

# Exploiting Curriculum Learning in Unsupervised Neural Machine Translation

Jinliang Lu<sup>1,2</sup> and Jiajun Zhang<sup>1,2</sup> \*

<sup>1</sup>National Laboratory of Pattern Recognition, Institute of Automation, CAS

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences  
{jinliang.lu, jjzhang}@nlpr.ia.ac.cn

## Abstract

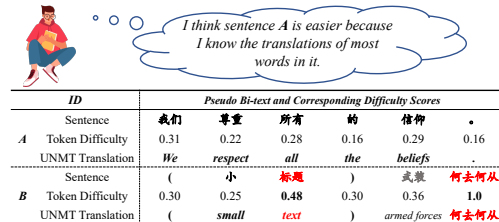
Back-translation (BT) has become one of the de facto components in unsupervised neural machine translation (UNMT), and it explicitly makes UNMT have translation ability. However, all the pseudo bi-texts generated by BT are treated equally as clean data during optimization without considering the quality diversity, leading to slow convergence and limited translation performance. To address this problem, we propose a curriculum learning method to gradually utilize pseudo bi-texts based on their quality from multiple granularities. Specifically, we first apply cross-lingual word embedding to calculate the potential translation difficulty (quality) for the monolingual sentences. Then, the sentences are fed into UNMT from easy to hard batch by batch. Furthermore, considering the quality of sentences/tokens in a particular batch are also diverse, we further adopt the model itself to calculate the fine-grained quality scores, which are served as learning factors to balance the contributions of different parts when computing loss and encourage the UNMT model to focus on pseudo data with higher quality. Experimental results on WMT 14 En↔Fr, WMT 16 En↔De, WMT 16 En↔Ro, and LDC En↔Zh translation tasks demonstrate that the proposed method achieves consistent improvements with faster convergence speed.<sup>1</sup>

## 1 Introduction

Unsupervised neural machine translation (UNMT) (Artetxe et al., 2018b; Lample et al., 2018a) has made significant progress (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020b; Tran et al., 2020) in recent years. It consists of three main components: the initialization of the cross-lingual pre-trained language model (PLM), denoising auto-encoder (AE) (Vincent et al., 2008), and

\*Corresponding author

<sup>1</sup>Our code is available in [https://github.com/JinliangLu96/CL\\_UNMT](https://github.com/JinliangLu96/CL_UNMT)



ID	Pseudo Bi-text and Corresponding Difficulty Scores					
Sentence	我们	尊重	所有	的	信仰	.
Token Difficulty	0.31	0.22	0.28	0.16	0.29	0.16
UNMT Translation	We	respect	all	the	beliefs	.
Sentence	(	小	标题	)	武装	何去何从
Token Difficulty	0.30	0.25	0.48	0.30	0.36	1.0
UNMT Translation	(	small	text	)	armed forces	何去何从

Figure 1: Difficulty scores in A are lower than B. And its translation is credible, making pseudo bi-text A better (red words in B are mis-translated or untranslated).

back-translation (BT) (Sennrich et al., 2016). BT generates pseudo bi-texts for training and explicitly enables its translation ability. However, pseudo bi-texts are quite diverse in quality, and the low-quality bi-texts are difficult to learn. Equally treating pseudo bi-texts as clean data would negatively influence the convergence process and harm the translation performance (Fadaee and Monz, 2018).

Recently, curriculum learning (CL) (Bengio et al., 2009), which aims to help the model learn from easy samples to the hard ones, has shown its effectiveness in speeding up the convergence and improving performance. Just as the name implies, the critical point of CL is *difficulty criteria*. Zhang et al. (2018) classify criteria in supervised machine translation into linguistic-inspired criteria (Kocmi and Bojar, 2017) and model-based criteria (Zhang et al., 2017, 2019; Zhou et al., 2020; Xu et al., 2020). Most of them are designed from the perspective of the source side in the pure parallel corpus. However, pseudo bi-texts produced by BT with monolingual sentences in UNMT contain different levels of noise, and low-quality samples with much noise would be difficult for the model to learn appropriately (Guo et al., 2018; Zhang et al., 2020). In this paper, we propose a CL method to gradually utilize pseudo bi-texts for UNMT from easy to hard, helping the model concentrating on the data with high quality from multiple granularities.

Intuitively, pseudo bi-text with high quality is

more accessible and suitable for UNMT optimization. Accordingly, we will measure the sample difficulty with bi-text quality. First, we apply the unsupervised cross-lingual word embedding (Lample et al., 2018b) to calculate the quality of bi-texts, which is in turn used to measure the sample quality. Then, samples will be fed into UNMT from easy to hard batch by batch based on their difficulty. Figure 1 illustrates that it is reasonable to bridge the bi-text quality and the sample difficulty.

