Unsupervised Neural Machine Translation with Cross-lingual Language Representation Agreement

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao

***Abstract*—Unsupervised cross-lingual language representation initialization methods such as unsupervised bilingual word embedding (UBWE) pre-training and cross-lingual masked language model (CMLM) pre-training, together with mechanisms such as denoising and back-translation, have advanced unsupervised neural machine translation (UNMT), which has achieved impressive results on several language pairs, particularly French-English and German-English. Typically, UBWE focuses on initializing the word embedding layer in the encoder and decoder of UNMT, whereas the CMLM focuses on initializing the entire encoder and decoder of UNMT. However, UBWE/CMLM training and UNMT training are independent, which makes it difficult to assess how the quality of UBWE/CMLM affects the performance of UNMT during UNMT training. In this paper, we first empirically explore relationships between UNMT and UBWE/CMLM. The empirical results demonstrate that the performance of UBWE and CMLM has a significant influence on the performance of UNMT. Motivated by this, we propose a novel UNMT structure with cross-lingual language representation agreement to capture the interaction between UBWE/CMLM and UNMT during UNMT training. Experimental results on several language pairs demonstrate that the proposed UNMT models improve significantly over the corresponding state-of-the-art UNMT baselines.**

***Index Terms*—Unsupervised neural machine translation, unsupervised bilingual word embedding, cross-lingual language model**

I. Introduction

NEURAL network-based supervised bilingual word embedding (BWE) methods have attracted much attention since 2013 [2]–[11]. Recently, many studies have shown that BWE can be learned only by depending on monolingual corpora, which indicates that bilingual supervision signals are not always essential for BWE [12]–[15]. For example, Conneau *et al*., [14] proposed an unsupervised BWE (UBWE) method and reported remarkable results for word-translation tasks. The success of UBWE enables machine translation to be modeled without any supervised bilingual signal. Artetxe *et al*. [16] and Lample *et al*. [17] first combined UBWE with a denoising auto-encoder and back-translation to build a novel unsupervised neural machine translation (UNMT), which relies only on monolingual corpora, and achieves impressive results on French-English and German-English translation tasks. More recently, Lample and Conneau extended the cross-lingual language representation from word embedding to a cross-lingual masked language model (CMLM) [18]. The empirical results demonstrated that the pre-trained CMLM improved the performance of UNMT.

In the existing UNMT, UBWE is first trained on monolingual corpora, and then initializes the word embedding layer in the encoder and decoder of UNMT. Furthermore, the encoder is pre-trained based on monolingual corpora by the CMLM method, and then the full parameters are used to initialize the entire encoder and decoder of UNMT instead of only initializing the word embedding layer. Despite their success on the UNMT, UBWE/CMLM training and UNMT training are independent, which makes it difficult to assess how the quality of UBWE/CMLM affects the performance of UNMT. In this paper, we first empirically explore the relationships between UNMT and UBWE/CMLM. As a result, there are two main empirical findings:

- 1) A positive correlation exists between the pre-trained UBWE quality and UNMT performance. UBWE quality significantly decreases (becomes worse) during UNMT training.
- 2) There is a positive correlation between the pre-trained CMLM quality and UNMT performance. CMLM quality significantly decreases (becomes worse) during UNMT training.

In fact, UBWE pre-trained UNMT and CMLM pre-trained UNMT are independent. The CMLM pre-training system performs better than the UBWE pre-training system. In this paper, we propose a general agreement method for UNMT systems, regardless of whether the UBWE pre-training system or CMLM pre-training system is used. In detail, we propose a new UNMT with cross-lingual language

This work is an extension of our paper "*Unsupervised Bilingual Word Embedding Agreement for Unsupervised Neural Machine Translation*" [1] presented at ACL-2019. In the ACL paper, we focused on UNMT with bilingual word embedding agreement only. In this paper, we extend the study to the general structure of UNMT with cross-lingual language representation agreement. The primary extensions are as follows: 1) We empirically investigate the relationship between CMLM and UNMT and propose two methods to train UNMT with CMLM agreement. 2) We introduce a knowledge distillation-based method to train the UNMT model with CMLM agreement. 3) We evaluate our method on the state-of-the-art WMT2019 German-Czech unsupervised translation task.

Haipeng Sun and Tiejun Zhao are with the Machine Intelligence and Translation Laboratory, School of Computer Science of Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: hpsun@hit-mtlab.net). Haipeng Sun was an internship research fellow at NICT under the supervision of Rui Wang when this work was conducted. Tiejun Zhao was supported by National Key Research and Development Program of China via grant 2017YFB1002102.

Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita are with the Advanced Speech Translation Research and Development Promotion Center, National Institute of Information and Communications Technology, Kyoto 619-0289, Japan (e-mail: wangrui@nict.go.jp). Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): "Unsupervised Neural Machine Translation in Universal Scenarios" and NICT tenure-track researcher startup fund "Toward Intelligent Machine Translation".

representation agreement to capture the interaction between UBWE/CMLM and UNMT during training. According to the aforementioned initialization methods, there are two cross-lingual language representation agreement approaches: UBWE agreement and CMLM agreement. For the UBWE agreement method, we propose UBWE agreement regularization and a UBWE adversarial training strategy to preserve UBWE quality during UNMT training. For the CMLM agreement method, we propose CMLM agreement regularization and a CMLM knowledge distillation strategy to maintain CMLM quality during UNMT training. The experimental results on several language pairs demonstrate that our proposed UNMT with cross-lingual language representation agreement significantly outperforms conventional UNMT.

II. Background of UNMT

Typically, UNMT mainly consists of three components: unsupervised cross-lingual language representation initialization, including UBWE pre-training or CMLM pre-training; a denoising auto-encoder; and back-translation.

Formally, given two monolingual corpora X and Y in two languages L_1 and L_2, ϕ_{L_1} and ϕ_{L_2} denote the data spaces of X and Y, respectively. To obtain a robust translation model, UNMT is first initialized by the UBWE or CMLM mechanism and is then trained by maximizing the objective function $\mathcal{L}_{entire}$ [19]:

$$\mathcal{L}_{entire} = \mathcal{L}_{deno} + \mathcal{L}_{back}, \tag{1}$$

where $\mathcal{L}_{deno}$ denotes the loss function for the denoising auto-encoder and $\mathcal{L}_{back}$ denotes the loss function for back-translation.

A. Initialization

There are primarily two types of initialization for UNMT:

1) Bilingual Word Embedding Pre-training: Compared with supervised neural machine translation (NMT) [20]–[24], UNMT has no bilingual supervised signals. Fortunately, UBWE [13]–[15] has successfully learned the equivalent translation between word pairs from two non-parallel monolingual corpora. Generally, UBWE initializes the embedding of the UNMT encoder and decoder. The naive translation knowledge provided by this pre-trained UBWE enables back-translation to generate pseudo-supervised bilingual signals [16], [17]. During the UNMT training process, the embeddings of the encoder and decoder change independently.

2) Cross-lingual Masked Language Model Pre-training: A universal cross-lingual encoder that encodes two monolingual sentences into a shared latent space is established by CMLM. The encoder and decoder of the UNMT model are initialized by the pre-trained cross-lingual encoder [18], except the multi-head attention over the output of the encoder stack in the decoder. Compared with UBWE pre-training, CMLM pre-training arguably provides more cross-lingual information for the UNMT model.

B. Denoising Auto-encoder

Without any restrictions, it is difficult for an auto-encoder to learn useful information about UNMT. In fact, it becomes a verbatim copying task [17]. To mitigate this problem, we follow a denoising auto-encoder strategy [25], and introduce noise in the form of random token deletions and reorderings in the input sentence to improve the learning ability of the UNMT model [26], [27]. In detail, with a certain probability, some tokens in the input sentence are deleted and the input sentence is slightly shuffled. The loss function of denoising auto-encoder training is formulated as follows:

$$\begin{aligned}\mathcal{L}_{deno} = &\mathbb{E}_{X\sim\phi_{L_1}}[-\log P_{L_1\to L_1}(X|N(X)] \\ &+ \mathbb{E}_{Y\sim\phi_{L_2}}[-\log P_{L_2\to L_2}(Y|N(Y)],\end{aligned} \tag{2}$$

where $N(X)$ and $N(Y)$ denote noisy versions of input sentences X and Y, respectively. The reconstruction probability in language L_1 (L_2) that encodes noisy input sentences $N(X)$ ($N(Y)$) and reconstructs them using the UNMT decoder in the same language [19] is denoted by $P_{L_1\to L_1}$ ($P_{L_2\to L_2}$).

