Denoising Pre-training for Machine Translation Quality Estimation with Curriculum Learning

Xiang Geng[1], Yu Zhang[1], Jiahuan Li[1], Shujian Huang[1]*, Hao Yang[2], Shimin Tao[2], Yimeng Chen[2], Ning Xie[2], Jiajun Chen[1]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Huawei Translation Services Center, Beijing, China
{gx, zhangy, lijh}@smail.nju.edu.cn, {huangsj, chenjj}@nju.edu.cn,
{yanghao30, taoshimin, chenyimeng, nicolas.xie}@huawei.com

Abstract

Quality estimation (QE) aims to assess the quality of machine translations when reference translations are unavailable. QE plays a crucial role in many real-world applications of machine translation. Because labeled QE data are usually limited in scale, recent research, such as DirectQE, pre-trains QE models with pseudo QE data and obtains remarkable performance. However, there tends to be inevitable noise in the pseudo data, hindering models from learning QE accurately. Our study shows that the noise mainly comes from the differences between pseudo and real translation outputs. To handle this problem, we propose CLQE, a denoising pre-training framework for QE based on curriculum learning. More specifically, we propose to measure the degree of noise in the pseudo QE data with some metrics based on statistical or distributional features. With the guidance of these metrics, CLQE gradually pre-trains the QE model using data from cleaner to noisier. Experiments on various benchmarks reveal that CLQE outperforms DirectQE and other strong baselines. We also show that with our framework, pre-training converges faster than directly using the pseudo data. We make our CLQE code available (https://github.com/NJUNLP/njuqe).

Introduction

Machine translation (MT) quality estimation (QE) is the task of estimating the quality of machine translations when reference translations are unavailable (Specia, Scarton, and Paetzold 2018). As shown in Table 1, QE focuses on predicting sentence-level scores and word-level tags given the sources and MT outputs. QE is of great practical use in real-world scenarios including reducing post-editing effort by filtering out low-quality translation results or pointing out possibly low-quality tokens (Specia 2011), improving the performance of MT systems by guiding the decoding process (Wang et al. 2020a), etc.

Despite the usefulness of QE, the collection and annotation of QE data are expensive, which hinders the performance of neural QE models (Kepler et al. 2019). Thus, the pre-training and fine-tuning strategy have been widely used to transfer bilingual knowledge from parallel corpora to the QE task. The predictor-estimator framework (Kim et al. 2017; Fan et al. 2019) pre-trains a feature extractor called predictor, which predicts each word in the reference given the source sentence and the target context.

Source	the exhibit provided new homes for the zoo 's siamangs and pygmy marmosets .
MT	**这个 (this)** 展览 (exhibit) 为 (for) 动物园 (zoo) 的 ('s) **壁炉 (fireplace)** **和 (and)** **俾格米 (pygmy)** **马 (horse)** **赛车 (car racing)** 提供了 (provided) 新的 (new) 家园 (homes) **.** ‖ HTER = 0.4667

Table 1: An example from the WMT20 English-Chinese (EN-ZH) QE dataset. For word-level tags, each token is labeled as 'OK' or 'BAD' (we show the 'BAD' tokens in bold font with an underline), representing whether it needs to be corrected. Sentence level score HTER (Snover et al. 2006) measures the whole effort to correct the MT manually.

Cui et al. (2021) argued that the differences between the predictor task and the QE task might cause problems for bilingual knowledge transfer. As a solution, they proposed the DirectQE framework, which generates pseudo QE data (pseudo MT outputs and their QE labels) and uses these data to pre-train whole QE models directly. To generate pseudo MT outputs with controllable errors, DirectQE randomly replaces target tokens of parallel sentences. These replaced tokens are assigned with QE label 'BAD' and others with 'OK'.

While DirectQE achieves remarkable performance, we notice that the generated pseudo data still have some noises. Although the pseudo labels are accurate, these pseudo MT outputs sometimes are quite different from real ones. To be specific, there could be more meaningless pseudo translation errors that a well-trained MT model may hardly generate. Besides, the characteristics, such as length distribution or domain, of pseudo MT outputs are different from real MTs. From the transfer learning perspective, the difference between the pre-trained distribution and the target distribution could result in performance degradation (negative transfer) (Tan et al. 2017).

To handle this problem, we propose CLQE, a denoising pre-training framework for QE based on curriculum learning

*Corresponding Author.

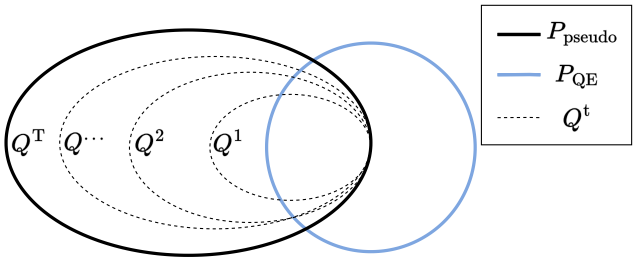


Figure 1: CLQE gradually pre-trains model from the cleaner subset Q^1 to the noisier whole pseudo set P_{pseudo}, where P_{QE} is the target distribution we want to learn.

