Unsupervised Bilingual Word Embedding Agreement for Unsupervised Neural Machine Translation

Haipeng Sun[1]*, **Rui Wang**[2], **Kehai Chen**[2],
Masao Utiyama[2], **Eiichiro Sumita**[2], **and Tiejun Zhao**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
hpsun@hit-mtlab.net, tjzhao@hit.edu.cn
{wangrui, khchen, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Unsupervised bilingual word embedding (UBWE), together with other technologies such as back-translation and denoising, has helped unsupervised neural machine translation (UNMT) achieve remarkable results in several language pairs. In previous methods, UBWE is first trained using non-parallel monolingual corpora and then this pre-trained UBWE is used to initialize the word embedding in the encoder and decoder of UNMT. That is, the training of UBWE and UNMT are separate. In this paper, we first empirically investigate the relationship between UBWE and UNMT. The empirical findings show that the performance of UNMT is significantly affected by the performance of UBWE. Thus, we propose two methods that train UNMT with UBWE agreement. Empirical results on several language pairs show that the proposed methods significantly outperform conventional UNMT.

1 Introduction

Since 2013, neural network based bilingual word embedding (BWE) has been applied to several natural language processing tasks (Mikolov et al., 2013; Faruqui and Dyer, 2014; Xing et al., 2015; Dinu et al., 2015; Lu et al., 2015; Wang et al., 2016; Artetxe et al., 2016; Smith et al., 2017; Wang et al., 2018). Recently, researchers have found that supervision is not always necessary (Cao et al., 2016; Zhang et al., 2017). Several unsupervised BWE (UBWE) methods (Conneau et al., 2018; Artetxe et al., 2018a) have been proposed and these have achieved impressive performance in word-translation tasks. The success of UBWE makes unsupervised neural machine translation (UNMT) possible. The combination of UBWE with denoising autoencoder and back-translation has led to UNMT that relies solely on monolingual corpora, with remarkable results reported for several language pairs such as English-French and English-German (Artetxe et al., 2018c; Lample et al., 2018a).

In previous methods, UBWE is first trained using non-parallel monolingual corpora. This pre-trained UBWE is then used to initialize the word embedding in the encoder and decoder of UNMT. That is, the training of UBWE and UNMT take place in separate steps. In this paper, we first empirically investigate the relationship between UBWE and UNMT. Our empirical results show that:

- 1) There is a positive correlation between the quality of the pre-trained UBWE and the performance of UNMT.
- 2) The UBWE quality significantly decreases during UNMT training.

Based on these two findings, we hypothesize that the learning of UNMT with UBWE agreement would enhance UNMT performance. In detail, we propose two approaches, UBWE agreement regularization and UBWE adversarial training, to maintain the quality of UBWE during NMT training. Empirical results on several language pairs show that the proposed methods significantly outperform the original UNMT. The remainder of this paper is organized as follows. In Section 2, we introduce the background of UNMT. The results of preliminary experiments are presented and analyzed in Section 3. In Section 4, we propose methods to jointly train UNMT with UBWE agreement. In Sections 5 and 6 , we describe experiments to evaluate the performance of our approach and analyze the results. Section 7 introduces some related work and Section 8 concludes the paper.

*Haipeng Sun was an internship research fellow at NICT when conducting this work.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1235–1245
Florence, Italy, July 28 - August 2, 2019.

2 Background of UNMT

There are three primary components of UNMT: UBWE initialization, denoising auto-encoder, and back-translation.

Consider a sentence X in language L_1 and a sentence Y in another language L_2. The data spaces of the L_1 sentence X and the L_2 sentence Y are denoted by ϕ_{L_1} and ϕ_{L_2}, respectively.

After initialization by UBWE, the encoders and decoders of L_1, L_2 are trained through denoising and back-translation. The objective function $\mathcal{L}_{all}$ of the entire UNMT model would be optimized as:

$$\mathcal{L}_{all} = \mathcal{L}_{auto} + \mathcal{L}_{bt}, \tag{1}$$

where $\mathcal{L}_{auto}$ is the objective function for auto-denoising, and $\mathcal{L}_{bt}$ is the objective function for back-translation.

2.1 Bilingual Word Embedding Initialization

Unlike supervised NMT (Bahdanau et al., 2015; Chen et al., 2017a,b, 2018a; Vaswani et al., 2017), there are no bilingual supervised signals in UNMT. Fortunately, UBWE (Zhang et al., 2017; Artetxe et al., 2018a; Conneau et al., 2018) successfully learned translation equivalences between word pairs from two monolingual corpora. Typically, UBWE initializes the embedding of the vocabulary for the encoder and decoder of UNMT. The pre-trained UBWE provides naive translation knowledge to enable the back-translation to generate pseudo-supervised bilingual signals (Artetxe et al., 2018c; Lample et al., 2018a). The embeddings of the encoder and decoder change independently during the UNMT training process.