However, the batch-based standard learning procedure is coarse-grained, and the qualities of pseudo bi-texts at sentence/word-level in a particular batch are also different, which should be addressed. To perform such fine-grained learning from easy to difficult, we borrow the idea from self-paced learning (Kumar et al., 2010), which is an adapted CL algorithm. Specifically, we first adopt the model to estimate the quality scores of pseudo bi-texts. Then, the scores are served as learning factors to balance the contributions of different parts when computing the training loss, encouraging the UNMT model to concentrate on the parts with higher quality.

In general, the contributions of this paper can be summarized as follows:

- We propose a multi-granularity CL method to improve UNMT. To the best of our knowledge, this is the first attempt to study the CL framework for UNMT.
- Through utilizing the quality of pseudo bi-text from multi-granularities, our method helps UNMT concentrate on the easy-to-learn part of data and optimize in the proper direction.
- Extensive experiments on WMT14 En↔Fr, WMT16 En↔De, WMT16 En↔Ro, and LDC En↔Zh translation tasks demonstrate that our method consistently outperforms the strong baselines with faster convergence speed.

## 2 Background of UNMT

The architecture of the current state-of-the-art UNMT is the same as supervised NMT model, except that the UNMT model simultaneously processes both translation directions. The training procedure comprises three main components: the initialization of cross-lingual PLM, denoising auto-encoder and back-translation.

**Cross-lingual PLM** is the auto-encoder that aims to encode the source sentences and target

sentences into a shared embedding space. The parameters are used to initialize the encoder and decoder in UNMT model before training.

**Denoising Auto-Encoder** is one of the crucial components for UNMT. It can improve the model learning ability through reconstructing the original sentences from the sentences with artificial noise, such as random deletion, swapping, or blanking. It is optimized by minimizing the following objective function:

$$\mathcal{L}_{auto} = \mathbb{E}_{x \sim \phi_{l_1}} [-\log P_{l_1 \rightarrow l_1}(x|C(x))] + \mathbb{E}_{y \sim \phi_{l_2}} [-\log P_{l_2 \rightarrow l_2}(y|C(y))] \quad (1)$$

where  $x$  and  $y$  indicate sentences sampled from monolingual dataset  $\phi_{l_1}$  and  $\phi_{l_2}$ .  $l_1$  and  $l_2$  are the two languages.  $C(\cdot)$  is the artificial noise function.

**Back Translation** is another essential component of UNMT, which explicitly ensure the model to have translation ability. First, each batch of monolingual sentences is translated into the other language by UNMT model  $M$ . Then,  $M$  applies the pseudo parallel sentences  $(M_{l_1 \rightarrow l_2}(x), x)$  and  $(M_{l_2 \rightarrow l_1}(y), y)$  into training. The process is called *on-the-fly back translation*. The objective function is:

$$\mathcal{L}_{bt} = \mathbb{E}_{x \sim \phi_{l_1}} [-\log P_{l_2 \rightarrow l_1}(x|M_{l_1 \rightarrow l_2}(x))] + \mathbb{E}_{y \sim \phi_{l_2}} [-\log P_{l_1 \rightarrow l_2}(y|M_{l_2 \rightarrow l_1}(y))] \quad (2)$$

In conclusion, the final loss during UNMT training can be written as follow:

$$\mathcal{L} = \mathcal{L}_{auto} + \mathcal{L}_{bt} \quad (3)$$

Even though strong UNMT models have been proposed in recent years, such as XLM (Conneau and Lample, 2019) and MASS (Song et al., 2019). The uneven quality of pseudo bi-text is still harmful. First, pseudo bi-texts are produced at each round. The translation performance in the early stages is pretty low and will affect the final results. Second, equally treating pseudo bi-texts with uneven quality can bring deviation to the optimization, slowing down the convergence speed and restricting translation performance.

## 3 Approach

In this section, we introduce the proposed CL method for UNMT. As shown in Figure 2, our method consists of two sub-modules that work at different levels:

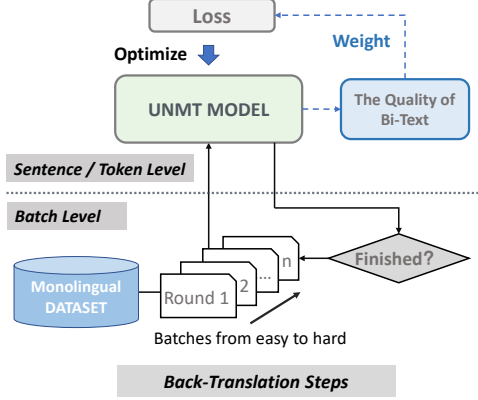


Figure 2: Illustration of our method. Batch level CL is shown below the black dash line, which controls the dataloader to prepare batches based on sample difficulty. Sentence/token level CL is illustrated above the black dash line, applying UNMT model to estimate the quality of pseudo bi-text and weight the training loss.