(Bengio et al. 2009). Firstly, we define several metrics for measuring how noisy the pseudo QE data are. More specifically, we use two model-free metrics to measure noise using statistical features, i.e., sentence length and word rarity, and other three metrics to estimate the distribution of real QE data using surrogate MT models.

Then, we introduce a competence-based curriculum to mitigate the adverse effects of the noisy pseudo data. As shown in Figure 1, CLQE starts pre-training from cleaner data, whose sample distribution Q^1 is determined by noise metrics and expected to be closer to the real one P_{QE}. The curriculum gradually exposes the model to noisier data as the model becomes more competent until the whole pre-training dataset P_{pseudo}. In this way, CLQE smoothly assigns larger weights for cleaner data and could minimize the upper bound of the expected risk on QE task theoretically (Gong et al. 2016).

We summarize our contributions: (1) We demonstrate that the pseudo QE data are noisy, due to the differences between pseudo and real MT outputs. (2) We propose CLQE, a denoising framework for pre-training on pseudo QE data, which measures the noise and helps the model learn better on the noisy data. (3) We achieve new SOTA on different QE datasets, and the analysis further reveals the superiority of CLQE.

Preliminaries

Machine Translation

We denote (X, Y) as a parallel pair, where X is the source sentence, and Y is the target sentence with the same meaning. MT model aims to learn a mapping function $f(Y|X;\theta)$ parameterized by θ using parallel data. Given a set of observed parallel pairs $\mathcal{D} = \{X^i, Y^i\}_{i=1}^N$, a standard training objective is to maximize the log-likelihood:

$$J_{\text{mt}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \log P(Y|X;\theta). \quad (1)$$

After training, the translation can be decoded as $\hat{Y} = \arg\max P(Y|X;\theta)$. In most studies, researchers evaluate the quality of the translation $\hat{Y}$ by comparing it to annotated references. However, the references are unavailable in most applications of MT systems.

Quality Estimation

Machine translation quality estimation assesses the translation quality of an MT system without access to reference translations. The quality can be evaluated in different grains, and we focus on the sentence level and the word level in this paper. Given a source sentence X and a machine translation $\hat{Y} = \{y_1, y_2, \ldots, y_{|\hat{Y}|}\}$ with $|\hat{Y}|$ words, the fine-grained word level labels is a sequence of $|\hat{Y}|$ tags $G = \{g_1, g_2, \ldots, g_{|\hat{Y}|}\}$, the tag g_j is usually a binary label ('OK' or 'BAD') representing whether the word y_j need to be corrected. The sentence-level score h is usually a real number in $[0, 1]$ representing the quality of the whole translation. Human-targeted Translation Edit Rate (HTER), calculated by the percentage of edits required to fix for $\hat{Y}$, has been widely used as the sentence-level score. Formally, we can organize QE dataset as $\mathcal{Q} = \{s^i\}_{i=1}^n$, where each instance $s^i = (X^i, \hat{Y}^i, G^i, h^i)$. In most translation directions, the real QE dataset is scarce and much smaller than the parallel dataset. Thus, a series of researches designed different pre-training methods to exploit the bilingual knowledge in parallel data.

DirectQE Framework

DirectQE proposes to construct pseudo QE task on the bilingual corpus to facilitate more direct knowledge transfer. Concretely, DirectQE first generates pseudo QE data using parallel data and then pre-trains model on these pseudo QE Data with QE objective.

Generating Pseudo QE Data. DirectQE trains a masked language model (Devlin et al. 2019) conditioned on source sentences as the pseudo data generator. Given parallel data, they randomly mask some percentage of the target tokens. The generator is trained to predict those masked tokens given source tokens and the rest of the target tokens.

The generating process is similar to the training process. Given a parallel pair (X, Y), DirectQE feeds the X and the masked Y into the trained generator. Then, DirectQE replaces each masked token with the token sampled from the output probability distribution of the generator. The generated sentence is regarded as pseudo MT, denoted as $\tilde{Y}$. To sample truly negative tokens efficiently, DirectQE randomly selects the tokens from those with the top k generation probability. Empirically, DirectQE sets $k = 10$. Therefore, to generate pseudo tags $\tilde{G}$, every changed token of Y is annotated as 'BAD', and the others are annotated as 'OK'. The pseudo sentence-level scores $\tilde{h}$ are further defined as the ratio of 'BAD' tokens. In this way, DirectQE can generate a large amount of pseudo QE data $\tilde{\mathcal{Q}} = \{\tilde{s}^i\}_{i=1}^N$ using parallel dataset $\mathcal{D}$, where $\tilde{s}^i = (X^i, \tilde{Y}^i, \tilde{G}^i, \tilde{h}^i)$.