C. Back-translation

The denoising auto-encoder does not achieve translation across two languages. Therefore, back-translation [28] is used to train an unsupervised translation system that relies solely on monolingual corpora. Specifically, given the sentences X and Y, the UNMT model PM in the previous iteration generates the sentences $Y_{PM}(X)$ and $X_{PM}(Y)$, respectively. The new UNMT model is trained by the pseudo-parallel sentence pairs $\{Y_{PM}(X), X\}$ and $\{X_{PM}(Y), Y\}$. Finally, the loss function of back-translation is minimized as follows:

$$\begin{aligned}\mathcal{L}_{back} = &\mathbb{E}_{X\sim\phi_{L_1}}[-\log P_{L_2\to L_1}(X|Y_{PM}(X)] \\ &+ \mathbb{E}_{Y\sim\phi_{L_2}}[-\log P_{L_1\to L_2}(Y|X_{PM}(Y)],\end{aligned} \tag{3}$$

where the translation probability across two languages that encodes pseudo input sentences $X_{PM}(Y)$ ($Y_{PM}(X)$) and generates them using the UNMT decoder in a different language [19] is denoted by $P_{L_1\to L_2}$ ($P_{L_2\to L_1}$).

III. Preliminary Experiments

A. Relationship between UNMT and UBWE

One similar language pair in the same language family (French-English) and one distant language pair in different language families (Japanese-English) were selected as the corpora to investigate the relationship between UNMT and UBWE. In Section V, we present the detailed experimental settings for UNMT and UBWE.

1) Effect of UBWE Quality on UNMT Performance: UNMT performance using UBWE pre-training with different accuracies is shown in Fig. 1. The `VecMap` [15] embeddings at different pre-training checkpoints were used to initialize the UNMT. The precision of word translation in the `MUSE` test set[1] using the first-ranked predicted candidate is indicated by Precision@1. Precision@1 of monolingual

[1]https://github.com/facebookresearch/MUSE

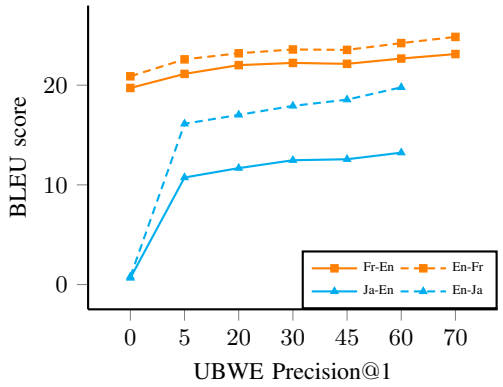


Fig. 1. UNMT performance using UBWE pre-training with different levels of accuracy.

embeddings, which was used in both languages before `VecMap` started training, was zero. For both language pairs, UNMT performance increased as UBWE Precision@1 increased. This demonstrates that the pre-trained UBWE quality influenced UNMT performance.

2) Trend in UBWE Quality during UNMT Training: The trend in the BLEU score and UBWE accuracy throughout the process of UNMT training is shown in Fig. 2. The embedding layer of the UNMT encoder and decoder was initialized by `VecMap`. The source embedding of the encoder and target embedding of the decoder at different checkpoints during UNMT training were used to compute the word translation accuracy on the `MUSE` test set.

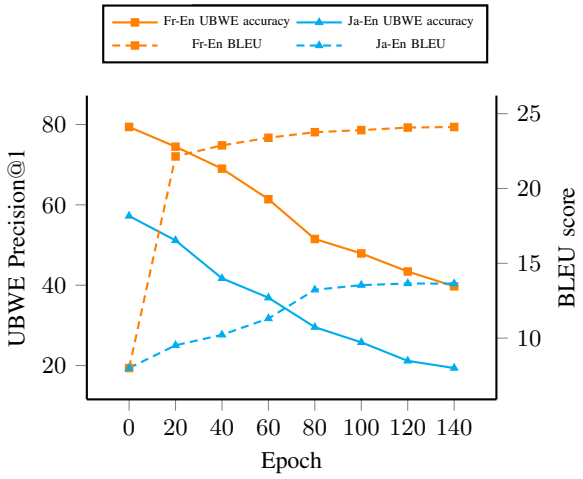


Fig. 2. BLEU score and UBWE accuracy over the entire UNMT training process.

Fig. 2 shows that UBWE accuracy degraded significantly over the entire UNMT training process on both language pairs.

3) Analysis: High pre-trained UBWE quality was essential to UNMT performance as the empirical results demonstrate. However, the UBWE quality degraded significantly as UNMT proceeded. We assume that preserving UBWE quality could improve UNMT performance. Therefore, we attempted to fix the encoder and decoder embedding based on the original baseline system (Baseline-fix). The performances of the original baseline system and Baseline-fix system were very similar as shown in Table I. Fixed embedding did not improve UNMT performance because it hindered UBWE from further participating in enhancing UNMT performance while maintaining UBWE accuracy.

TABLE I
RESULTS FOR UNMT.

Method	Fr-En	En-Fr	Ja-En	En-Ja
Baseline	24.43 (0.07)	25.30 (0.08)	13.85 (0.19)	21.52 (0.17)
Baseline-fix	24.16 (0.03)	25.14 (0.11)	13.62 (0.20)	21.82 (0.11)

Note: The first number in each column denotes the expectation of results over three runs; the number in parentheses denotes the standard deviation of results over three runs.

B. Relationship between UNMT and CMLM

The English-French language pair was chosen to investigate the relationship between UNMT and CMLM. In Section V, we present the complete experimental settings for UNMT and CMLM.

1) Effect of CMLM Quality on UNMT Performance: Fig. 3 shows the performance of the UNMT model using CMLM with different perplexity levels. To obtain the CMLM with different perplexity levels, we used the CMLM at different pre-training checkpoints to initialize UNMT. As CMLM perplexity decreased (became better), UNMT performance increased in both directions. This demonstrated that the quality of the pre-trained CMLM was essential for UNMT.

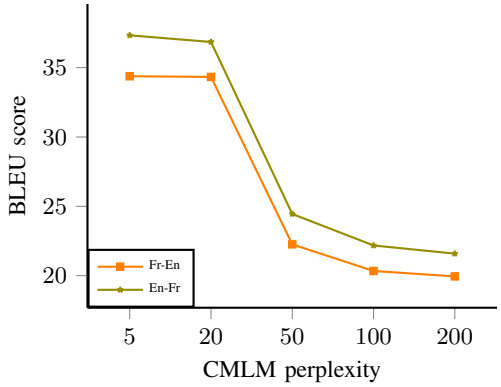


Fig. 3. UNMT performance using CMLM with different perplexity levels. Note that this perplexity denotes the average value of the perplexity scores for both languages.

2) Trend in CMLM Quality during UNMT Training: Fig. 4 shows the trend in the BLEU score and CMLM perplexity over the entire UNMT training process. The parameters of the pre-trained CMLM were used to initialize the parameters for the encoder and decoder of UNMT [18]. We used the encoder of UNMT to calculate the CMLM perplexity on the monolingual test set during UNMT training. The test set was the same as that used to calculate the BLEU scores during UNMT training.

Before UNMT training, the CMLM perplexity on the En test set was 5.75 and the CMLM perplexity on the Fr test set was 3.60. However, once UNMT training began, the

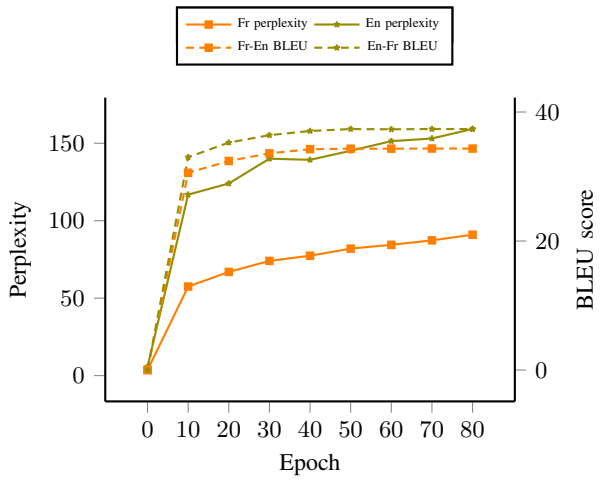


Fig. 4. BLEU score and CMLM perplexity over the entire UNMT training process.

CMLM perplexity increased (became worse) rapidly. In the early epoch of UNMT training, CMLM perplexity reached 110 on the En test set and 50 on the Fr test set. CMLM perplexity increased significantly on both monolingual datasets over the course of UNMT training. As shown in Fig. 3, the quality of the language model had an important influence on machine translation performance. We believe that jointly training UNMT and CMLM improved UNMT performance because CMLM training enriched the source representation[2].

IV. Methods

Based on previous empirical results and analyses on the relationship between UNMT and UBWE/CMLM in Section III, we propose general cross-lingual language representation agreement methods, that is, UBWE agreement and CMLM agreement, to improve translation performance.