- 1) At batch level, we aim to optimize the dataloader so as to load the samples for training from easy to difficult batch by batch (§ 3.1);
- 2) At sentence/token level, we attempt to improve the parameter optimization procedure by using an adapted CL algorithm self-pace learning (Kumar et al., 2010), which calculates fine-grained difficulty scores and encourages the optimizer to pay more attention on easy-to-learn sentences/tokens (§ 3.2).

### 3.1 Batch Level CL

In this section, we introduce the CL method which controls the dataloader to load samples from easy to hard at the batch level. First, we describe the cross-lingual difficulty definition for the measurement of training samples. Then, we explain the sample loading schedule for UNMT.

#### 3.1.1 Difficulty Criterion Definition

As mentioned above, difficulty criterion is essential for CL. Traditional criteria, such as sentence length or word rarity, cannot reflect the practical complexity of pseudo bi-text.

We first use cross-lingual similarity to calculate the word-level bi-text quality, which is in turn used to define the word-level difficulty. Then, we weight the word-level difficulties by importance to get the sentence-level difficulty.

Specifically, pre-trained monolingual word embedding of language  $X$  and  $Y$  are first mapped

into the same latent space through MUSE (Lample et al., 2018b) toolkit and cross-lingual embedding matrices  $Z_X, Z_Y$  are obtained. Next, sentence  $x^i = \langle x_1^i, x_2^i, \dots, x_n^i \rangle$  is mapped into a sequence of vectors  $x^i = [x_1^i, x_2^i, \dots, x_n^i]$  through  $Z_X$ . Then, the difficulty of word  $x_j^i$  can be calculated, which is represented by the shortest distance from it to the target language space  $Z_Y$ :

$$d(x_j^i) = 1 - \max_{z_k \in Z_Y} \cos(x_j^i, z_k) \quad (4)$$

where  $z_k$  indicates an arbitrary word embedding in  $Z_Y$ . Considering the contribution of different words, sentence-level difficulty calculation incorporates importance weighting (indicated by `tfidf` score). Sentence length is further applied as the penalty. To sum up, the formula can be written as:

$$d(x^i) = \frac{\sum_{j=1}^n \text{tfidf}(x_j^i) \cdot d(x_j^i)}{\sum_{j=1}^n \text{tfidf}(x_j^i)} \cdot \log(n) \quad (5)$$

Finally, the difficulties are normalized to  $[0, 1]$  by minmax normalization, employed during the batch preparation.

#### 3.1.2 Sample Loading Schedule

The second question in CL is how to design the sample loading schedule, which determines how complex samples the UNMT can accept at specific steps. We follow the competence definition designed by Platanios et al. (2019), which indicates the capacity of the model:

$$c(t) = \min(1, \sqrt[p]{\frac{t}{T}(1 - c_0^p) + c_0^p}) \quad (6)$$

where  $c_0$  is the initial competence,  $p$  (set as 2 in our experiments) is the coefficient to control the curriculum schedule. At step  $t$ , sentences with  $d(x^i) \leq c(t)$  become accessible to the model. And  $T$  determines the step when all the sentences become available for the model.

We compute UNMT competence at specific steps for efficiency. At the beginning of the training process, all the available samples ( $d(x^i) \leq c_0$ ) are grouped into batches. Then, the batches are shuffled and successively transported into the model. When all of them are used up, the next phase will start with the update of  $c_t$ , sentence selection, and the batch preparation. Through the learning schedule, UNMT gradually receives the samples from easy to hard batch by batch.

### 3.2 Sentence/Token Level CL

While batch level CL controls dataloader to help UNMT learn from easy samples to the hard ones gradually, the qualities and difficulties of sentences and words in a particular batch are also different. However, such fine-grained operations are not suitable for the dataloader. To address this problem, we apply the UNMT to estimate the quality of pseudo bi-text at sentence/word level. Then, the quality scores are employed to regulate training loss, helping UNMT automatically focus on the words and sentences with high quality.

#### 3.2.1 Cross-lingual PLM based Pseudo Bi-text Quality Estimation (CP)

Cross-lingual PLMs are proven to be effective on reference-free evaluation in machine translation (Qi, 2019; Yankovskaya et al., 2019; Kim et al., 2019; Zhao et al., 2020; Takahashi et al., 2020). Actually, the encoder of UNMT, which is initialized by the cross-lingual PLM, should also be able to judge the quality of pseudo bi-texts. Furthermore, since sentences/token level CL works on optimization, we apply the model to calculate dynamic quality estimation related to the current learning state instead of utilizing static scores.