Pre-training and Fine-tuning. DirectQE pre-trains the QE model with QE objective on pseudo QE dataset $\tilde{\mathcal{Q}}$ instead of a surrogate objective. DirectQE jointly learns sentence-level objective J_{sen} and word-level objective J_{word} as a multi-task problem:

$$J_{\text{qe}} = J_{\text{sen}} + J_{\text{word}}. \quad (2)$$

J_{sen} and J_{word} can be regarded as a regression problem and a sequence labeling problem, respectively:

$$J_{\text{sen}} = \log P(\tilde{h}|X, \tilde{Y}; \theta), \quad (3)$$

$$J_{\text{word}} = \log P(\tilde{G}|X, \tilde{Y}; \theta). \quad (4)$$

When pre-training and fine-tuning, the training instances in Eq. 3 and Eq. 4 are sampled from pseudo QE dataset $\tilde{\mathcal{Q}}$ and real QE dataset $\mathcal{Q}$ respectively. Following Cui et al. (2021), the pre-trained model is selected with the pseudo validation set for further fine-tuning.

CLQE framework

In this section, we start with the analysis for understanding the noise in pseudo data, then define several noise metrics according to the analysis and finally introduce how to control the schedule of pseudo data for denoising pre-training with curriculum learning.

The Noise[1]

The noise of pseudo data may have two sources: noisy labels in pseudo QE data; the differences between pseudo translations and real MTs.

The noisy label problem in pseudo QE data generated by DirectQE is mainly manifested as false negatives: a correct replaced word is labeled as 'BAD'. However, this rarely happens due to the negative sampling strategy of DirectQE. Empirically, we manually check 20 random pseudo QE data, and the word label accuracy is 0.972 (95% confidence interval 0.957-0.987). Therefore, this study focuses on pseudo translations instead of pseudo labels.

We notice that there are the following aspects that pseudo MTs could be different from real MTs: (1) While the outputs of most advanced MT models are autoregressive fashion[2], the DirectQE framework is fully non-autoregressive. Thus, these pseudo MTs involve more errors due to the lack of dependency between masked tokens, such as meaningless repeats. The frequency of repeat in pseudo MTs is significantly higher than that of real MTs (0.560 vs. 0.012 consecutive repeated tokens per sentence). Besides, to sample truly 'BAD' tokens, low probability tokens are preferred by designed sample strategies. As a result of these reasons, the pseudo MTs are less fluent than real MTs (perplexity 135.21 vs. 66.61 calculated by GPT-2 (Radford et al. 2019)). (2) Most parallel corpus is collected from the news domain, but the QE data are from the Wikipedia domain. This leads to the pseudo data generated from parallel corpus differs from real QE data in terms of statistical features, such as sentence lengths. We plot the cumulative density function of lengths for pseudo/real QE data in Figure 2. We can see that the length distribution of pseudo MTs is also different from that of real MTs.

[1]These analyses are conducted on the WMT20 EN-ZH QE task.

[2]Note that most non-autoregressive models use translations produced by autoregressive models for knowledge distillation (Bao et al. 2021).

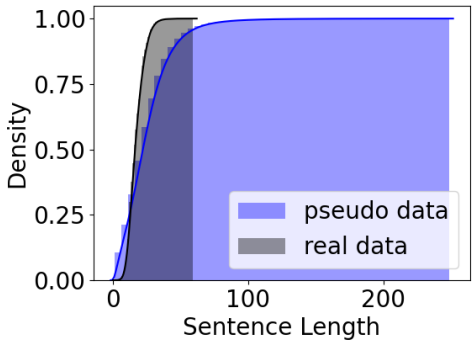


Figure 2: The cumulative density function of lengths for pseudo/real data on the WMT20 EN-ZH task.

Noise Metrics[3]

Inspired by statistical QE methods (Specia et al. 2013), and curriculum learning methods for MT (Platanios et al. 2019), we propose two model-free noise metrics based on statistical features of source sentences as follows.

Sentence length. The pseudo MTs of longer source sentences could be noisier. Since more tokens need to be modified by the generator, these pseudo MTs are more likely to involve irrelevant errors. Besides, the real MTs of longer source sentences could contain complex translation errors, e.g., improper sentence structures, which are difficult to simulate by the generator. Formally, we can define the noise of a pseudo data as $d_{\text{length}} = |X|$.

Word rarity. Similar to long sentences, generating proper pseudo MTs for source sentences with more rare words is difficult. Moreover, rare words have fewer related candidates for replacement. For source sentence $X = \{x_1, x_2, \dots x_{|X|}\}$, we calculate the word rarity as follows:

$$d_{\text{rarity}} = -\sum_{i=1}^{|X|} \log \hat{P}(x_i), \quad (5)$$

where $\hat{P}(x_i)$ denote the frequency of the word x_i given the source corpus of the parallel dataset $\mathcal{D}$.

Modeling the translation distribution is a more direct way to measure the difference between pseudo and real MTs. However, the target MT model to be estimated is unavailable in the black-box setting. Thus, we propose the following alternatives.