Typically, to ensure UBWE/CMLM agreement, loss function $\mathcal{L}_{agreement}$ is added during the UNMT training process. The entire UNMT loss function is reformulated as follows:

$$\mathcal{L}_{entire} = \mathcal{L}_{deno} + \mathcal{L}_{back} + \lambda \mathcal{L}_{agreement}. \tag{4}$$

A. Training UNMT with UBWE Agreement

Regarding UBWE agreement, we propose two UBWE agreement methods, that is, UBWE agreement regularization and UBWE adversarial training, to help UNMT and UBWE to interact during UNMT training to improve translation performance. The architecture of UNMT with UBWE agreement is shown in Fig. 5 and 6.

1) UBWE Agreement Regularization: UBWE agreement regularization was induced during back-translation based on the existing UNMT architecture to preserve UBWE accuracy between the encoder and decoder embeddings over the course of UNMT training as shown in Fig. 5. UBWE accuracy was measured using the similarity function Function(L_1, L_2) of the encoder and decoder embeddings. The objective function $\mathcal{L}_{agreement}$ is formulated as follows:

$$\begin{aligned}\mathcal{L}_{agreement} &\triangleq \mathcal{L}_{AR} \\ &= \text{Function}(L_1, L_2) \\ &= \text{Function}(enc_{L_1}, dec_{L_2}) \\ &+ \text{Function}(enc_{L_2}, dec_{L_1}),\end{aligned} \tag{5}$$

where all encoder L_1 and L_2 embeddings are denoted by enc_{L_1} and enc_{L_2}, respectively, and all decoder L_1 and L_2 embeddings are denoted by dec_{L_1} and dec_{L_2}, respectively. Thus, the similarity between the encoder and decoder word embeddings is formulated by applying the cosine similarity measure as

$$\begin{aligned}&\text{Function}(enc_{L_1}, dec_{L_2}) \\ &\approx \frac{1}{|Dict|}\sum_{i}^{|Dict|}(1 - \cos(enc_{x_i}, dec_{y_i})),\end{aligned} \tag{6}$$

$$\begin{aligned}&\text{Function}(enc_{L_2}, dec_{L_1}) \\ &\approx \frac{1}{|Dict|}\sum_{j}^{|Dict|}(1 - \cos(enc_{y_j}, dec_{x_j})),\end{aligned} \tag{7}$$

where the word x_i embedding in encoder L_1 is denoted by enc_{x_i} and the word y_i embedding in decoder L_2 is denoted by dec_{y_i}. The word y_j embedding of encoder L_2 is denoted by enc_{y_j} and the word y_j embedding of decoder L_1 is denoted by dec_{x_j}. $|Dict|$ denotes the dictionary size. Other similarity measures, such as the Euclidean distance, can also be applied to our proposed UBWE agreement regularization method.

Before calculating Function(enc_{L_1}, dec_{L_2}) and Function(enc_{L_2}, dec_{L_1}), we created a synthetic word-pair dictionary to evaluate UBWE accuracy over the entire UNMT training process because we did not have a test or development dataset to use as a bilingual dictionary. Cross-domain similarity local scaling (CSLS) [14] was used to evaluate UBWE accuracy; it can also be considered as the similarity between source and target word embeddings:

$$\begin{aligned}\text{CSLS}(x_i, y_i) &= 2 \cdot \cos(enc_{x_i}, dec_{y_i}) \\ &- r(x_i) - r(y_i),\end{aligned} \tag{8}$$

$$r(x_i) = \frac{1}{K}\sum_{y \in \mathcal{N}(x_i)} \cos(enc_{x_i}, dec_y), \tag{9}$$

$$r(y_i) = \frac{1}{K}\sum_{x \in \mathcal{N}(y_i)} \cos(enc_x, dec_{y_i}), \tag{10}$$

where the K nearest neighbor of the source word x_i is denoted by $y \in \mathcal{K}(x_i)$, and the K nearest neighbor of the source word y_i is denoted by $x \in \mathcal{K}(y_i)$. We chose a subset as a synthetic dictionary because the entire vocabulary was large. The most precise word pairs $\{x_i, y_i\}$ were selected as the synthetic dictionary $Dict_{x \to y}$ by ranking the CSLS scores. The same approach was used to achieve the opposite word pairs $Dict_{y \to x} = \{y_j, x_j\}$. The size of the two synthetic word-pair dictionaries was the same.

[2]In fact, the denoising auto-encoder produced the same effect as a language model trained in the encoder and decoder. CMLM training further enriched the source representation trained in the encoder.

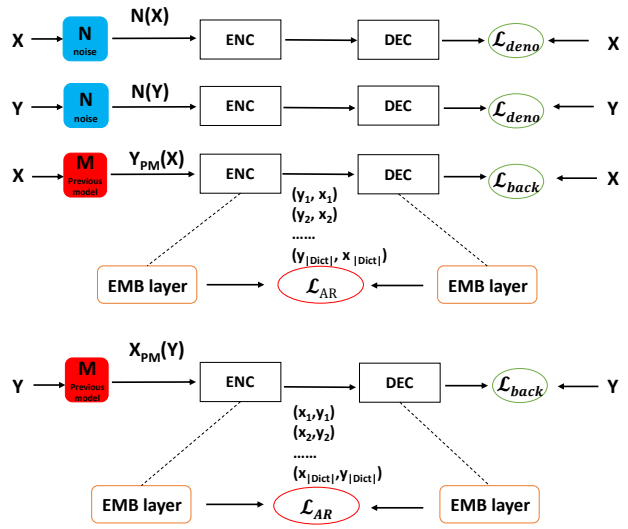


Fig. 5. Architecture of UNMT with UBWE agreement regularization.

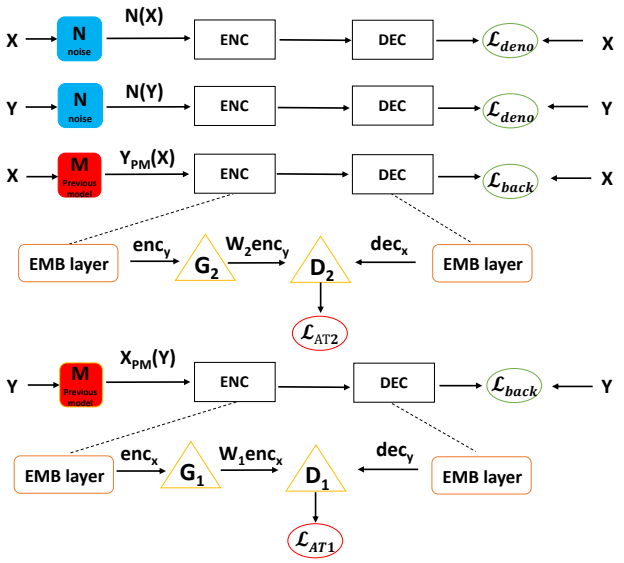


Fig. 6. Architecture of UNMT with UBWE adversarial training.

During back-translation training, we adopted the similarity function in Eq. 5 as UBWE agreement regularization. Note that before each epoch of the UNMT training process, the synthetic dictionary was regenerated dynamically.

2) UBWE Adversarial Training: A transformation matrix was used to project the source word embedding space to the target space in the existing UBWE methods. For UBWE adversarial training as shown in Fig. 6, an adversarial method was used to learn this transformation matrix [14]. The generator is computed as

$$G_1 = W_1 enc_x, \tag{11}$$

where enc_x denotes the source word embedding of the L_1 encoder, dec_y denotes the target word embedding of the L_2 decoder, and W_1 denotes the transformation matrix that projects enc_x to dec_y. Discriminator D_1 is a multilayer perceptron that outputs the probability that the word embeddings belong to a language. D_1 needs to learn how to distinguish which language $W_1 enc_x$ and dec_y come from. W_1 needs to learn to make $W_1 enc_x$ close to dec_y to confuse discriminator D_1; that is, the probability of selecting the precise language between samples generated by G_1 and original word embeddings is maximized to train D_1. $\log(1 - D_1(G_1(enc_x)))$ is minimized to train generator G_1. Therefore, $V(G_1, D_1)$ is a two-player minimax game [29] that is optimized as

$$\begin{aligned}\min_{G_1}\max_{D_1} V(G_1, D_1) &= \mathbb{E}_{dec_y}[\log D_1(dec_y)] \\ &+ \mathbb{E}_{enc_x}[\log(1 - D_1(G_1(enc_x)))].\end{aligned} \tag{12}$$