2.2 Denoising Auto-encoder

The auto-encoder is difficult to learn useful knowledge for UNMT without some constraints. Otherwise, it would become a copying task that learned to copy the input words one by one (Lample et al., 2018a). To alleviate this problem, we utilize the same strategy of denoising auto-encoder (Vincent et al., 2010), and noise in the form of random token swaps is introduced in this input sentence to improve the model learning ability (Hill et al., 2016; He et al., 2016). The denoising auto-encoder, which encodes a noisy version and reconstructs it with the decoder in the same language, is optimized by minimizing the objective function:

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

where $C(X)$ and $C(Y)$ are noisy versions of sentences X and Y, $P_{L_1\to L_1}$ ($P_{L_2\to L_2}$) denotes the reconstruction probability in the language L_1 (L_2).

2.3 Back-translation

The denoising auto-encoder acts as a language model that has been trained in one language and does not consider the final goal of translating between two languages. Therefore, back-translation (Sennrich et al., 2016) was adapted to train translation systems in a true translation setting based on monolingual corpora. Formally, given the sentences X and Y, the sentences $Y_P(X)$ and $X_P(Y)$ would be produced by the model at the previous iteration. The pseudo-parallel sentence pair $(Y_P(X), X)$ and $(X_P(Y), Y)$ would be obtained to train the new translation model. Finally, the back-translation process is optimized by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_{bt} = &\mathbb{E}_{X\sim\phi_{L_1}}[-logP_{L_2\to L_1}(X|Y_P(X)] \\ &+ \mathbb{E}_{Y\sim\phi_{L_2}}[-logP_{L_1\to L_2}(Y|X_P(Y)],\end{aligned} \tag{3}$$

where $P_{L_1\to L_2}$ ($P_{L_2\to L_1}$) denotes the translation probability across two languages.

3 Preliminary Experiments

To investigate the relationship between UBWE and UNMT, we empirically choose one similar language pair (English-French which are in the same language family) and one distant language pair (English-Japanese which are in the different language families) as the corpora. The detailed experimental settings for UBWE and UNMT are given in Section 5.

3.1 Effect of UBWE Quality on UNMT Performance

Figure 1 shows the UNMT performance using UBWE with different levels of accuracy. To obtain UBWE with different accuracy levels, we used the VecMap (Artetxe et al., 2018a) embedding at different checkpoints to pre-train UNMT.[1]

[1] Accuracy "0" indicates only monolingual embeddings were used on each language before VecMap training started.

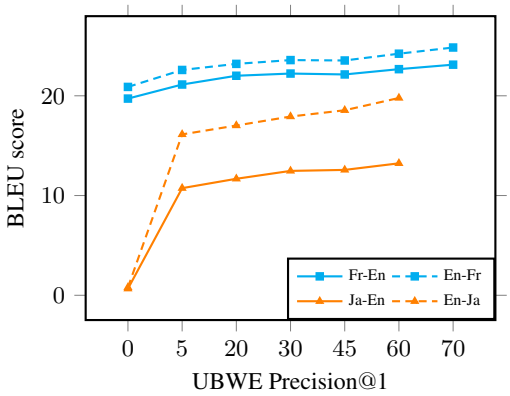


Figure 1: UNMT performance using UBWE with different levels of accuracy.

Precision@1 indicates the accuracy of word translation using the top-1 predicted candidate in the MUSE test set[2].

As the UBWE accuracy increased, the NMT performance of both language pairs increased. This indicates that the quality of pre-trained UBWE is important for UNMT.

3.2 Trend of UBWE Quality during UNMT Training

Figure 2 shows the trend in UBWE accuracy and BLEU score as UNMT proceeds through the training stage. VecMap was used to pre-train the word embedding for the encoder and decoder of UNMT. We used source embedding of encoder and target embedding of decoder to calculate the word translation accuracy on the MUSE test set during UNMT training.