Specifically, sentence  $x^i = \langle x_1^i, x_2^i, \dots, x_n^i \rangle$  is sampled from the monolingual dataset, its *on-the-fly* translation is  $\hat{y}^i = \langle \hat{y}_1^i, \hat{y}_2^i, \dots, \hat{y}_m^i \rangle$ . We apply the encoder to obtain the hidden states  $\mathbf{H}_{x^i} = [\mathbf{h}_{x_1^i}, \mathbf{h}_{x_2^i}, \dots, \mathbf{h}_{x_n^i}]$  and  $\mathbf{H}_{\hat{y}^i} = [\mathbf{h}_{\hat{y}_1^i}, \mathbf{h}_{\hat{y}_2^i}, \dots, \mathbf{h}_{\hat{y}_m^i}]$ . Then, the hidden states are employed to estimate the quality of pseudo bi-text.

**Token-Level Translation Quality (TTQ):** For token  $x_j^i$ , we use the greedy matching strategy to match it to the most similar token in  $\hat{y}^i$ . The corresponding quality of  $x_j^i$  is represented by the cosine similarity, which can be formulated as:

$$w = \max_{v \in \{1, 2, \dots, m\}} \cos(\mathbf{h}_{x_j^i}, \mathbf{h}_{y_v^i}) \quad (7)$$

$$\hat{\alpha}_j^i = w^k \quad (8)$$

where  $k$  is hyper-parameter for the quality gap scaling. To stabilize the training process and maintain the same loss scale as the conventional model, we normalize the quality scores by *softmax*:

$$\alpha_j^i = \frac{\exp(\hat{\alpha}_j^i)}{\sum_{t=1}^n \exp(\hat{\alpha}_t^i)} \quad (9)$$

**Sentence-Level Translation Quality (STQ):** We take the average of the token hidden states as the

sentence-level features, written as  $\mathbf{h}_{x^i}$  and  $\mathbf{h}_{\hat{y}^i}$ . The sentence-level quality can be calculated as:

$$u = \cos(\mathbf{h}_{x^i}, \mathbf{h}_{\hat{y}^i}) \quad (10)$$

$$\hat{\beta}^i = u^k \quad (11)$$

Similarly, sentence-level quality scores are also normalized by *softmax*:

$$\beta^i = \frac{\exp(\hat{\beta}^i)}{\sum_{t=1}^M \exp(\hat{\beta}^t)} \quad (12)$$

where  $M$  represents the batch size.

#### 3.2.2 JS-Divergence based Confidence Estimation (JS)

An alternative of CP is *Two-Pass* JS-divergence, which can reflect the difference between token distributions. It can be formulated as

$$JS(p||q) = \frac{1}{2}KL(p||r) + \frac{1}{2}KL(q||r) \quad (13)$$

where  $p$  and  $q$  represent the distributions of tokens at each force-decoding step with different *dropout*, and  $r = (p + q)/2$ .

**Token-Level JS Score**  $\alpha_j^i$  is the JS score of  $j$ -th token in sentence  $i$  during force-decoding.

**Sentence-Level JS Score**  $\beta^i$  is represented by the mean of token-level JS confidence in the  $i$ -th sentence.

Both of  $\alpha_j^i$  and  $\beta^i$  are multiplied by  $k$  power and normalized by *softmax*.

#### 3.2.3 Training Strategy

Higher score indicates better quality. So the corresponding tokens or sentences should contribute more when computing loss, helping UNMT optimize in the reasonable direction. Therefore, we apply the quality scores to regulate the training loss. The loss of  $i$ -th sentence can be calculate as:

$$\mathcal{L}_i = - \sum_{j=1}^n \alpha_j^i \log P(x_j^i | \hat{y}^i, x_{<j}^i; \theta) \quad (14)$$

And the total loss of mini-batch is:

$$\mathcal{L} = \sum_{i=1}^M \beta^i \mathcal{L}_i \quad (15)$$

During the training, CP can be only employed in BT steps, while JS can be employed not only in BT steps but also AE steps because it actually measures the model confidence. In our experiments, JS and CP are respectively applied in AE steps and BT steps. Further analyses also compare the performance of different estimation methods.

## 4 Datasets and Experiment Settings

### 4.1 Datasets

**Pre-training:** For En-Fr, En-De, En-Ro, we download pre-trained language models from XLM<sup>2</sup> and MASS<sup>3</sup> toolkits. For En-Zh, we train a standard XLM model from scratch. The monolingual data consists of WMT 2008-2019 News Crawl dataset (5M Chinese sentences in total and 5M English sentences uniformly selected for equality).

**UNMT:** For En-Fr, En-De, En-Ro, we respectively keep 2M (1M English, 1M the other language) sentences for training from WMT News Crawl. For En-Zh, we extract Chinese sentences from the first half of the 2M parallel sentences in LDC, and English sentences from the other half. WMT *newstest 2013/2014*, *newstest 2013/2016*, *newsdev/newstest 2016* and NIST03/NIST06 as validation/test sets for En-Fr, En-De, En-Ro, and En-Zh, respectively.

### 4.2 Settings

**CL Settings:** For difficulty computation, MUSE<sup>4</sup> is applied to map the monolingual word embeddings<sup>5</sup> into the common space.  $c_0 = 0.01$  for En-De, En-Ro and En-Zh,  $c_0 = 0.1$  for En-Fr.  $T$  is approximately estimated by the step when UNMT baseline reaches 90% BLEU (Papineni et al., 2002) on the valid set.