G_2 and D_2 are similar to G_1 and D_1, respectively. Generator G_1 and discriminator D_1 are formulated as:

$$\begin{aligned}\mathcal{L}_{G_1} &= \mathbb{E}_{enc_x}[-\log(D_1(W_1 enc_x))] \\ &+ \mathbb{E}_{dec_y}[-\log(1 - D_1(dec_y)],\end{aligned} \tag{13}$$

$$\begin{aligned}\mathcal{L}_{D_1} &= \mathbb{E}_{enc_x}[-\log(1 - D_1(W_1 enc_x))] \\ &+ \mathbb{E}_{dec_y}[-\log(D_1(dec_y)].\end{aligned} \tag{14}$$

Generator G_2 and discriminator D_2 are formulated as:

$$\begin{aligned}\mathcal{L}_{G_2} &= \mathbb{E}_{enc_y}[-\log(D_2(W_2 enc_y))] \\ &+ \mathbb{E}_{dec_x}[-\log(1 - D_2(dec_x)],\end{aligned} \tag{15}$$

$$\begin{aligned}\mathcal{L}_{D_2} &= \mathbb{E}_{enc_y}[-\log(1 - D_2(W_2 enc_y))] \\ &+ \mathbb{E}_{dec_x}[-\log(D_2(dec_x)],\end{aligned} \tag{16}$$

where enc_y denotes the source word embedding of the L_2 encoder, dec_x denotes the target word embedding of the L_1 decoder, and W_2 denotes the transformation matrix that projects enc_y to dec_x. After UBWE adversarial training is induced in UNMT, the $\mathcal{L}_{agreement}$ loss function is formulated as

$$\mathcal{L}_{agreement} \triangleq \mathcal{L}_{AT} = \mathcal{L}_{AT1} + \mathcal{L}_{AT2}, \tag{17}$$

where $\mathcal{L}_{AT1} = \mathcal{L}_{D_1} + \mathcal{L}_{G_1}$ and $\mathcal{L}_{AT2} = \mathcal{L}_{D_2} + \mathcal{L}_{G_2}$. During back-translation training, our proposed $L_{agreement}$ (L_{AR} or L_{AT}) are added into L_{entire} in Eq. 4, as shown in Figs. 5 and 6.

B. Training UNMT with CMLM Agreement

Regarding CMLM agreement, we propose two CMLM agreement mechanisms, that is, CMLM agreement regularization and CMLM knowledge distillation, to enable UNMT and CMLM to interact during training to improve translation performance.

1) CMLM Agreement Regularization: We induce CMLM agreement regularization based on the existing UNMT architecture during back-translation to enrich source representation during UNMT training. The architecture of UNMT with CMLM agreement mechanisms is illustrated in Fig. 7.

Inspired by Devlin *et al.* [30], and Lample and Conneau [18], we propose training a masked language model during UNMT training. Following Devlin *et al.* [30], we randomly sampled 15% of all tokens in each sentence. We replaced

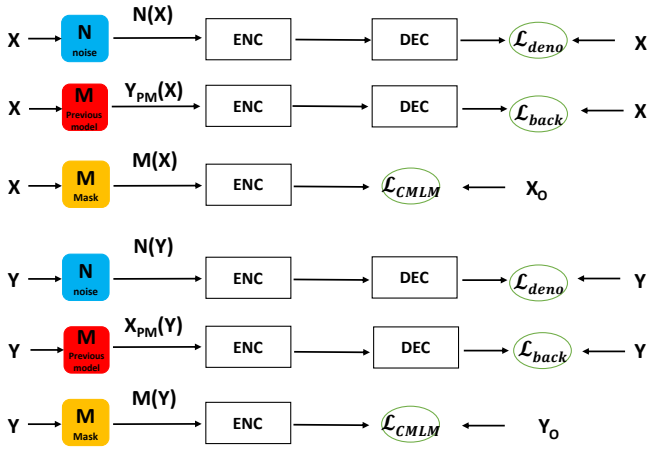


Fig. 7. Architecture of UNMT with CMLM agreement regularization.

80% of these sampled tokens with the [MASK] token, and replaced 10% of these sampled tokens with a random word. The remaining 10% of tokens remained unchanged.

The language model, more precisely the encoder of the UNMT model, was optimized by minimizing the objective function:

$$\begin{aligned}\mathcal{L}_{agreement} &\triangleq \mathcal{L}_{CMLM} \\ &= \mathbb{E}_{X\sim\phi_{L_1}}[-\log P_{L_1}(X_O|M(X)] \\ &+ \mathbb{E}_{Y\sim\phi_{L_2}}[-\log P_{L_2}(Y_O|M(Y)],\end{aligned} \tag{18}$$

where $M(X)$ and $M(Y)$ are X and Y with masked tokens, and X_O and Y_O denote the original tokens before masking, respectively. P_{L_1} (P_{L_2}) denotes the language model prediction probability in language L_1 (L_2) that encodes sentences with some masked tokens $M(X)$ ($M(Y)$) and predicts these masked tokens with the same encoder in the same language.

2) CMLM Knowledge Distillation: UNMT achieved much better performance after the introduction of CMLM pre-training. However, the existing CMLM did not further participate in UNMT training as a language model, but only in the initialization of UNMT. We propose introducing the existing CMLM into UNMT training through the knowledge distillation method [31], [32]. We propose teaching the student language model (the encoder of UNMT) training described in Section IV-B1 using the existing CMLM as a teacher. We first describe the strategy of knowledge distillation in language model training and then introduce language model training as CMLM agreement regularization into UNMT training. The student language model (the encoder of UNMT) not only matches the output of the ground-truth masked tokens but also matches the probability output of the teacher language model. The one-hot target has a small information entropy. Conversely, the soft probability distribution of the target distilled by the teacher model contains a large amount of information about the relationship between different classes. We can apply knowledge distillation to improve CMLM training. Thus, this language model was optimized by minimizing the objective function:

$$\begin{aligned}\mathcal{L}_{agreement} &\triangleq \mathcal{L}_{KD} \\ &= \mathcal{L}_{KD_X} + \mathcal{L}_{KD_Y},\end{aligned} \tag{19}$$

$$\begin{aligned}\mathcal{L}_{KD_X} &= (1-\alpha)\cdot\mathbb{E}_{X\sim\phi_{L_1}}[-\log P_{L_1}(X_O|M(X)] \\ &+ \alpha\cdot T^2\cdot\mathbb{E}_{X\sim\phi_{L_1}}[KL(X_S, X_T)],\end{aligned} \tag{20}$$

$$\begin{aligned}\mathcal{L}_{KD_Y} &= (1-\alpha)\cdot\mathbb{E}_{Y\sim\phi_{L_2}}[-\log P_{L_2}(Y_O|M(Y)] \\ &+ \alpha\cdot T^2\cdot\mathbb{E}_{Y\sim\phi_{L_2}}[KL(Y_S, Y_T)],\end{aligned} \tag{21}$$

where α is a hyper-parameter that adjusts the weight of two loss functions. T is the temperature. A higher temperature causes a softer probability distribution to be obtained. $KL(\cdot)$ denotes KL divergence. X_S and X_T denote the soft probability distribution of L_1 tokens of the student and teacher after encoding masked L_1 tokens, respectively. Y_S and Y_T denote the soft probability distribution of L_2 tokens of the student and teacher after encoding masked L_2 tokens, respectively. $\mathcal{L}_{KD_X}$ ($\mathcal{L}_{KD_Y}$) denotes the L_1 (L_2) language model loss function.

V. Experiments

A. Datasets

To make our experiments comparable with previous work, we considered four language pairs: English-German (En-De), English-French (En-Fr), German-Czech (De-Cs), and English-Japanese (En-Ja). For the first three language pairs, 30 million sentences from the WMT monolingual News Crawl datasets were extracted to train the UNMT systems for each language. Note that, the De-Cs translation direction, which was the first time introduced in WMT2019, is the only WMT unsupervised translation task to date. En-Ja is a distant language pair. It is substantially more difficult to train UBWE on distant language pairs than on similar European language pairs [33]. Furthermore, English and Japanese are in distinct language families and have different word orderings. Thus, UNMT performed particularly poorly on En-Ja if only monolingual corpora are used[3]. As a result, we used shuffled parallel corpora which were 3.0 million ASPEC corpus sentence pairs for En-Ja to construct simulated experiments.

WMT newstest2014 for En-Fr, WMT newstest2016 for En-De, WMT newstest2019 for De-Cs, and WAT-2018 ASPEC testset for En-Ja were used to evaluate UNMT performance.