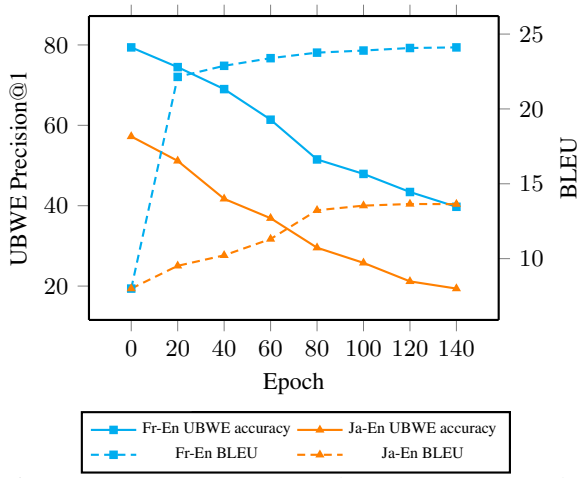


Figure 2: UBWE accuracy and BLEU score over the course of UNMT training.

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

Regardless of the language, the UBWE performance decreased significantly over the course of UNMT training, as shown in Figure 2.

3.3 Analysis

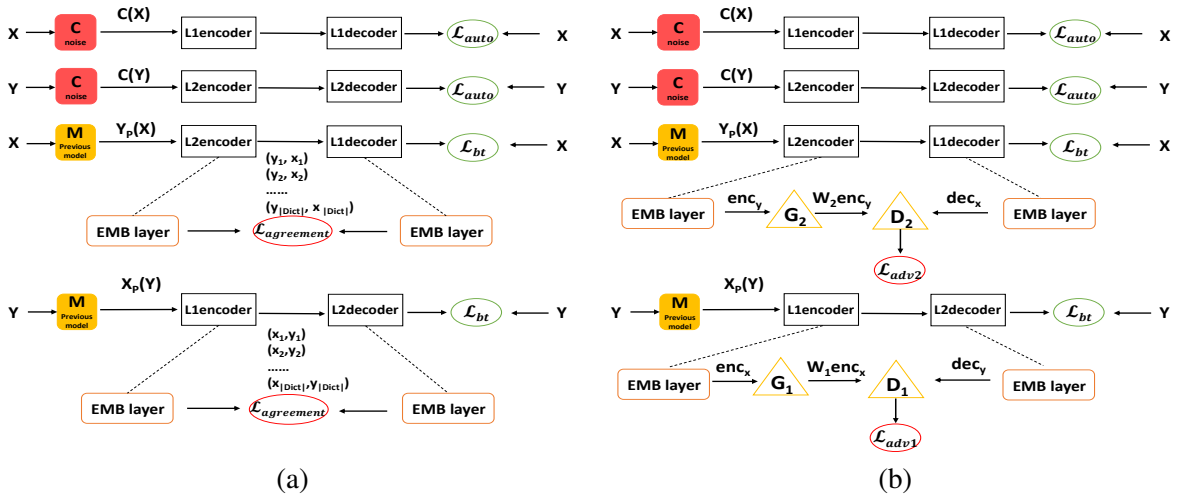
The empirical results in this section show that the quality of pre-trained UBWE is important to UNMT. However, the quality of UBWE decreases significantly during UNMT training. We hypothesize that maintaining the quality of UBWE may enhance the performance of UNMT. In this subsection, we analyze some possible solutions to this issue.

Use fixed embedding? As Figure 2 shows, the UBWE performance decreases significantly during the UNMT training process. Therefore, we try to fix the embedding of the encoder and decoder on the basis of the original baseline system (Baseline-fix). Table 1 shows that the performance of the Baseline-fix system is quite similar to that of the original baseline system. In other words, Baseline-fix prevents the degradation of UBWE accuracy; however, the fixed embedding also prevents UBWE from further improving UNMT training. Therefore, the fixed UBWE does not enhance the performance of UNMT.

Methods	Fr-En	En-Fr	Ja-En	En-Ja
Baseline	24.50	25.37	14.09	21.63
Baseline-fix	24.22	25.26	13.88	21.93

Table 1: Results of UNMT

Use byte pair encoding (BPE) to increase shared subwords? For English-French and English-German UNMT, Lample et al. (2018b) concatenated two bilingual corpora into a single monolingual corpus. They adopted BPE to enlarge the number of shared subwords in the two languages. The pre-trained monolingual subword embedding was used as the initialization for UNMT. Because there are many shared subwords in these similar language pairs, this method achieves better performance than other UBWE methods. However, this initialization does not work for distant language pairs such as English-Japanese and English-Chinese, where there are few shared subwords. Using word-based embedding in UNMT is more universal. In addition, word-based embedding are easy to combine with UBWE technology. Therefore, we do not adopt BPE in the proposed method.



(a) (b)

Figure 3: (a) Architecture of UNMT with UBWE Agreement Regularization; (b) Architecture of UNMT with UBWE Adversarial Training.