**UNMT Settings:** During training, mini-batches are limited to 2000 tokens and maximum sequence length is 100 tokens. Adam with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.998$ ,  $lr = 0.0001$  is employed for optimization. When decoding, we use beam size as 4 and length penalty as 1.0 for each language pair. 4-gram BLEU score computed by *multi-bleu.perl*<sup>6</sup> script is reported for comparison.

## 5 Experimental Results

### 5.1 Translation Quality

Table 1 shows the UNMT results on different translation tasks. XLM and MASS are the baseline results<sup>7</sup>. Our proposed method consistently out-

<sup>2</sup><https://github.com/facebookresearch/XLM>

<sup>3</sup><https://github.com/microsoft/MASS>

<sup>4</sup><https://github.com/facebookresearch/MUSE>

<sup>5</sup>we download fasttext embeddings pretrained on wiki, <https://fasttext.cc/>

<sup>6</sup><https://github.com/moses-smt/mosesdecoder>

<sup>7</sup>With the limitation of resources, the size of our training datasets is less than 2% of the ones used in (Conneau and

performs the strong baselines, demonstrating the effectiveness of our method. Furthermore, removing either batch-level CL or sentence/word-level CL decreases the translation improvements on most language pairs, indicating the two parts are complementary.

Another interesting finding is that sentence/word level CL is more effective on similar languages, such as En-Fr, En-De, while single batch-level CL is suitable for the distant language pair like En-Zh. We assume that cross-lingual PLM on similar languages could provide hidden states with accurate semantic information, precisely estimating the quality of pseudo bi-text. In contrast, distant languages cannot fully take the advantage, while heuristic difficulty criteria help more.

### 5.2 Convergence Speed

Most curriculum learning methods aim to accelerate convergence speed while improving performance. We visualize the average loss of training samples and the learning curve to compare the convergence speed on WMT Ro→En *newstest2016* in Figure 3. Both of the curves indicate that our method achieves convergence at a higher speed.

The left part of Figure 3 shows the loss curves. At the beginning of the training process, the average losses of different methods decrease with different speeds. However, the loss curves of the batch level CL and the baseline almost coincide at the end. When adding sentence/word level CL, the model achieves a lower loss than baseline, demonstrating the rationality of our weighted learning objective.

On the other hand, the learning curves, which are represented by the BLEU on valid set, clearly describe the efficiency of our method. As shown in the right part of Figure 3, XLM baseline reaches convergence at step 31k, while our approach achieves the same performance at step 10k, indicating that our methods are 3.1 times faster.

The acceleration ratios for different languages are recorded in Table 2. Our methods significantly accelerate the training process. Considering the time exhausted in the computation of quality estimation, we also calculate the time acceleration. The records indicate that our methods can achieve equivalent performance with less training time.

Lample, 2019). Therefore, the baseline results are a bit lower.

Model	En-Fr		En-De		En-Ro		En-Zh	
	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En	En→Zh	Zh→En
XLM	35.89	33.58	26.21	32.51	33.48	30.97	12.97	26.42
+ Both Level	<b>36.31</b>	<b>33.97</b>	<b>27.22</b>	<b>33.26</b>	<b>35.05</b>	<b>32.00</b>	<b>13.70</b>	<b>28.18</b>
w/o s/t level	35.70	33.77	26.27	32.69	34.04	31.78	<b>13.70</b>	27.33
w/o batch level	35.91	33.90	27.01	33.21	34.72	31.58	13.30	27.04
MASS	34.97	32.98	26.93	32.20	34.32	31.58	-	-
+ Both Level	<b>35.36</b>	<b>33.40</b>	<b>27.53</b>	<b>32.62</b>	<b>34.86</b>	<b>32.27</b>	-	-

Table 1: BLEU scores of different UNMT methods for translations to and from English. Experiments on XLM are listed above the double lines and experiments on MASS are listed below it. "s/t" means "sentence/token".

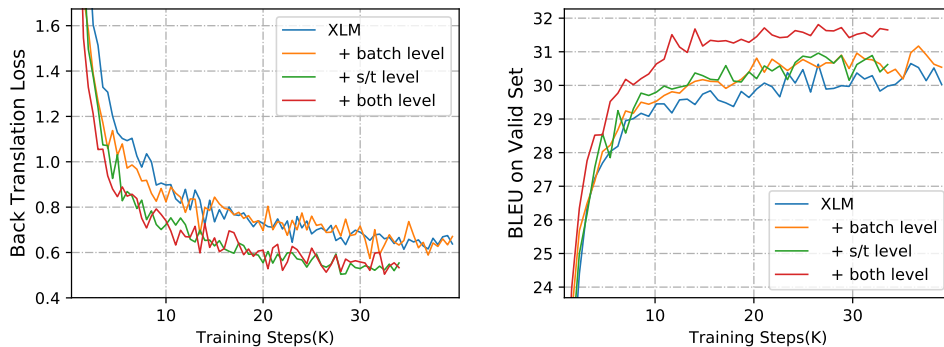


Figure 3: Average loss on the training process (left) and learning curves on valid data set (right) of WMT En-Ro. Our method achieves lower loss and higher BLEU score with faster convergence speed.