B. UBWE Settings

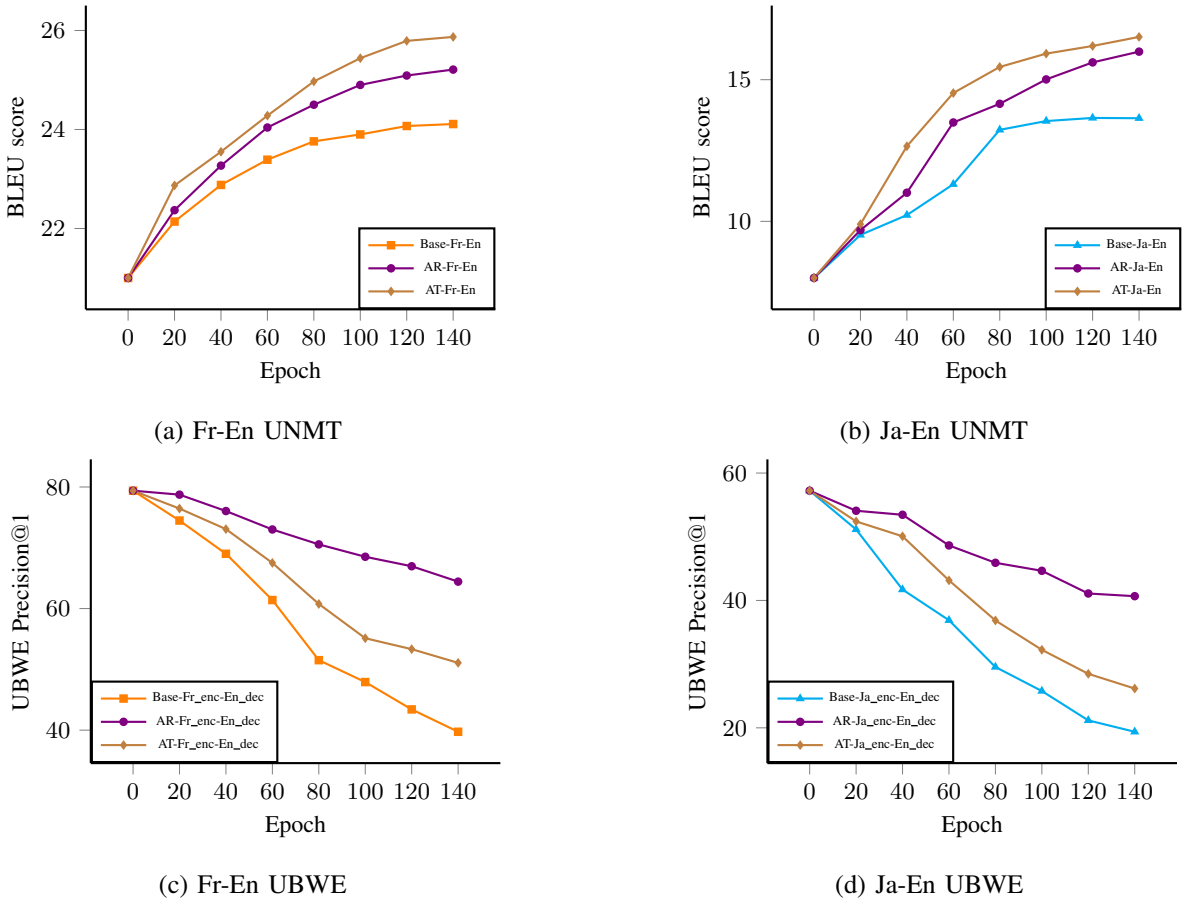
In the training process for UBWE, the fastText toolkit[4] [34] (default settings) was used to train word embeddings on the monolingual corpora mentioned above for each language. Before bilingual projection, the word embeddings were normalized to unit length and mean centered. Two monolingual word embeddings were projected into one space using VecMap[5] [15].

The word translation precision in the MUSE test set using the first-ranked predicted candidate was selected as the criterion to measure UBWE quality.

[3]The results on En-Ja pure monolingual corpora were analyzed in Section VI-A.

[4]https://github.com/facebookresearch/fastText

[5]https://github.com/artetxem/vecmap



(a) Fr-En UNMT (b) Ja-En UNMT

(c) Fr-En UBWE (d) Ja-En UBWE

Fig. 8. Trends in the BLEU score and UBWE quality for the baseline (Base), UBWE agreement regularization (AR), and UBWE adversarial training (AT) during UNMT training on the Fr-En and Ja-En datasets.

C. UNMT Settings

For UNMT training, we chose two baselines: UNMT with UBWE pre-training (UNMT-UBWE baseline) and UNMT with CMLM pre-training (UNMT-CMLM baseline).

1) UNMT-UBWE Baseline Settings: For the UNMT-UBWE baseline, the transformer-based UNMT toolkit[6] was used to train UNMT systems and we followed the settings of Lample *et al*. [19]. Specifically, four encoder and decoder layers were used and three encoder and decoder layers were shared for both languages. The dimension size of hidden layers was 512. The batch size for training UNMT systems was 32 and the Adam optimizer [35] was used to optimize the UNMT model parameters. The initial learning rate was 0.0001 and $\beta_1 = 0.5$. Two monolingual corpora were concatenated to achieve a 60K vocabulary. Because we had to evaluate UBWE quality more accurately, we did not adopt BPE for this baseline. Each UNMT system was trained on one P100 GPU for 140 epochs (approximately 500K iterations)[7].

2) UNMT-CMLM Baseline Settings: For the UNMT-CMLM baseline, we used a transformer-based UNMT[8] and followed the settings of UNMT [18] because it has been shown to achieve state-of-the-art results: six layers for the encoder and decoder. The dimension of the hidden layers was 1024. The Adam optimizer [35] was used to optimize the model parameters. The initial learning rate was 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. To apply the existing language model, we used a shared vocabulary for both languages with 60K subword tokens based on BPE [39] for this baseline. The cross-lingual language model was used to pre-train the encoder and decoder of the entire UNMT model. For the Fr-En language pair, we used the language model and vocabulary provided by Lample and Conneau [18]. To investigate the effect of CMLM quality on UNMT performance, we also trained a language model for the Fr-En language pair. For the De-Cs language pair, that is, the WMT2019 unsupervised machine translation task, we first trained the language model before training the UNMT model. For the CMLM knowledge distillation method, α was set to 0.1 and T was set to 2. We performed 80 epochs[9] (approximately 120K iterations) to train each model on eight V100 GPUs.

We used the case-sensitive BLEU score calculated using the $multi-bleu.perl$ script from `Moses`[10] to evaluate UNMT performance. We followed the strategy mentioned by Lample *et al*. [17] to select the final model. We computed the BLEU

[6]https://github.com/facebookresearch/UnsupervisedMT

[7]Note that the definition of a UNMT epoch is different from that of an NMT epoch. The setting in Lample *et al*. [19]'s toolkit, for which there were 3,500 iterations in one epoch, was used to train the UNMT systems.

[8]https://github.com/facebookresearch/XLM

[9]We followed the settings in Lample and Conneau [18]'s toolkit, that is, 1500 iterations as one epoch.

[10]https://github.com/moses-smt/mosesdecoder

TABLE II
PERFORMANCE (BLEU SCORE) OF UNMT WITH UBWE AGREEMENT.

Method	Fr-En	En-Fr	De-En	En-De	Ja-En	En-Ja
Artetxe *et al*. [16]	15.56	15.13	n/a	n/a	n/a	n/a
Lample *et al*. [17]	14.31	15.05	13.33	9.64	n/a	n/a
Yang *et al*. [36]	15.58	16.97	14.62	10.86	n/a	n/a
Lample *et al*. [19]	24.20	25.10	21.00	17.20	n/a	n/a
UNMT-BWE Baseline	24.50	25.37	21.23	17.06	14.09	21.63
+ UBWE agreement regularization	25.21++	27.86++	22.38++	18.04++	16.36++	23.01++
+ UBWE adversarial training	**25.87++**	**28.38++**	**22.67++**	**18.29++**	**17.22++**	**23.64++**

Note: "++" after a score denotes that our proposed method was much better than the original UNMT system with significance level p <0.01 [37].

TABLE III
PERFORMANCE (BLEU SCORE) OF UNMT WITH CMLM AGREEMENT.

Method	Fr-En	En-Fr	De-Cs	Cs-De	Ja-En	En-Ja
Lample and Conneau [18]	33.30	33.40	n/a	n/a	n/a	n/a
Single UNMT model of the WMT2019 first-ranked team [38]	n/a	n/a	15.50	n/a	n/a	n/a
UNMT-MLM Baseline	34.38	36.63	16.10	15.80	20.45	32.66
+ CMLM agreement regularization	35.14++	37.65++	17.50++	17.00++	21.01	33.35
+ CMLM knowledge distillation	**35.38++**	**37.87++**	**17.80++**	**17.50++**	**21.55++**	**33.87++**

score between the original L_1 (L_2) monolingual sentences and their reconstructions $L_1 \rightarrow L_2 \rightarrow L_1$ ($L_2 \rightarrow L_1 \rightarrow L_2$). Finally, the model that had the best average BLEU score in both translation directions was chosen. At the beginning of training, both our proposed UBWE agreement and CMLM agreement were induced as objective functions. In Section VI, we discuss the selection of some parameters.