4 Train UNMT with UBWE Agreement

Based on previous empirical findings and analyses, we propose two joint agreement mechanisms, i.e., UBWE agreement regularization and UBWE adversarial training, that enable UBWE and UNMT to interact during the training process, resulting in improved translation performance. Figure 3 illustrates the architecture of UNMT and the proposed agreement mechanisms.

Generally, during UNMT training, an objective function $\mathcal{L}_{BWE}$ is added to ensure UBWE agreement. The general UNMT objective function can be reformulated as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{auto} + \mathcal{L}_{bt} + \lambda \mathcal{L}_{BWE}. \quad (4)$$

4.1 UBWE Agreement Regularization

On the basis of the existing architecture of UNMT, we induce UBWE agreement regularization during back-translation to maintain the UBWE accuracy in the encoder and the decoder during UNMT training. The similarity function $\text{Similarity}(L_1, L_2)$ of the encoder and decoder embeddings is used to measure the UBWE accuracy and the objective function $\mathcal{L}_{BWE}$ is

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

where enc_{L_1} and enc_{L_2} denote all word embeddings of encoders L_1 and L_2, respectively, dec_{L_1} and dec_{L_2} denote all word embeddings of decoders L_1 and L_2, respectively.

As there is no test or development data set that can be employed as a bilingual dictionary in UNMT, before computing $\text{Similarity}(L_1, L_2)$, we need to generate a synthetic word-pair dictionary to measure the UBWE accuracy during NMT training. Motivated by Conneau et al. (2018), we use the cross-domain similarity local scaling (CSLS) to measure the UBWE accuracy. This can also be viewed as the similarity between the source word embedding and the target word embedding.

$$\begin{aligned} \text{CSLS}(x_i, y_i) = 2 \cdot \cos(enc_{x_i}, dec_{y_i}) \\ - r(x_i) - r(y_i), \end{aligned} \quad (6)$$

$$r(x_i) = \frac{1}{K} \sum_{y \in \mathcal{N}(x_i)} \cos(enc_{x_i}, dec_y), \quad (7)$$

$$r(y_i) = \frac{1}{K} \sum_{x \in \mathcal{N}(y_i)} \cos(enc_x, dec_{y_i}), \quad (8)$$

where $y \in \mathcal{N}(x_i)$ denotes the K nearest neighborhood of the source word x_i, and similarly for $x \in \mathcal{N}(y_i)$. enc_{x_i} denotes the embedding of word x_i in encoder L_1 and dec_{y_i} denotes the word embedding of y_i in decoder L_2.

As the size of the entire vocabulary is large, we select a subset as the synthetic word-pair dictionary. By ranking the CSLS, we can select the most accurate word pairs $\{x_i, y_i\}$ as the synthetic dictionary $Dict_{x->y}$. The opposite word pairs $Dict_{y->x} = \{y_j, x_j\}$ could be obtained by the

same method. enc_{y_j} denotes the embedding of word y_j by encoder L_2 and dec_{x_j} denotes the embedding of word x_j by decoder L_1. Both dictionary sizes are set to $|Dict|$. Therefore, the similarity between the word embeddings in the encoder and decoder is measured as

$$\begin{aligned}&\text{Similarity}(enc_{L_1}, dec_{L_2}) \\ &\approx \frac{1}{|Dict|}\sum_{i}^{|Dict|}(1-\cos(enc_{x_i}, dec_{y_i})).\end{aligned} \tag{9}$$

$$\begin{aligned}&\text{Similarity}(enc_{L_2}, dec_{L_1}) \\ &\approx \frac{1}{|Dict|}\sum_{j}^{|Dict|}(1-\cos(enc_{y_j}, dec_{x_j})).\end{aligned} \tag{10}$$

The above similarity between word pairs in $Dict$ is used for UBWE agreement regularization during back-translation. Note that the synthetic word-pair dictionary is dynamically selected in each epoch of UNMT training.

4.2 UBWE Adversarial Training

In UBWE, there is a transformation matrix to project the source word embedding to the target word embedding. Motivated by Conneau et al. (2018), we induce a transformation matrix using an adversarial approach . The generator is estimated as:

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

where enc_x is the L_1 the encoder word embedding, dec_y is the corresponding L_2 decoder word embedding, and W_1 is the transformation matrix that project the embedding space of enc_x onto that of dec_y. The discriminator D_1 is a multi-layer perceptron representing the probability that the word embedding comes from this language. It is trained to discriminate the language to which the word embedding between $W_1 enc_x$ and dec_y belongs. W_1 is trained to confuse the discriminator D_1 by making $W_1 enc_x$ and dec_y increasingly similar. In other words, we train D_1 to maximize the probability of choosing the accurate language between the original word embedding and samples from G_1. The generator G_1 is trained to minimize $\log(1-D_1(G_1(enc_x)))$. Thus, the two-player minimax game (Goodfellow et al., 2014) with value function $V(G_1, D_1)$ is optimized as:

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

D_2 and G_2 are similar to D_1 and G_1. The objective functions for the discriminator D_1 and generator G_1 can be written as:

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

$$\begin{aligned}\mathcal{L}_{G_1} &= \mathbb{E}_{enc_x}[-\log(D_1(G_1))] \\ &+ \mathbb{E}_{dec_y}[-\log(1-D_1(dec_y)].\end{aligned} \tag{14}$$

$\mathcal{L}_{D_2}$ and $\mathcal{L}_{G_2}$ are similar to $\mathcal{L}_{D_1}$ and $\mathcal{L}_{G_1}$. After inducing UBWE adversarial training into UNMT, the $\mathcal{L}_{BWE}$ objective function is minimized as

$$\mathcal{L}_{BWE} \triangleq \mathcal{L}_{adv} = \mathcal{L}_{adv1} + \mathcal{L}_{adv2}, \tag{15}$$

where $\mathcal{L}_{adv1} = \mathcal{L}_{G_1} + \mathcal{L}_{D_1}$ and $\mathcal{L}_{adv2} = \mathcal{L}_{G_2} + \mathcal{L}_{D_2}$. The proposed L_{BWE} ($L_{agreement}$ or L_{adv}) is added to the L_{all} in Eq. 4 during back-translation of UNMT training as shown in Figure 3.

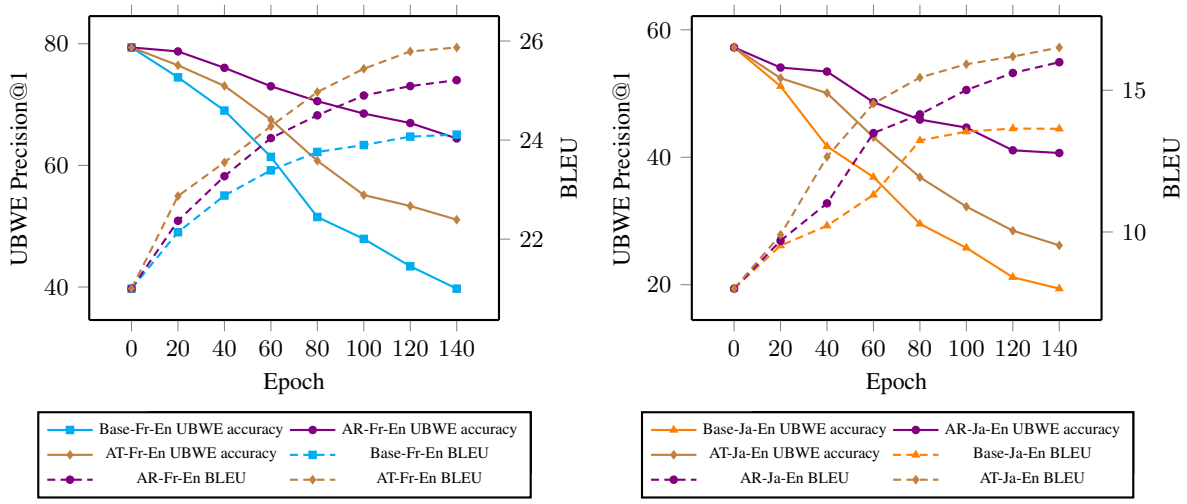
5 Experiments

5.1 Datasets

The proposed methods were evaluated on three language pairs: French-English (Fr-En), German-English (De-En), and Japanese-English (Ja-En). Fr-En and De-En are similar European language pairs. We used 30 million sentences from the WMT monolingual News Crawl datasets from 2007 to 2013. Ja-En is a distant languages pair and so UBWE training is much more difficult than for similar European language pairs (Søgaard et al., 2018). In addition, Japanese and English are different language families and their word orderings are quite different. As a result, the performance of Ja-En UNMT is too poor to further empirical study if only pure monolingual data are used. Therefore, we constructed simulated experiments using shuffled parallel sentences, i.e., 3.0M sentence pairs from the ASPEC corpus for Ja-En. We reported the results on WMT newstest2014 for Fr-En, WMT newstest2016 for De-En, and WAT-2018 ASPEC testset for Ja-En.