Language Direction	Our method	
	Step Acc.	Time Acc.
En→De	5.91x	4.86x
De→En	2.46x	1.95x
En→Ro	2.78x	2.15x
Ro→En	3.08x	2.41x

Table 2: Acceleration on steps and time upon WMT En-De *newstest2016* and WMT En-Ro *newstest2016*. The acceleration is calculated by the ratio of the steps(time) when the baseline model reaches convergence to the steps(time) when our methods achieve equivalent translation quality.

## 6 Analysis

### 6.1 Correlation Between the Difficulty and Improvements

Even though our methods improve across all the language pairs, it remains a question which part of sentences contribute more to the performance. Figure 4 shows the BLEU improvements at different difficulty intervals on WMT En→Ro *newstest2016*. The difficulty is represented by the definition described in section 3.1.1. We find that our approach outperforms XLM baseline in different difficulty intervals. The easiest sentences (<2%) have significant improvements, which owes to the emphasis on the easy samples during training. In contrast,

hard sentences (>70%) have limited performance gains.

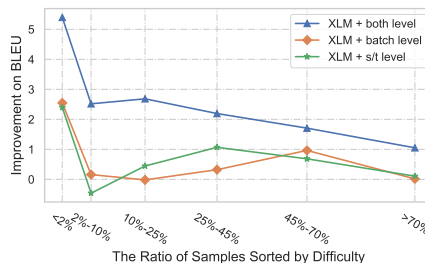


Figure 4: Improvements of BLEU at different difficulty intervals on WMT En→Ro *newstest2016*.

Figure 5 shows the relationship between the improvement of sentence-level BERTScore (Zhang\* et al., 2020) and the difficulty distribution. The larger points are sparsely distributed on the left side, indicating that simple sentences achieve significant improvements. And the minor points are concentrated in the lower right corner, meaning that complex sentences yield slight performance improvement.

This finding is different from related works on supervised NMT (Xu et al., 2020; Liu et al., 2020a), which prove that curriculum learning is beneficial for complex samples. We suspect the reason lies

Method	De→En	$\Delta_{\text{BLEU}}$	Ro→En	$\Delta_{\text{BLEU}}$
XLM	32.51	-	30.97	-
+BT <sub>Length</sub>	32.21	-0.30	31.04	+0.07
+BT <sub>Rarity</sub>	32.08	-0.43	30.95	-0.02
+BT <sub>Ours</sub>	<b>32.69</b>	<b>+0.18</b>	<b>31.78</b>	<b>+0.81</b>

Table 3: The comparison of different difficulty criteria on WMT De→En *newstest2016* and WMT En→Ro *newstest2016*.  $\Delta_{\text{BLEU}}$  represents the performance increase or decrease compared with XLM baseline.

in the particularity of our method, considering the quality of bi-texts instead of the pure difficulty. Therefore, we think our method helps the UNMT model mainly strengthen its essential translation ability.

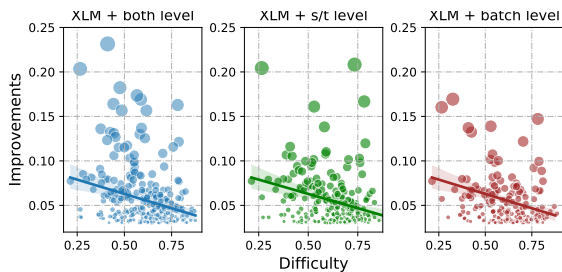


Figure 5: The relationship between difficulty distribution and improvements of BERTScore. The size of points indicate corresponding improvements.

## 6.2 Comparison of Difficulty Criteria

To verify the effectiveness of our difficulty definition, we compare it with traditional difficulty criteria (such as length and rarity) on the single batch level CL. As shown in Table 3, the proposed definition achieves better performance than both length and rarity. By contrast, traditional artificial difficulties do not improve the UNMT translation quality and may even cause damage.

We assume that traditional difficulty criteria are not appropriate for UNMT because many noises exist in pseudo bi-text during BT steps, which significantly changes the distribution of sentence-level difficulty. By comparison, our difficulty definition considers word-level translation difficulties. Intuitively, words with lower difficulties can be translated at higher quality, producing pseudo bi-text with fewer noises and easy to learn. Therefore, our difficulty definition implicitly describes the influence of noise, making the learning schedule more suitable for UNMT.