D. UNMT Performance with UBWE Agreement

The trends in the BLEU score and UBWE quality during UNMT training on En-Fr and En-Ja are shown in Figure 8. Our findings are as follows:

1) UBWE quality decreased during UNMT training for all the UNMT systems. This is in agreement with our findings in Section III.

2) UBWE quality in the UNMT systems using our proposed two UBWE agreement methods decreased more slowly than in the conventional UNMT baseline system. This indicates that both our proposed UBWE agreement methods significantly alleviated the degradation of UBWE quality.

3) Compared with UBWE adversarial training, UBWE agreement regularization was better at alleviating the degradation of UBWE quality.

The detailed BLEU scores for the UNMT systems on the En-Fr, En-De, and En-Ja test sets are presented in Table II. Our observations are as follows:

1) Our re-implemented baseline was a strong UNMT system. It achieved similar performance to the state-of-the-art approach of Lample *et al*. [19].

2) For all language pairs, our proposed UBWE agreement approaches achieved 1–3 BLEU scores more than the original UNMT baseline system.

3) Regarding the two UBWE agreement methods, UBWE adversarial training achieved slightly better translation performance than UBWE agreement regularization, although UBWE agreement regularization better preserved UBWE quality. This may be because UBWE adversarial training has more interaction with the UNMT system because it could jointly train with the UNMT system. However, we only added UBWE agreement regularization to the UNMT training objective to maintain UBWE accuracy.

E. UNMT Performance with CMLM agreement

Figure 9 shows the trend in the BLEU score during UNMT training on Fr-En and En-Fr and perplexity of En and Fr monolingual data. Our observations are as follows:

1) For both En and Fr, the perplexity increased during UNMT training. This is in agreement with our findings in Section III.

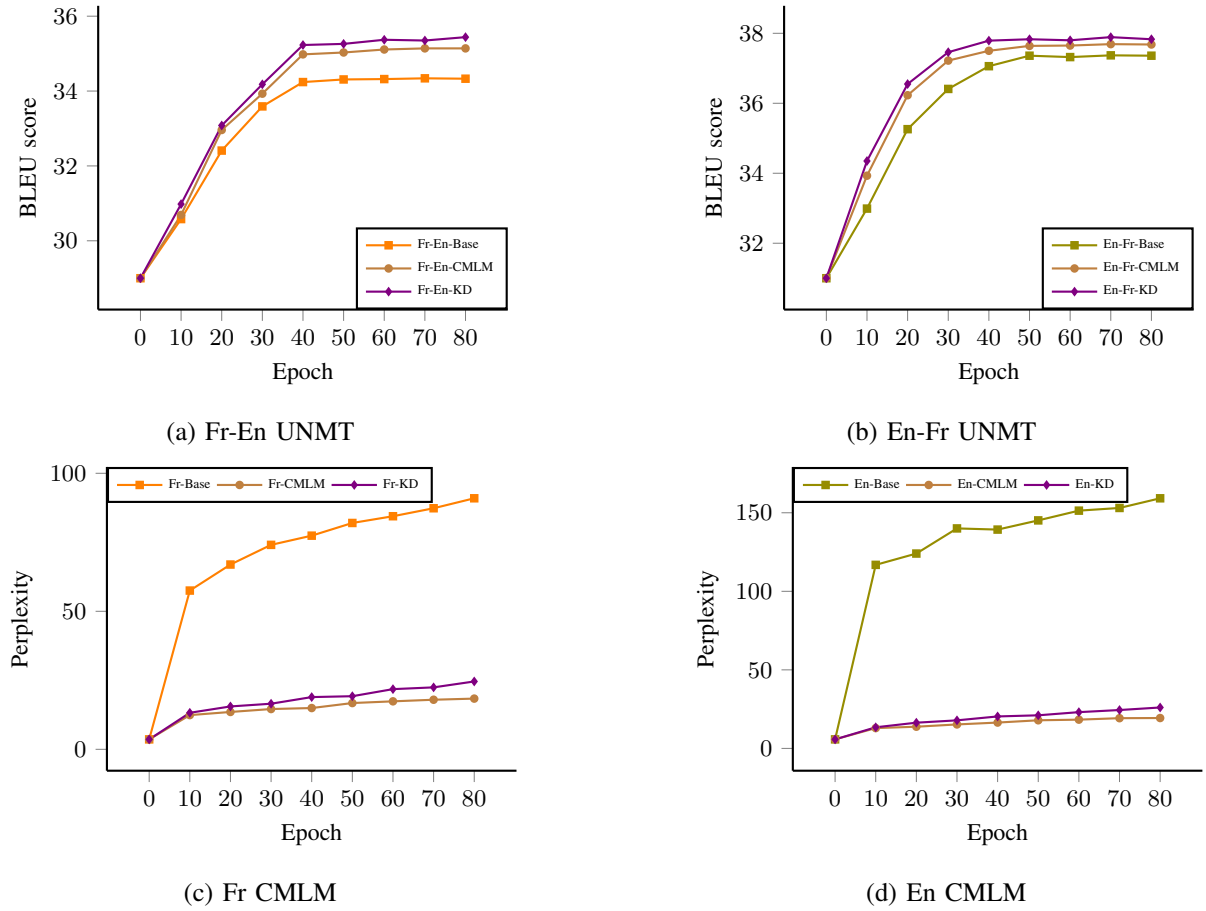
2) For the system with CMLM agreement regularization and CMLM knowledge distillation, the perplexity was much lower than that in the conventional UNMT baseline system. This indicates that our proposed CMLM agreement methods significantly alleviated high perplexity.

3) Regarding two CMLM agreement methods, CMLM agreement regularization was better at mitigating high perplexity than CMLM knowledge distillation.

The detailed BLEU scores for the UNMT systems on the En-Fr and De-Cs test sets are presented in Table III. Our observations are as follows:

1) Our re-implemented UNMT baseline performed better than the state-of-the-art method of Lample and Conneau [18] on En-Fr and the single UNMT model of the WMT2019 first-ranked team[11] on De-Cs. This indicates that our re-implemented baseline was a strong UNMT system.

[11]http://matrix.statmt.org/systems/show/4346



(a) Fr-En UNMT (b) En-Fr UNMT

(c) Fr CMLM (d) En CMLM

Fig. 9. Trends in the BLEU score and CMLM quality for the baseline (Base), CMLM agreement regularization (CMLM), and CMLM knowledge distillation (KD) during UNMT training on Fr-En.

2) For all three language pairs, our proposed CMLM agreement methods significantly outperformed the original UNMT baseline system by approximately 1 BLEU score.

3) Regarding the two CMLM agreement methods, CMLM knowledge distillation performed slightly better than CMLM agreement regularization by the BLEU score, although CMLM agreement regularization was better at maintaining low perplexity. This may be because introducing a existing language model as a teacher through the CMLM knowledge distillation method achieved better source representation over the course of UNMT training compared with the CMLM agreement regularization method.

VI. Discussion

A. Distant Language Pair Analysis

We also conducted our proposed UBWE agreement approach on the pure monolingual corpora for En-Ja. We used the same WMT monolingual News Crawl dataset described in Section V-A for En and 20M in-house news-domain monolingual corpus for Ja. For the En-Ja testset, we employed experienced translators with a linguistic background to annotate the Ja reference based on the newstest2012 and newstest2013 En reference. This new En-Ja testset will be released soon.

TABLE IV
Performance (BLEU score) of UNMT with UBWE agreement on different Ja-En datasets.

Method	newstest2012		newstest2013	
	Ja-En	En-Ja	Ja-En	En-Ja
UNMT-BWE Baseline	1.85	2.35	1.76	2.73
+ UBWE agreement regularization	2.19	2.67	2.09	3.10
+ UBWE adversarial training	2.37	2.69	2.29	3.12

As shown in Table IV, the baseline system performed particularly poorly on the Ja-En translation task. Our proposed UBWE agreement approaches also achieved approximately 0.4 BLEU score more than the original UNMT baseline system. This indicates that our proposed agreement methods are robust.

B. Effect of the Dictionary Size

The size of the synthetic word-pair dictionary and different similarity measures for the proposed UBWE agreement regularization method were evaluated on the En-Fr task. As shown in Table V, regardless of the synthetic word-pair dictionary size and similarity measures, the proposed method achieved better performance than the original UNMT baseline system. This demonstrates that our proposed agreement regularization method was effective. Regarding two similarity

measures, the cosine similarity measure was slightly better than the Euclidean distance according to the BLEU score.