Generation probability of a surrogate MT model. Using parallel data $\mathcal{D}$, we can train a surrogate model $\tilde{\theta}_{\text{mt}}$ with Eq. 1. We define the noise of pseudo MT $\tilde{Y}$ as the negative of the generation probability parameterized by $\tilde{\theta}_{\text{mt}}$:

$$d_{\text{prob}} = -\log P(\tilde{Y}|X; \tilde{\theta}_{\text{mt}}). \quad (6)$$

[3]Note that negation is used in metrics based on probability because a low probability denotes a higher noise score.

We assume that this generation probability could be similar to the one of the target model. Moreover, it can detect translation errors due to the lack of dependency between masked tokens.

Generation probability of a fine-tuned MT model. In MT domain adaptation, a common strategy is fine-tuning a general domain MT model on a small in-domain corpus (van der Wees, Bisazza, and Monz 2017). Inspired by this idea, we fine-tune the surrogate model $\tilde{\theta}_{\text{mt}}$ using the sources $X \in \mathcal{Q}$ as input and the MTs $\hat{Y} \in \mathcal{Q}$ as the ground truth. It can also be seen as the knowledge distillation strategy for non-autoregressive translations (Bao et al. 2021). Please note that we aim to simulate the behavior of the target MT model rather than improve the translation performance. Here, we define the noise metric based on fine-tuned model $\hat{\theta}_{\text{mt}}$:

$$d_{\text{qe prob}} = -\log P(\tilde{Y}|X; \hat{\theta}_{\text{mt}}). \tag{7}$$

Cross entropy difference. Cross entropy difference (CED) (Axelrod, He, and Gao 2011) is used to measure domain relevance of a parallel pair in Wang, Caswell, and Chelba (2019). Following Wang, Caswell, and Chelba (2019), we further propose the noise metric with CED:

$$d_{\text{ced}} = d_{\text{qe prob}} - d_{\text{prob}}. \tag{8}$$

The CED could detect these pseudo MTs, which are more likely generated by the target model rather than the surrogate model.

Denoising Curriculum

The basic idea of the proposed denoising curriculum is that QE models are gradually pre-trained from cleaner pseudo data to noisier pseudo data. To this end, we design Algorithm 1 following the competence-based curriculum learning (Platanios et al. 2019). Specifically, we normalize the noise score as $\hat{d}(s^i) \in [0, 1]$ with cumulative density function (CDF). We assume that the model competence $c(t) \in (0, 1]$ increases with pre-training as follows:

$$c_{\text{linear}}(t) = \min(1, t\frac{1 - c_0}{T} + c_0), \tag{9}$$

where T denotes the length of the curriculum, c_0 is the initial competence without training. With the increase of competence, noisier data that satisfies $\hat{d}(s^i) < c(t)$ will be continually added into the pre-training. For stable and efficient pre-training, we organize the training batch according to sentence length as recommended in most MT implements.

In preliminary experiments, we also test CLQE with the square root competence function c_{sqrt} in Platanios et al. (2019). We find that c_{linear} works better than c_{sqrt}. The reason may be that the competence of c_{sqrt} increases too fast while the model has not handled the given data yet (considering the pre-training QE task is hard).

Algorithm 1: Denoising pre-training for machine translation quality estimation.

Input:Pseudo QE dataset $\tilde{\mathcal{Q}}$, noise scoring function d, competence function c.
Output:Pre-trained model θ.

```
Compute the noise score d(s^i) for each sample s^i = (X^i, Ỹ^i, G̃^i, h̃^i) ∈ Q̃
Normalize noise scores as d̂(s^i) ∈ [0, 1] with cumulative density function (CDF)
for t = 1 ... do
    Compute model competence c(t)
    Select a pre-training subset Q̃^t = {s^i | d̂(s^i) < c(t), s^i ∈ Q̃}
    Shuffle Q̃^t, and then sort Q̃^t by sentence lengths
    while Q̃^t ≠ ∅ do
        Pop batch B from sorted Q̃^t
        Update θ with Eq. 2 on batch B
    end while
end for
```

Experiments

Setting Description

Datasets. We employ WMT19 and WMT20/WMT21[4] QE dataset for English-German (EN-DE) and English-Chinese (EN-ZH) direction respectively. The size of training, development, and test sets are 13K/1K/1K, 7K/1K/1K, and 8K/1K/1K for WMT19, 20, and 21 QE tasks, respectively. For parallel data, we randomly sample about 3M parallel pairs provided by WMT QE Shared Task for each translation direction, which are much larger than QE dataset.