5.2 UBWE Settings

For UBWE training, we first used the monolingual corpora described above to train



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

Figure 4: The trends of UBWE quality and BLEU score for baseline (Base), UBWE agreement regularization (AR), and UBWE adversarial training (AT) during UNMT training on the Fr-En and Ja-En dataset

Methods	De-En	En-De	Fr-En	En-Fr	Ja-En	En-Ja
Artetxe et al. (2018c)	n/a	n/a	15.56	15.13	n/a	n/a
Lample et al. (2018a)	13.33	9.64	14.31	15.05	n/a	n/a
Yang et al. (2018)	14.62	10.86	15.58	16.97	n/a	n/a
Lample et al. (2018b)	21.0	17.2	24.2	25.1	n/a	n/a
UNMT Baseline	21.23	17.06	24.50	25.37	14.09	21.63
+ UBWE agreement regularization	22.38++	18.04++	25.21++	27.86++	16.36++	23.01++
+ UBWE adversarial training	**22.67++**	**18.29++**	**25.87++**	**28.38++**	**17.22++**	**23.64++**

Table 2: Performance (BLEU score) of UNMT. "++" after a score indicates that the proposed method was significantly better than the UNMT baseline at significance level p <0.01.

the embeddings for each language independently with fastText[3](Bojanowski et al., 2017) (default settings). The word embeddings were normalized by length and mean centered before bilingual projection. We then used VecMap[4](Artetxe et al., 2018a) (default settings) to project two monolingual word embeddings into one space.

To evaluate the quality of UBWE, we selected the accuracy of word translation using the top-1 predicted candidate in the MUSE test set as the criterion.

5.3 UNMT Settings

In the training process for UNMT, we used the transformer-based UNMT toolkit[5] and the settings of Lample et al. (2018b). That is, we used four layers in both the encoder and the decoder. Three out of the four encoder and decoder layers were shared between the source and target languages. The dimension of the hidden layers was set to 512. Training used a batch-size of 32 and the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.0001, $\beta_1 = 0.5$. The vocabulary size was set to 60k by concatenating the source and target corpora. We performed 140 epochs[6] (approximately 500K iterations) to train every model. The case-sensitive BLEU score computed with the $multi - bleu.perl$ script from Moses[7] was used as the evaluation metric. For model selection, we followed the strategy described by Lample et al. (2018a). That is, the BLEU score computed between the original source sentences

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

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

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

[6] The definition of epoch in UNMT is different from that in NMT. We followed the settings in Lample et al. (2018b)'s toolkit, i.e., 3500 iterations as one epoch.

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

and their reconstructions was used as the criterion. We selected the model that had the highest average BLEU score over the two translation directions.

For the proposed methods, both UBWE agreement regularization and UBWE adversarial training were added as objective functions at the beginning of UNMT training. The detailed parameter settings are discussed in Section 6.

5.4 Performance

Figure 4 shows the trend in UBWE quality and BLEU score during UNMT training on Fr-En and Ja-En. Our observations are as follows:

1) For all systems, the UBWE accuracy decreases during UNMT training. This is consistent with our finding in the preliminary experiments.

2) For the system with UBWE agreement regularization and UBWE adversarial training, UBWE accuracy decreased much more slowly than in the original baseline system. This indicates that the proposed methods effectively mitigated the degradation of UBWE accuracy.

3) Regarding the two proposed methods, UBWE agreement regularization was better at mitigating the degradation of UBWE accuracy than UBWE adversarial training.

Table 2 presents the detailed BLEU scores of the UNMT systems on the De-En, Fr-En, and Ja-En test sets. Our observations are as follows:

1) Our re-implemented baseline performed similarly with the state-of-the-art method of Lample et al. (2018b). This indicates that the baseline is a strong system.

2) The proposed methods significantly outperformed the corresponding baseline in all the language pairs by 1∼3 BLEU scores.

3) Regarding the two proposed methods, UBWE adversarial training performed slightly better than UBWE agreement regularization by BLEU score, although UBWE agreement regularization was better at maintaining UBWE accuracy. The reason may be that agreement regularization is just added to the training objective of UNMT. In comparison, UBWE adversarial training is jointly trained with UNMT, thus has more interaction with UNMT model.

6 Discussion

We now analyze the effect of the hyper-parameters. There are two primary factors that affect the performances of the proposed methods: the synthetic word-pair dictionary size for UBWE agreement regularization and λ for UBWE adversarial training.

6.1 Effect of Dictionary Size

We first evaluated the impact of the synthetic word-pair dictionary size $|Dict|$ during UBWE agreement regularization training on the Fr-En task. As indicated by Table 3, almost all models with different dictionary sizes outperformed the baseline system. This indicates that the proposed method is robust.