TABLE V
EFFECT ON THE DICTIONARY SIZE AND DIFFERENT SIMILARITY MEASURES

Dict Size	Cosine		Euclidean Distance	
	Fr-En BLEU	En-Fr BLEU	Fr-En BLEU	En-Fr BLEU
Baseline	24.50	25.37	24.50	25.37
20K	25.15	27.18	25.06	27.31
10K	25.10	27.48	25.06	27.51
5K	25.14	27.58	25.17	27.44
3K	25.21	27.86	25.26	27.57
1K	25.25	27.40	25.18	27.21
500	25.13	27.07	25.04	26.89

Based on the cosine similarity measure, the relationship between UBWE accuracy and dictionary size was also investigated. The larger the dictionary size, the slower the degradation of UBWE quality, as shown in Fig. 10. This indicates that a larger dictionary size could achieve better UBWE agreement. However, as Table V demonstrates, a higher BLEU score could not be achieved by a larger dictionary size. The best translation performance was achieved by the UNMT model with a 3000 word-pair dictionary.

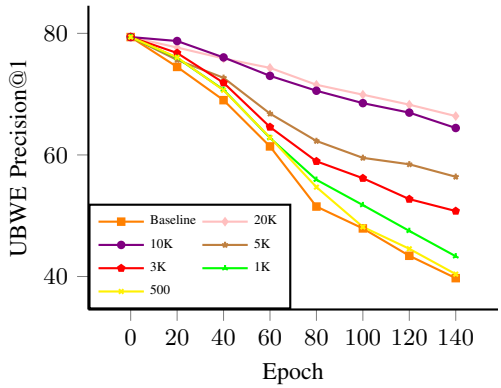


Fig. 10. UBWE accuracy with respect to dictionary size on the Fr-En test set during UNMT training.

C. Effect of Hyper-parameter λ

The influence of hyper-parameter λ in Eq. (4) on UNMT translation performance on the En-Fr task was empirically investigated for the UBWE adversarial training and CMLM knowledge distillation methods, as shown in Figs. 11 and 12. Over the course of UNMT training, the role of $\mathcal{L}_{agreement}$ was affected by the selection of λ. $\mathcal{L}_{agreement}$ played a less important role than other loss terms during UNMT training if λ was small. The larger the λ value, the more important the role of $\mathcal{L}_{agreement}$. As shown in Figs. 11 and 12, UNMT performance for our proposed agreement methods improved for almost all λ values ranging from 0.01 to 10, and the best translation performance was achieved with a balanced $\lambda = 1$ for both our proposed methods.

D. Combination Study

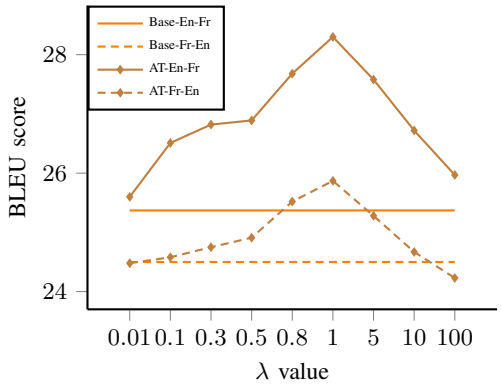


Fig. 11. Effect of hyper-parameter λ for the UBWE adversarial training (AT) model on the En $\leftrightarrow$ Fr dataset.

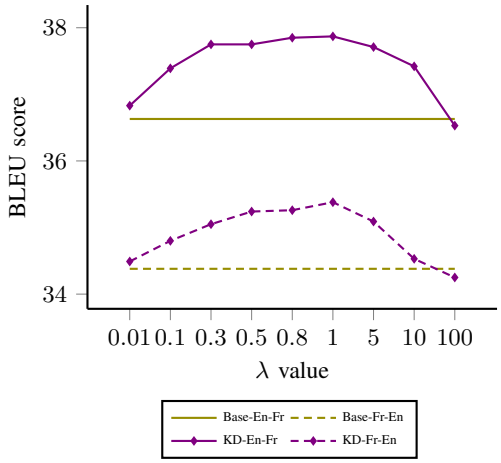


Fig. 12. Effect of hyper-parameter λ for CMLM knowledge distillation (KD) model on the En $\leftrightarrow$ Fr dataset.

We also investigated the performance of UNMT system using the combination of our proposed UBWE agreement and CMLM agreement method. In detail, we selected the UBWE adversarial training and CMLM knowledge distillation methods, respectively. We conducted the experiments based on CMLM pre-training UNMT system on the En-Fr and En-Ja language pairs.

TABLE VI
PERFORMANCE (BLEU SCORE) OF UNMT WITH OUR PROPOSED AGREEMENT METHODS

Method	Fr-En	En-Fr	Ja-En	En-Ja
UNMT-MLM Baseline	34.38	36.63	20.45	32.66
+ UBWE adversarial training	34.97	37.56	20.99	33.41
+ CMLM knowledge distillation	35.38	37.87	21.55	33.87
+ Combination of both methods	35.48	38.14	21.64	34.17

As shown in Table VI, both the UBWE adversarial training and CMLM knowledge distillation methods outperformed the UNMT baseline system. The CMLM knowledge distillation method achieved a larger improvement in terms of the BLEU score. Moreover, the combination of both agreement methods further improved the translation performance.

E. Comparison with Word-based CMLM Pre-training

We also investigated the difference between word-based and BPE-based CMLM pre-training. The UNMT baseline system based on words performed similarly to the baseline based on BPE. Regardless of whether the BPE-based or word-based UNMT system was used, our proposed CMLM agreement methods significantly outperformed the corresponding baseline for both language pairs by approximately 1 BLEU score, as shown in Table VII. This indicates that our proposed CMLM agreement methods are robust.

TABLE VII
PERFORMANCE (BLEU SCORE) OF UNMT WITH DIFFERENT LEVELS OF TOKENS.

Method	BPE		word	
	Fr-En	En-Fr	Fr-En	En-Fr
UNMT-MLM Baseline	34.38	36.63	34.32	36.38
+ CMLM agreement regularization	35.14	37.65	35.19	37.21
+ CMLM knowledge distillation	35.38	37.87	35.36	37.43

F. Efficiency

The efficiency of our proposed agreement methods was analyzed and the results are presented in Table VIII. The number of parameters did not increase with UBWE agreement regularization, CMLM agreement regularization, and CMLM knowledge distillation. Compared with existing parameters, only a few parameters were added using the BWE adversarial training method. These proposed methods were trained almost at the same speed. Moreover, UNMT decoding was not affected by our proposed agreement methods. Therefore, the speed of UNMT training was not affected by our cross-lingual language representation agreement methods.

TABLE VIII
ANALYSIS OF PARAMETERS AND TRAINING SPEED (NUMBER OF PROCESSED WORDS PER SECOND ON ONE P100 ON $\#1 \sim 3$ AND NUMBER OF PROCESSED WORDS PER SECOND ON ONE V100 ON $\#4 \sim 6$).

#	Method	Parameters	Speed
1	UNMT-BWE Baseline	120,141K	3784
2	+UBWE agreement regularization	120,141K	3741
3	+UBWE adversarial training	120,764K	3733
4	UNMT-MLM Baseline	304,903K	3327
5	+ CMLM agreement regularization	304,903K	3308
6	+ CMLM knowledge distillation	304,903K	3184

VII. RELATED WORK

Supervised BWE takes advantage of similarities between the source and target language to learn a linear transformation matrix [2]. However, an essential practical issue related to the lack of a bilingual dictionary arises for many language pairs. UBWE has attracted great interest. A self-learning framework was proposed to train BWE with a 25-word bilingual dictionary, or even no dictionary [11], [15]. A generative adversarial network was induced to learn UBWE [13], [14].

Recently, UNMT relying on non-parallel monolingual corpora [16], [17] was proposed through UBWE [14], [15] pre-training and other mechanisms, such as denoising and back-translation. The source sentences were encoded to achieve a shared latent representation using a shared encoder. Two decoders for every language were used by Artetxe *et al*. [16] and a single shared decoder was used by Lample *et al*. [17]. A weight-sharing mechanism between two independent encoders instead of a shared encoder was proposed by Yang *et al*. [36]. It preserved the uniqueness and intrinsic characteristics of each language. Lample *et al*. [19] concatenated two monolingual corpora as one monolingual corpus and used this monolingual embedding to initialize the UNMT model to achieve impressive performance on several similar language pairs.

Moreover, Lample *et al*. [19] and Artetxe *et al*. [40] used unsupervised statistical machine translation (USMT). The performance of unsupervised machine translation was further improved by combining UNMT and USMT [41]–[43].

More recently, language model pre-training has been shown to be effective for improving many natural language processing tasks [30], [44], [45]. Lample and Conneau [18] initialized the UNMT model with a pre-training cross-lingual language model to achieve state-of-the-art UNMT performance. It has been demonstrated that supervised machine translation benefits from a language model based on neural network [46]–[48]. However, language model training was not introduced into UNMT training in previous works. In this study, we jointly trained a language model and UNMT model to improve UNMT performance.