Method	En→De	En→Ro	Speed
AE <sub>None</sub> + BT <sub>None</sub>	26.21	33.48	3183 (1.00x)
AE <sub>JS</sub> + BT <sub>JS</sub>	26.28	33.72	2119 (0.67x)
AE <sub>VAR</sub> + BT <sub>VAR</sub>	26.51	33.38	1454 (0.46x)
AE <sub>None</sub> + BT <sub>CP</sub>	27.01	34.35	2961 (0.93x)
AE <sub>JS</sub> + BT <sub>CP</sub>	<b>27.22</b>	<b>34.72</b>	2475 (0.78x)

Table 4: The comparison of different estimation methods on WMT En→De *newstest2016* and WMT Ro→En *newstest2016*. ST methods are listed below the dash line. Average speed (tokens/s) is measured on NVIDIA V100 and numbers in brackets is the fraction compared with XLM baseline.

## 6.3 Comparison of Different Estimation Methods

We also compare the effect of our quality estimation approach with different confidence estimation methods. This part of the experiments is conducted without batch level CL for more evident results. Table 4 shows that AE<sub>JS</sub> + BT<sub>CP</sub> yields the best results among the methods, indicating the proposed estimation method is more engaging for UNMT. On the other hand, we find that single CP helps while single JS almost does not affect. Uncertainty-based model confidence AE<sub>VAR</sub> + BT<sub>VAR</sub> (VAR is the abbreviation of variance)<sup>8</sup>, which is proven to be helpful in supervised NMT (Wang et al., 2019; Wan et al., 2020), achieves only limited performance improvements in UNMT. Besides, the computation of VAR is time-consuming, slowing down the training efficiency by 54% at each step.

BT step is of vital importance for the translation ability of UNMT, which can be described as a rough imitation of NMT steps. Uncertainty-based confidence estimation is practical when bi-texts are pure. However, when the information provided by the particular bi-text is not equal, great deviation would be brought into the estimation of VAR or  $\mathbb{E}$ . By contrast, the quality of bi-text is much essential under this circumstance. We think that is the reason why CP yields higher translation performance.

## 6.4 STQ Versus TTQ

As we described in section 3.2.1, fine-grained quality scores are estimated on the sentence-level (STQ) and the token-level (TTQ). We also compare their influence on translation performance. As shown in Table 5, both STQ and TTQ improve the translation quality on WMT *newstest2016* En-De and

<sup>8</sup>Computing VAR for each token needs  $Q$ -Pass forward computation with different dropout,  $Q$  is set as 5 in the experiments.

Method	En-De		En-Ro	
	en→de	de→en	en→ro	ro→en
XLM	26.21	32.51	33.48	30.97
AE <sub>JS</sub> + BT <sub>CP</sub>	<b>27.22</b>	<b>33.26</b>	<b>34.72</b>	<b>31.58</b>
w/o BT <sub>STQ</sub>	26.69	32.54	34.42	31.34
w/o BT <sub>TTQ</sub>	26.92	32.88	34.59	31.43

Table 5: Comparison of sentence-level (STQ) and token-level (TTQ) quality estimation. STQ performs better than TTQ.

WMT *newstest2016* En-Ro. Interestingly, STQ outperforms TTQ. We suspect that cross-lingual PLM can estimate sentence-level quality more accurately than the token-level. Intuitively, the combination of STQ and TTQ achieves better results.

## 6.5 Effect of $k$

Fine-grained quality scores in the proposed method are calculated by the  $k$ -th power of cosine similarity. Therefore, we compare the translation performance with different  $k$ . The results are shown in Figure 6. Histogram illustrates that UNMT yields the best performance when  $k = 2$ . However, when  $k > 2$ , the translation quality slightly decreases. We assume that appropriate  $k$  can help our approach accurately reflect the translation quality, benefiting the UNMT performance.

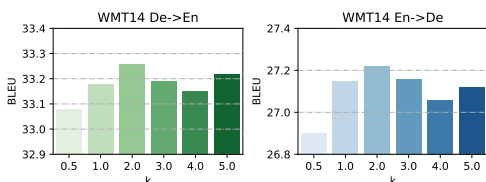


Figure 6: The effect of  $k$  on the performance upon WMT En-De *newstest2016*.

## 7 Related Work

### 7.1 UNMT

Lample et al. (2018a) and Artetxe et al. (2018b) propose UNMT using monolingual corpus only, which established on the progress of cross-lingual word embedding projection (Artetxe et al., 2018a; Lample et al., 2018b). Recent years, UNMT rapidly develops with the help of pre-trained language models. Conneau and Lample (2019) releases the first cross-lingual PLM, named XLM, greatly improving the UNMT performance. Song et al. (2019), Liu et al. (2020b), and Tran et al. (2020) designs different seq2seq pre-training strategy, achieving state-of-the-art UNMT performance.