Dict Size	Fr-En BLEU	En-Fr BLEU
Baseline	24.50	25.37
20K	25.15	27.18
10K	25.10	27.48
5K	25.14	27.58
3K	25.21	27.86
1K	25.25	27.40
500	25.13	27.07

Table 3: Effect on Dictionary Size

We also investigated the relationship between dictionary size and UBWE accuracy. As shown in Fig. 5, a larger dictionary size results in a slower decrease in UBWE accuracy. This indicates that a larger dictionary size helps estimate a better UBWE agreement. However, larger dictionary size did not always obtain a higher BLEU as shown in Table 3. The model with a dictionary size of 3000 achieved the best performance.

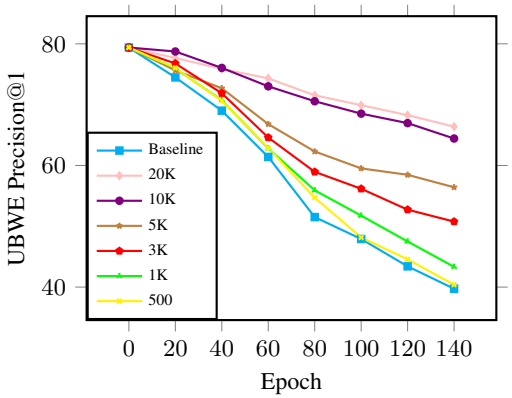


Figure 5: UBWE accuracy with respect to dictionary size on the Fr-En test set during UNMT training.

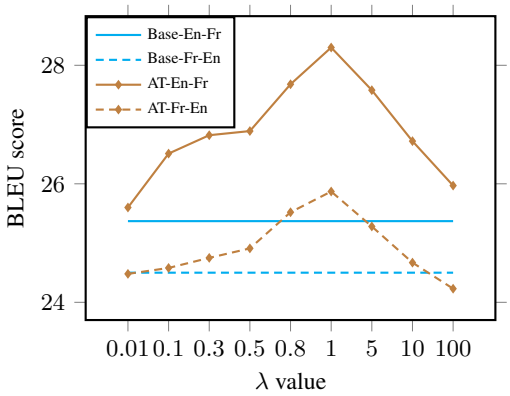


Figure 6: Effect of Hyper-parameter λ for UBWE adversarial training (AT) model on the En $\leftrightarrow$ Fr dataset.

6.2 Effect of Hyper-parameter λ

In Figure 6, we empirically investigated how the hyper-parameter λ in Eq. (4) affects the UNMT performance on the Fr-En task. The selection of λ influences the role of the $\mathcal{L}_{BWE}$ across the entire UNMT training process. Larger values of λ cause the $\mathcal{L}_{BWE}$ to play a more important role than the back-translation and denoising loss terms. The smaller the value of λ, the less important are the $\mathcal{L}_{BWE}$. As the Fig. 6 shows, λ ranging from 0.01 to 10 nearly all enhanced UNMT performance and a balanced $\lambda = 1$ achieved the best performance.

6.3 Efficiency

We now discuss the efficiency of our proposed methods. Table 4 indicates that UBWE agreement regularization does not increase the number of parameters. UBWE adversarial training adds very few parameters. The training speed of these methods is almost the same. In addition, the proposed methods do not affect the UNMT decoding. Thus, our proposed methods do not affect the speed of the model.

	Parameters	Speed
Baseline	120,141K	3784
UBWE agreement regularization	120,141K	3741
UBWE adversarial training	120,764K	3733

Table 4: Analysis on parameters and training speed (number of processed words per second on one P100).

7 Related Work

The supervised BWE (Mikolov et al., 2013), which exploits similarities between the source language and the target language through a linear transformation matrix, serves as the basis for many NLP tasks, such as machine translation (Bahdanau et al., 2015; Vaswani et al., 2017; Chen et al., 2018b; Zhang and Zhao, 2019), dependency parsing (Zhang et al., 2016; Li et al., 2018), semantic role labeling (He et al., 2018; Li et al., 2019). However, the lack of a large word-pair dictionary poses a major practical problem for many language pairs. UBWE has attracted considerable attention. For example, Artetxe et al. (2017) proposed a self-learning framework to learn BWE with a 25-word dictionary, and Artetxe et al. (2018a) extended previous work without any word dictionary via fully unsupervised initialization. Zhang et al. (2017) and Conneau et al. (2018) proposed UBWE methods via generative adversarial network training.