Models. Following Cui et al. (2021), we implement the DirectQE generator using a small transformer (Vaswani et al. 2017), whose encoder and decoder both have 6 layers with 256 hidden neurons. To achieve strong baselines, we use the XLM-R large model (Conneau et al. 2020), a pre-trained cross-lingual sentence encoder, as the QE model. We input both sources and pseudo/real MTs by concatenating them. The representation of the beginning-of-sentence token $\langle s \rangle$ is used to predict the HTER score. We average the representations of sub-tokens for predicting the tag of the whole word. We use the standard transformer base setting as in (Vaswani et al. 2017) for surrogate MT models. EN-ZH task is the glass-box setting so that we directly use the target MT model instead of the fine-tuned surrogate model to score the noise $d_{\text{qe prob}}$. We learn the BPE vocabulary (Sennrich, Haddow, and Birch 2016) with 30K steps using parallel data for the generator and surrogate MT models.

Implementation and reproducibility. Our implementation is built on the open source toolkit Fairseq(-py) (Ott et al. 2019). We provide our implementation online.[5] All experiments are performed on NVIDIA V100 GPUs. We set the initial competence $c_0 = 0.05$ and total duration of curriculum learning $T = 5$ epochs. Other details can be found in supplementary materials.

Evaluation metrics. For sentence-level QE, the performance is measured by Pearson correlation coefficient (the

[4]https://www.statmt.org/wmt##, ## can be 19, 20, 21.

[5]https://github.com/NJUNLP/njuqe

Dataset	Method	Sent-level Test			Word-level Test	
		Pearson↑	MAE↓	RMSE↓	MCC↑	F1-MULT↑
WMT19 EN-DE	DirectQE(Cui et al. 2021)	55.08	11.25	16.33	-	39.71
	NMT+TER	57.11	10.86	17.04	46.48	49.04
	CBSQE	57.52	10.69	16.14	46.53	49.80
	Rarity	57.84	10.77	16.92	47.20	49.99
	Prob	58.40	10.47	16.24	47.08	50.25
	QE Prob	58.72	10.63	**16.11**	47.19	50.43
	CED	**59.98**	**10.42**	16.15	**47.75**	**50.98**
WMT20 EN-ZH	DirectQE(Cui et al. 2021)	60.83	-	-	-	53.06
	DirectQE (ours)	64.82	12.77	16.02	50.05	55.69
	Length	64.62	12.70	16.06	50.45	55.99
	Rarity	65.75	12.64	15.95	50.29	55.75
	Prob	65.06	12.66	15.91	50.57	56.07
	QE Prob *	**66.34**	**12.46**	**15.74**	**51.72**	**57.10**
	CED *	65.96	12.61	15.89	50.94	56.31
WMT21 EN-ZH	DirectQE (ours)	32.12	24.99	30.49	30.23	37.05
	Length	32.72	24.53	29.60	30.39	37.19
	Rarity	32.32	24.53	30.62	30.46	37.25
	Prob	33.41	23.95	28.30	30.87	37.86
	QE Prob *	**33.91**	**22.66**	**27.74**	**31.76**	**38.24**
	CED *	33.13	22.77	28.50	30.98	38.04

Table 2: Main results on different QE datasets. * indicates that $d_{\text{qe prob}}$ is calculated by the target MT model.

primary metric for WMT19, 20, and 21), mean absolute error (MAE), and root mean square error (RMSE). For word-level QE, the performance is measured by F1-MULT (the primary metric for WMT19) and Matthews correlation coefficient (MCC, the primary metric for WMT20 and 21). F1-MULT is the multiplication of F1-scores for the 'OK' and 'BAD' words.

Method	Pearson↑	F1-MULT↑
QE-BERT	52.60	40.61
SOURCE	54.74	-
UNBABEL	57.18	47.52
DirectQE	57.25	-
Our Single	**59.98**	**50.98**

Table 3: Ensemble results on WMT19 EN-DE dataset. The results are collected from their original papers (Kim et al. 2019; Zhou, Zhang, and Hu 2019; Kepler et al. 2019; Cui et al. 2021).

Method	Pearson↑	MCC↑	MCC (w/ gap)↑
IST-Unbabel	65.1	38.2	57.5
HW-TSC	-	42.8	58.7
NICT	-	44.9	58.2
Tencent	66.4	-	-
NiuTrans	67.5	48.4	61.0
Our Ensemble	**67.86**	**52.87**	**62.85**

Table 4: Ensemble results on WMT20 EN-ZH dataset. The results are collected from their original papers (Moura et al. 2020; Wang et al. 2020b; Rubino 2020; Wu et al. 2020; Hu et al. 2020).