Knowledge distillation [31], [32] refers to a method that uses a more complex teacher model that has been trained to guide a smaller student model training, thereby reducing the model size and computing resources while maintaining the accuracy of the original teacher model. It has been proven to be effective for improving NMT [49]–[51]. The existing translation model as a teacher was introduced to guide a student translation model in these studies. In the present study, we introduced the existing language model as a teacher to enrich source representation to improve UNMT performance.

VIII. CONCLUSION

Unsupervised cross-lingual language representation initialization is a fundamental mechanism of UNMT. In existing approaches, the pre-trained UBWE or CMLM is only used to initialize the parameters of UNMT before UNMT training. In this work, we found that not only at the UNMT initialization step but also during the training period, the quality of UBWE or CMLM had a significant influence on UNMT performance. Based on this empirical finding, we proposed several methods to train UNMT with UBWE or CMLM agreement. The experimental results on several language pairs demonstrated that our proposed UBWE/CMLM agreement methods alleviated the degradation of UBWE/CMLM quality. Both of our proposed agreement methods significantly improved UNMT performance.

REFERENCES

[1] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, and T. Zhao, "Unsupervised bilingual word embedding agreement for unsupervised neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 1235–1245.

[2] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *CoRR*, vol. abs/1309.4168, 2013.

[3] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014, pp. 462–471.

[4] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized word embedding and orthogonal transform for bilingual word translation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, 2015, pp. 1006–1011.

[5] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *Proceedings of the Third International Conference on Learning Representations*, San Diego, California, 2015.

[6] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, "Deep multilingual correlation for improved word embeddings," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, 2015, pp. 250–256.

[7] R. Wang, H. Zhao, S. Ploux, B. Lu, and M. Utiyama, "A bilingual graph-based semantic model for statistical machine translation," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, USA, 2016, pp. 2950–2956.

[8] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 2016, pp. 2289–2294.

[9] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," in *Proceedings of the Fifth International Conference on Learning Representations*, Toulon, France, 2017.

[10] R. Wang, H. Zhao, S. Ploux, B. Lu, M. Utiyama, and E. Sumita, "Graph-based bilingual word embedding for statistical machine translation," *ACM Trans. Asian & Low-Resource Lang. Inf. Process.*, vol. 17, no. 4, pp. 31:1–31:23, 2018.

[11] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 451–462.

[12] H. Cao, T. Zhao, S. Zhang, and Y. Meng, "A distribution-based model to learn bilingual word embeddings," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 2016, pp. 1818–1827.

[13] M. Zhang, Y. Liu, H. Luan, and M. Sun, "Adversarial training for unsupervised bilingual lexicon induction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017.

[14] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada, April 2018.

[15] M. Artetxe, G. Labaka, and E. Agirre, "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)*, Melbourne, Australia, 2018, pp. 789–798.

[16] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada, April 2018.

[17] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," in *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada, 2018.

[18] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019.

[19] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 5039–5049.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

[21] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao, "Neural machine translation with source dependency representation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 2846–2852.

[22] K. Chen, T. Zhao, M. Yang, and L. Liu, "Translation prediction with source dependency-based context representation," in *AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017, pp. 3166–3172.

[23] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, "Syntax-directed attention for neural machine translation," in *AAAI Conference on Artificial Intelligence*, New Orleans, Lousiana, USA, 2018, pp. 4792–4798.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[26] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego California, USA, 2016, pp. 1367–1377.

[27] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W. Ma, "Dual learning for machine translation," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, Barcelona, Spain, 2016, pp. 820–828.

[28] R. Sennrich, B. Haddow, and A. Birch, "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)*, Berlin, Germany, 2016.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014, pp. 2672–2680.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186.

[31] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 2006, pp. 535–541.

[32] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[33] A. Søgaard, S. Ruder, and I. Vulić, "On the limitations of unsupervised bilingual dictionary induction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018, pp. 778–788.

[34] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the Third International Conference on Learning Representations*, San Diego, California, 2015.

[36] Z. Yang, W. Chen, F. Wang, and B. Xu, "Unsupervised neural machine translation with weight sharing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018, pp. 46–55.

[37] M. Collins, P. Koehn, and I. Kučerová, "Clause restructuring for statistical machine translation," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*,

Ann Arbor, Michigan, Jun. 2005, pp. 531–540. [Online]. Available: https://www.aclweb.org/anthology/P05-1066

[38] B. Marie, H. Sun, R. Wang, K. Chen, A. Fujita, M. Utiyama, and E. Sumita, "NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy, Aug. 2019, pp. 294–301. [Online]. Available: https://www.aclweb.org/anthology/W19-5330

[39] R. Sennrich, B. Haddow, and A. Birch, "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)*, Berlin, Germany, 2016, pp. 1715–1725.

[40] M. Artetxe, G. Labaka, and E. Agirre, "Unsupervised statistical machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3632–3642.

[41] B. Marie and A. Fujita, "Unsupervised neural machine translation initialized by unsupervised statistical machine translation," *CoRR*, vol. abs/1810.12703, 2018.

[42] S. Ren, Z. Zhang, S. Liu, M. Zhou, and S. Ma, "Unsupervised neural machine translation with SMT as posterior regularization," *CoRR*, vol. abs/1901.04112, 2019.

[43] M. Artetxe, G. Labaka, and E. Agirre, "An effective approach to unsupervised machine translation," *CoRR*, vol. abs/1902.01313, 2019.

[44] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, Jun. 2018, pp. 2227–2237.

[45] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018.

[46] R. Wang, H. Zhao, B.-L. Lu, M. Utiyama, and E. Sumita, "Neural network based bilingual language model growing for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 189–195.

[47] R. Wang, M. Utiyama, A. Finch, L. Liu, K. Chen, and E. Sumita, "Sentence selection and weighting for neural machine translation domain adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1727–1741, Oct 2018.

[48] B. Marie, R. Wang, A. Fujita, M. Utiyama, and E. Sumita, "Nict's neural and statistical machine translation systems for the wmt18 news translation task," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 453–459.

[49] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016, pp. 1317–1327.

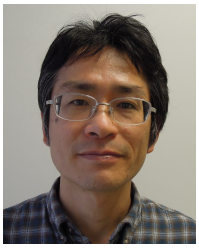
[50] M. Freitag, Y. Al-Onaizan, and B. Sankaran, "Ensemble distillation for neural machine translation," *CoRR*, vol. abs/1702.01802, 2017.

[51] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T. Liu, "Multilingual neural machine translation with knowledge distillation," *CoRR*, vol. abs/1902.10461, 2019.

Haipeng Sun received his B.S. degree in mathematics and applied mathematics from Harbin Institute of Technology in 2013 and his M.S. degree in computer science from Harbin Institute of Technology in 2015. He has been a Ph.D candidate at Harbin Institute of Technology since 2015. He has been an internship research fellow at the National Institute of Information and Communications Technology, Japan since 2018. His research interests include machine translation and natural language processing.

Rui Wang is a tenure-track researcher at the National Institute of Information and Communications Technology, Japan. He received his B.S. degree from Harbin Institute of Technology in 2009, his M.S. degree from the Chinese Academy of Sciences in 2012, and his Ph.D. degree from Shanghai Jiao Tong University in 2016, all of which are in computer science. He was a joint Ph.D. at Centre Nationnal de la Recherche Scientifique, France in 2014. His research interests are machine translation and natural language processing.

Kehai Chen received his B.S. degree from Xi'an University of Technology in 2010, his M.S. degree from the University of Chinese Academy of Sciences in 2013, and his Ph.D. degree from Harbin Institute of Technology in 2018, all of which were in computer science. He has been a researcher at the National Institute of Information and Communications Technology, Japan since 2018. His research interests include machine translation and natural language processing.



Masao Utiyama is an executive researcher at the National Institute of Information and Communications Technology, Japan. He completed his doctoral dissertation at the University of Tsukuba in 1997. His main research field is machine translation.

Eiichiro Sumita received his B.S. and M.S. degrees in computer science from the University of Electro-Communications, Japan in 1980 and 1982, respectively and his Ph.D degree in engineering from Kyoto University, Japan, in 1999. He has been the Director of the Multilingual Translation Laboratory at the National Institute of Information and Communication Technology since 2006. He worked at Advanced Telecommunications Research Institute International from 1992 to 2009 and IBM Research-Tokyo from 1980 to 1991. His research interests include machine translation and e-learning.

Tiejun Zhao is a professor at the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include: natural language understanding, content-based web information processing, and applied artificial intelligence. He has published three academic books and 60 papers in journals and conference proceedings in the last 3 years. He has been a PC member at ACL, COLING over the last 5 years and was also appointed as an MT Track Co-chair for COLING 2014.