-lingual transformers[J]. arXiv preprint arXiv:2011.01536, 2020
15. Martins A F T, Astudillo R, Hokamp C, et al. Unbabel’s participation in the wmt16 word-level translation quality estimation shared task[C]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016: 806-811.
16. Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

17. Fan K, Wang J, Li B, et al. “Bilingual Expert” Can Find Translation Errors[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6367-6374.