Main Results

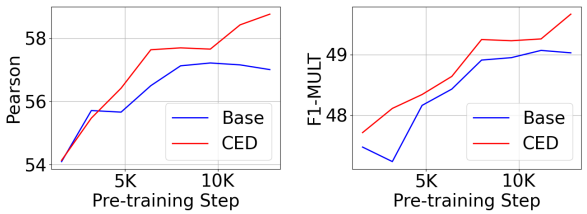
Table 2 shows the results on the WMT19 EN-DE and WMT20/21 EN-ZH QE tasks. The DirectQE we reproduce outperforms the vanilla DirectQE (Cui et al. 2021) by a large margin because we introduce extra cross-lingual knowledge from XLM-R and use larger models. Compared to the strong baseline, the proposed CLQE still demonstrates superior performance, whichever noise metrics are employed. Two model-free noise metrics slightly increase the performance, suggesting they are good options when there are limited computational resources. On the WMT19 EN-DE QE task, which is the black-box setting, the QE Prob metric outperforms the Prob metric, implying that fine-tuning helps imitate the target MT model. CED achieves the best result on the WMT19 EN-DE QE task and increases the Pearson by **+2.87** and the F1-MULT by **+1.94**. However, fine-tuning failed on the EN-ZH QE tasks, which are the glass-box setting. Fine-tuning the surrogate model on the QE training set does not increase the performance on the QE development set. That may be because the surrogate model is trained on the subset of the parallel corpus provided by WMT and used for training the target MT model. QE Prob calculated by the target MT model achieves the best result on EN-ZH tasks. QE Prob increases the Pearson by **+1.52/+1.79** and the MCC by **+1.67/+1.53** on WMT20 and WMT21 tasks, respectively.

Ensemble Results

The ensemble method has been widely used for WMT QE shared tasks. We also report our ensemble results in Table 3 and Table 4. Since WMT21 focused on the multilingual setting, WMT21 submissions are not comparable. However, we still provide our ensemble result of WMT21 in supplementary materials for reference. Following Kepler et al. (2019), we learn ensemble weights of different models using the de-

Metric	Accuracy↑
Length	0.569
Rarity	0.568
Prob	0.795
QE Prob	0.948
CED	0.963

Table 5: Accuracy of discriminating the pseudo and real QE data using different noise metrics.



(a) Pearson score. (b) F1 MULT.

Figure 3: Pre-training step vs. Pearson score (a) / F1 MULT (b) of different models on the real WMT19 EN-DE QE test set.

velopment set for sentence-level ensemble. For word-level ensemble, the tag with the most votes by different models for each word is regarded as the output. We only use the models listed in Table 2 for our ensemble systems. As shown in Table 3, our single model achieves better performance than other ensemble systems. That confirms we build a strong benchmark for the single model. In Table 4, we observe that our system outperforms the best results of the WMT20 EN-ZH QE shared task on both sentence and word level[6]. Please note that these systems use extra glass-box features for predicting. The ensemble theory in Krogh and Vedelsby (1995) points out that the diversity of models is related to the performance. Thus, the competitive ensemble performance implies that different noise metrics result in different curricula.

Analysis

In this section, we further investigate the factors contributing to the improvements and how the proposed method affects the pre-training.

Impact of Noise Metrics

As shown in Table 2, different noise metrics result in various performances. We are curious about how these metrics affect performance. Assuming that real QE data should be cleaner than pseudo data. Thus, better noise metrics should better discriminate between real and pseudo data. We design a discrimination task to measure the ability to detect noise using different metrics. Specifically, we randomly sample a pseudo instance for each instance in the WMT19 EN-DE QE test set, and the metrics are expected to assign a lower score for a real instance. We summary the discrimination accuracy in Table 5. Length and Rarity are slightly better than randomly guessing, while QE Prob and CED obtain significant accuracy under the black-box setting. Interestingly, we observe that the ranks of QE performance in Table 2 are similar to the ranks of noise detection performance in Table 5. The CDF of different noise metrics for pseudo/real data in supplementary materials also shows that noise metrics can clearly discriminate pseudo data from real data.

[6]WMT20 calculated the performance of word tags and gap tags together. Gap tags denote whether there are missing words between every two words. Thus, different from DirectQE (Cui et al. 2021), we also report the results with gap tags. We concatenate the representations of every two words for predicting the gap tags.

Method	Subset	Pearson↑	F1-MULT↑
w/o curriculum	25%	56.63	48.82
	50%	57.42	49.09
	75%	58.52	49.94
	100%	57.11	49.04
reverse curriculum	100%	58.08	49.64
denoising curriculum	100%	59.82	51.01

Table 6: Comparison of denoising curriculum, reverse curriculum, and data filtering method on WMT19 EN-DE QE dataset using CED metric.

Impact of Denoising Curriculum

We carry out experiments on the WMT19 EN-DE task to prove that denoising curriculum helps the QE model better exploit the pseudo data. Specifically, we compare the denoising curriculum with the data filtering methods and reverse curriculum. For the data filtering method, we sort the pseudo data from cleaner to nosier according to d_{ced}, and then select the top 25%/50%/75% least noisy pseudo data. We pre-train QE models on these subsets without curriculum learning. For the reverse curriculum, the pseudo data are presented to QE models from nosier to cleaner. The results are summarized in Table 6. With only 50% data, we achieve a similar performance of the model pre-trained on all pseudo data. This confirms that some pseudo data are harmful to the downstream QE task. Although the data filtering model with top 75% cleaner data outperforms the baseline, it is still much lower than CLQE. That is because these filtered pseudo data may contain some useful information. For example, only part of pseudo MT are noisy. The idea of CLQE is to reduce these noisy data's negative impacts instead of ignoring them. The curriculum also achieves better results than the reverse curriculum. Intuitively, curriculum improves performance when the dataset is challenging or noisy, while the reverse curriculum works on the clean dataset (Chang, Learned-Miller, and McCallum 2017). From the perspective of the optimization problem, if we start the pre-training with cleaner data, we may reach an excellent initial point which is essential for better generalization ability (Bengio et al. 2009). Theoretically, Gong et al. (2016) show that starting pre-training from cleaner examples guides the learning toward the expected target distribution.