Even though various models are proposed, the key of UNMT is still the cross-lingual ability. Sun et al. (2019) uses an agreement method to train UNMT with bilingual word embedding agreement. Ren et al. (2019) ameliorates the cross-lingual ability of BERT (Devlin et al., 2019) through predicting n-gram translation of masked tokens, benefiting the UNMT performance. Chronopoulou et al. (2020) modifies the predefined vocabulary of XLM for UNMT with limited monolingual corpus.

However, most of previous work focuses the cross-lingual ability of word embedding or PLM but ignores the efficiency of the training process in UNMT.

### 7.2 Curriculum Learning in NMT

Curriculum learning (Bengio et al., 2009) is motivated by the learning strategy of biological organisms which orders the training samples in an easy-to-hard manner (Elman, 1993). It has recently shown its effectiveness on machine translation tasks by changing the order of training samples. Kocmi and Bojar (2017) examine the effects of particular orderings of sentence pairs on the NMT training in one epoch. Platanios et al. (2019) propose competence-based curriculum learning framework, selecting samples at each step based on the difficulty and competence. Liu et al. (2020a) use the norm of word embedding to modify the competence-based curriculum learning, improving the performance of supervised NMT. Zhou et al. (2020) apply uncertainty into the difficulty and competence design. Wan et al. (2020) adopt self-paced learning (Kumar et al., 2010) for NMT, replacing curriculum learning and yielding better performance. Xu et al. (2020) propose dynamic curriculum learning strategy for low-resource NMT. However, curriculum learning for UNMT is still unexploited and our work is the first attempt.

## 8 Conclusion

In this paper, we propose a multi-granularity CL method to improve UNMT. Specifically, a novel cross-lingual difficulty definition is first proposed to help UNMT learn from easy samples to the hard ones at batch level. Then, the qualities of pseudo bi-text at sentence/word-level are estimated by the model itself to regulate the loss function, automatically helping UNMT optimize in the appropriate direction. Empirical results show that our method outperforms the strong baselines on different lan-



guage pairs with faster convergence speed. Further analyses confirm that our CL methods at different levels are helpful and complementary with each other, indicating the suitability for UNMT. In the future, we will explore its ability on multilingual machine translation and other cross-lingual generation tasks.

## Acknowledgements

The research work described in this paper has been supported by the Natural Science Foundation of China under Grant No. U1836221.

## References

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. [A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. [Unsupervised neural machine translation](#). In *International Conference on Learning Representations*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 41–48, New York, NY, USA. Association for Computing Machinery.
- Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. [Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.
- Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.
- Marzieh Fadaee and Christof Monz. 2018. [Back-translation sampling by targeting difficult words in neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.
- Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. 2018. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150.
- Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019. [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)*, pages 85–89, Florence, Italy. Association for Computational Linguistics.
- Tom Kocmi and Ondřej Bojar. 2017. [Curriculum learning and minibatch bucketing in neural machine translation](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.
- M. Kumar, Benjamin Packer, and Daphne Koller. 2010. [Self-paced learning for latent variable models](#). In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. [Unsupervised machine translation using monolingual corpora only](#). In *International Conference on Learning Representations*.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. [Word translation without parallel data](#). In *International Conference on Learning Representations*.
- Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020a. [Norm-based curriculum learning for neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 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)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hou Qi. 2019. [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)*, pages 95–100, Florence, Italy. Association for Computational Linguistics.
- Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. [Explicit cross-lingual pre-training for unsupervised machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, Hong Kong, China. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tiejun Liu. 2019. [Mass: Masked sequence to sequence pre-training for language generation](#). In *International Conference on Machine Learning*, pages 5926–5936.
- Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. [Unsupervised bilingual word embedding agreement for unsupervised neural machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245, Florence, Italy. Association for Computational Linguistics.
- Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. [Automatic machine translation evaluation using source language inputs and cross-lingual language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.
- Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. [Cross-lingual retrieval for iterative self-supervised training](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. [Extracting and composing robust features with denoising autoencoders](#). In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, page 1096–1103, New York, NY, USA. Association for Computing Machinery.
- Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. [Self-paced learning for neural machine translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080, Online. Association for Computational Linguistics.
- Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. [Improving back-translation with uncertainty-based confidence estimation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.
- Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020. [Dynamic curriculum learning for low-resource neural machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. [Quality estimation and translation metrics via pre-trained word and sentence embeddings](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.
- Dakun Zhang, Jungi Kim, Josep Crego, and Jean Senellart. 2017. [Boosting neural machine translation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 271–276, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. [An empirical exploration of curriculum learning for neural machine translation](#). *arXiv preprint arXiv:1811.00739*.
- Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. [Curriculum learning for domain adaptation in 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)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuchao Zhang, Xian Wu, Fanglan Chen, Liang Zhao, and Chang-Tien Lu. 2020. Self-paced robust learning for leveraging clean labels in noisy data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6853–6860.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. [On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. [Uncertainty-aware curriculum learning for neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.