Recently, several UBWE methods (Conneau et al., 2018; Artetxe et al., 2018a) have been applied to UNMT (Artetxe et al., 2018c; Lample et al., 2018a). These rely solely on monolingual corpora in each language via UBWE initialization, denoising auto-encoder, and back-translation. A shared encoder was used to encode the source sentences and decode them from a shared latent space (Artetxe et al., 2018c; Lample et al., 2018a). The difference is that Lample et al. (2018a) used a single shared decoder and Artetxe et al. (2018c) leveraged two independent decoders for each language. Yang et al. (2018) used two independent encoders for each language with a weight-sharing mechanism to overcome the weakness of retaining the uniqueness and internal characteristics of each language. Lample et al. (2018b) achieved remarkable results in several similar languages such as English-French by concatenating two bilingual corpora as one monolingual corpus and using monolingual embedding pre-training in the initialization step. This initialization achieves better performance than other UBWE methods. However, it does not work in some distant language pairs such as English-Japanese. This is why we did not use this initialization process for UBWE in our method.

In addition, an alternative unsupervised method based on statistical machine translation (SMT) was proposed (Lample et al., 2018b; Artetxe et al., 2018b). The unsupervised machine translation performance was improved through combining UNMT and unsupervised SMT (Marie and Fujita, 2018; Ren et al., 2019; Artetxe et al., 2019). More recently, Lample and Conneau (2019) achieved

better UNMT performance through introducing the pretrained language model. Neural network based language model has been shown helpful in supervised machine translation (Wang et al., 2014; Wang et al., 2018; Marie et al., 2018). We think that the proposed agreement mechanism can work with the pretrained language model.

8 Conclusion

UBWE is a fundamental component of UNMT. In previous methods, the pre-trained UBWE is only used to initialize the word embedding of UNMT. In this study, we found that the performance of UNMT is significantly affected by the quality of UBWE, not only in the initialization stage, but also during UNMT training. Based on this finding, we proposed two joint learning methods to train UNMT with UBWE agreement. Empirical results on several language pairs show that the proposed methods can mitigate the decrease in UBWE accuracy and significantly improve the performance of UNMT.

Acknowledgments

The corresponding authors are Rui Wang and Tiejun Zhao. This work was partially conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan. 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".

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. 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)*, pages 451–462, Vancouver, Canada.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *CoRR*, abs/1902.01313.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827, Osaka, Japan.

Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, Akihiro Tamura, Eiichiro Sumita, and Tiejun Zhao. 2017a. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2852, Copenhagen, Denmark.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2018a. Syntax-directed attention for neural machine translation. In *AAAI Conference on Artificial Intelligence*, pages 4792–4798, New Orleans, Lousiana, USA.

Kehai Chen, Tiejun Zhao, Muyun Yang, and Lemao Liu. 2017b. Translation prediction with source dependency-based context representation. In *AAAI Conference on Artificial Intelligence*, pages 3166–3172, San Francisco, California, USA.

Kehai Chen, Tiejun Zhao, Muyun Yang, Lemao Liu, Akihiro Tamura, Rui Wang, Maosao Utiyama, and Eiichro Sumita. 2018b. A neural approach to source

dependence based context model for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):266–280.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Third International Conference on Learning Representations*, San Diego, California.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, Montreal, Quebec, Canada.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 820–828, Barcelona, Spain.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. 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*, pages 1367–1377, San Diego California, USA.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*, San Diego, California.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. *CoRR*, abs/1901.05280.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. 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*, pages 250–256, Denver, Colorado.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. 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*, pages 453–459, Belgium, Brussels.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with SMT as posterior regularization. *CoRR*, abs/1901.04112.

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)*, Berlin, Germany.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the Fifth International Conference on Learning Representations*, Toulon, France.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. 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)*, pages 778–788, Melbourne, Australia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.

Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2018. Sentence selection and weighting for neural machine translation domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1727–1741.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2014. 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)*, pages 189–195, Doha, Qatar.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A bilingual graph-based semantic model for statistical machine translation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2950–2956, New York, USA.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2018. Graph-based bilingual word embedding for statistical machine translation. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.*, 17(4):31:1–31:23.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. 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*, pages 1006–1011, Denver, Colorado.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia.

Huan Zhang and Hai Zhao. 2019. Minimum divergence vs. maximum margin: An empirical comparison on seq2seq models. In *Proceedings of the Seventh International Conference on Learning Representations*, New Orleans, USA.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. 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.

Zhisong Zhang, Hai Zhao, and Lianhui Qin. 2016. Probabilistic graph-based dependency parsing with convolutional neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1382–1392, Berlin, Germany.