Source 1	Once introduced , these changes should contribute to improve the capacity of the High Commissioner to discharge her mandate .
Pseudo MT 1	**如果 (if) 改变 (change)** 一经 (once) **开始 (start)** , **以后 (after) 使 (make)** 高级专员 (High Commissioner) **履行 (fulfill)** 完成 (discharge) **家园 (home)** 的 (of) 能力 (capacity) **进行 (to) 促进作用 (enhancement)** 。(.) ‖ $d_{\text{qe prob}}$ = 6.93
Source 2	They should be clearly separated from humanitarian centres established , for example , for returning refugees .
Pseudo MT 2	**这些 (this)** 中心 (centres) 应 (should) 与 (from) 诸如 (for example) 为 (for) **返回 (return)** **家园 (home) 而 (while)** 建立的 (established) 人道主义 (humanitarian) 中心 (centres) **明确地 (definitely)** 分开 (separated) 。(.) ‖ $d_{\text{qe prob}}$ = 2.56

Table 7: Pseudo QE data for WMT20 EN-ZH QE task.

Convergence Speed

Lots of research report curriculum learning makes training converge faster (Platanios et al. 2019). As discussed before, real data are different from pseudo data. Thus, the pseudo validation set could not reflect the convergence speed on real data. To examine this property for CLQE, we save the pre-training checkpoint every 1.6K steps and fine-tune these checkpoints on real QE data. The sentence- and word-level results on the real WMT19 EN-DE QE test set are shown in Figure 3(a) and 3(b). It can be seen that the proposed CED convergences are significantly faster on the real QE test set than the baseline. Besides, we notice that CED achieves similar performance in Table 2 with only 13K steps (compared to 40K steps). Thus, monitoring the pre-training on real QE data could be better, though this strategy requires more computational and storage resources. We also provide convergence speed analysis on the WMT20 EN-ZH dataset in the appendix.

Low-Resource

Intuitively, CLQE will work better on noisier data. Fewer parallel data results in noisier pseudo data since the generator trained with fewer data will be weaker. To confirm that, we have performed experiments on the WMT19 EN-DE dataset using CED scores with only 100K parallel data. Results show that CLQE also outperforms DirectQE in the low-resource setting (CED vs. DirectQE: Pearson 54.58 vs. 53.01, F1-MULT 48.01 vs. 47.24).

A Case Study

We still take the instance in Table 1 as an example. For this instance, DirectQE fails to predict 'homes' as 'OK', while QE Prob does it right. To explain the possible reason, we list some related pseudo data in Table 7. Intuitively, the ridiculous translation error 'discharge home' in pseudo MT 1 is useless for QE and increases the risk of harmful inductive bias. 'return home' in pseudo MT 2 seems to simulate the translation error generated by MT model. Our method successfully detects the noise with $d_{\text{qe prob}}$ and reduces the negative impacts of pseudo MT 1 using curriculum learning.

Related Works

Although curriculum learning has been widely used for many NLP tasks such as machine translation (Platanios et al. 2019), sequence refinement (Agrawal and Carpuat 2022), the NLP community has not yet investigated its application to QE. Traditional QE methods (Specia et al. 2013), and curriculum learning methods for MT (Platanios et al. 2019) use sentence length and word rarity of the source sentence to quantify the complexity of translating it. In this research, we show that they also reflect the complexity of generating pseudo MTs and successfully use them for measuring the noise in pseudo data. Recently, Fomicheva et al. (2020) regarded the generation probabilities of real MTs as the uncertainty feature and directly used it to estimate their quality. Instead, we use the generation probabilities of pseudo MTs for measuring the difference from real MTs and use this feature for guiding the pre-training progress. Wang, Caswell, and Chelba (2019) used CED for domain adaption, where the surrogate MT model was fine-tuned on a small in-domain parallel data. Our fine-tuning is closer to the knowledge distillation using the target MT model as a teacher.

Conclusion

Due to the scarcity of QE data, pre-training with pseudo QE data has become increasingly important. In this study, we highlight that pseudo QE data are noisy and discuss the noise source. We present a novel framework called CLQE for denoising pre-training with pseudo QE data. We define how to measure noise and organize the presenting order of pseudo QE data from cleaner to noisier. Experiments and analyses demonstrate the effectiveness of our method.

There are several potential directions for bridging the gap between pseudo and real MTs. For example, we can use reference-based automatic evaluation methods to annotate the MTs generated by the surrogate MT model so that there will be little difference between pseudo and real translations. Considering the performance of automatic evaluation methods is imperfect, the pseudo labels may not be trustworthy in this way. Another possible way is to generate real-like translations while maintaining negative sampling. This method could significantly improve the generation cost since it involves a complex decoding process. We leave these studies to future research.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding

author. This work is supported by National Science Foundation of China (No. 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02).

References

Agrawal, S.; and Carpuat, M. 2022. An Imitation Learning Curriculum for Text Editing with Non-Autoregressive Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7550–7563. Dublin, Ireland: Association for Computational Linguistics.

Axelrod, A.; He, X.; and Gao, J. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 355–362. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Bao, Y.; Huang, S.; Xiao, T.; Wang, D.; Dai, X.; and Chen, J. 2021. Non-Autoregressive Translation by Learning Target Categorical Codes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5749–5759. Online: Association for Computational Linguistics.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chang, H.-S.; Learned-Miller, E.; and McCallum, A. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Cui, Q.; Huang, S.; Li, J.; Geng, X.; Zheng, Z.; Huang, G.; and Chen, J. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12719–12727.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fan, K.; Wang, J.; Li, B.; Zhou, F.; Chen, B.; and Si, L. 2019. "Bilingual Expert" Can Find Translation Errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6367–6374.

Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; and Specia, L. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555.

Gong, T.; Zhao, Q.; Meng, D.; and Xu, Z. 2016. Why curriculum learning & self-paced learning work in big/noisy data: A theoretical perspective. *Big Data & Information Analytics*, 1(1): 111.

Hu, C.; Liu, H.; Feng, K.; Xu, C.; Xu, N.; Zhou, Z.; Yan, S.; Luo, Y.; Wang, C.; Meng, X.; Xiao, T.; and Zhu, J. 2020. The NiuTrans System for the WMT20 Quality Estimation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, 1018–1023. Online: Association for Computational Linguistics.

Kepler, F.; Trénous, J.; Treviso, M.; Vera, M.; Góis, A.; Farajian, M. A.; Lopes, A. V.; and Martins, A. F. T. 2019. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In *Proceedings of the Fourth Conference on Machine Translation*.

Kim, H.; Jung, H.-Y.; Kwon, H.; Lee, J.-H.; and Na, S.-H. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1): 1–22.

Kim, H.; Lim, J.-H.; Kim, H.; and Na, S.-H. 2019. QE BERT: Bilingual BERT Using Multi-task Learning for Neural Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*.

Krogh, A.; and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems*.

Moura, J.; Vera, M.; van Stigt, D.; Kepler, F.; and Martins, A. F. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, 1029–1036.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Platanios, E. A.; Stretcu, O.; Neubig, G.; Poczos, B.; and Mitchell, T. 2019. Competence-based Curriculum Learning for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1162–1172. Minneapolis, Minnesota: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *Technical report, OpenAI.*

Rubino, R. 2020. NICT Kyoto Submission for the WMT'20 Quality Estimation Task: Intermediate Training for Domain and Task Adaptation. In *Proceedings of the Fifth Conference on Machine Translation*, 1042–1048. Online: Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics.

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Specia, L. 2011. Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Specia, L.; Scarton, C.; and Paetzold, G. H. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1): 1–162.

Specia, L.; Shah, K.; De Souza, J. G.; and Cohn, T. 2013. QuEst-A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 79–84.

Tan, B.; Zhang, Y.; Pan, S.; and Yang, Q. 2017. Distant domain transfer learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

van der Wees, M.; Bisazza, A.; and Monz, C. 2017. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1400–1410. Copenhagen, Denmark: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems 30*.

Wang, K.; Wang, J.; Ge, N.; Shi, Y.; Zhao, Y.; and Fan, K. 2020a. Computer Assisted Translation with Neural Quality Estimation and Auotmatic Post-Editing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2175–2186.

Wang, M.; Yang, H.; Shang, H.; Wei, D.; Guo, J.; Lei, L.; Qin, Y.; Tao, S.; Sun, S.; Chen, Y.; et al. 2020b. Hw-tsc's participation at wmt 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, 1056–1061.

Wang, W.; Caswell, I.; and Chelba, C. 2019. Dynamically Composing Domain-Data Selection with Clean-Data Selection by "Co-Curricular Learning" for Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1282–1292. Florence, Italy: Association for Computational Linguistics.

Wu, H.; Wang, Z.; Ma, Q.; Wen, X.; Wang, R.; Wang, X.; Zhang, Y.; Yao, Z.; and Peng, S. 2020. Tencent submission for WMT20 Quality Estimation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, 1062–1067. Online: Association for Computational Linguistics.

Zhou, J.; Zhang, Z.; and Hu, Z. 2019. SOURCE: SOURce-Conditional Elmo-style Model for Machine Translation Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*.