Explicit Cross-lingual Pre-training for Unsupervised Machine Translation

Shuo Ren[†‡]*, **Yu Wu** [§], **Shujie Liu**[§], **Ming Zhou**[§], **Shuai Ma**[†‡]
[†]SKLSDE Lab, Beihang University, Beijing, China
[‡]Beijing Advanced Innovation Center for Big Data and Brain Computing, China
[§]Microsoft Research Asia, Beijing, China
[†]{shuoren,mashuai}@buaa.edu.cn [§]{Wu.Yu,shujliu,mingzhou}@microsoft.com

Abstract

Pre-training has proven to be effective in unsupervised machine translation due to its ability to model deep context information in cross-lingual scenarios. However, the cross-lingual information obtained from shared BPE spaces is inexplicit and limited. In this paper, we propose a novel cross-lingual pre-training method for unsupervised machine translation by incorporating explicit cross-lingual training signals. Specifically, we first calculate cross-lingual n-gram embeddings and infer an n-gram translation table from them. With those n-gram translation pairs, we propose a new pre-training model called Cross-lingual Masked Language Model (CMLM), which randomly chooses source n-grams in the input text stream and predicts their translation candidates at each time step. Experiments show that our method can incorporate beneficial cross-lingual information into pre-trained models. Taking pre-trained CMLM models as the encoder and decoder, we significantly improve the performance of unsupervised machine translation. Our code is available at https://github.com/Imagist-Shuo/CMLM.

1 Introduction

Unsupervised machine translation has become an emerging research interest in recent years (Artetxe et al., 2017; Lample et al., 2017, 2018; Artetxe et al., 2018b; Marie and Fujita, 2018; Ren et al., 2019; Lample and Conneau, 2019). The common framework of unsupervised machine translation builds two initial translation models at first (i.e., source to target and target to source), and then does iterative back-translation (Sennrich et al., 2016a; Zhang et al., 2018) with the two models using pseudo data generated by each other. The initialization process is crucial to the final translation performance as pointed in Lample et al. (2018), Artetxe et al. (2018b) and Ren et al. (2019).

Previous approaches benefit mostly from cross-lingual n-gram embeddings, but recent work proves that cross-lingual language model pre-training could be a more effective way to build initial unsupervised machine translation models (Lample and Conneau, 2019). However, in their method, the cross-lingual information is mostly obtained from shared Byte Piece Encoding (BPE) (Sennrich et al., 2016b) spaces during pre-training, which is inexplicit and limited. Firstly, although the same BPE pieces from different languages may share the same semantic space, the semantic information of n-grams or sentences in different languages may not be shared properly. However, cross-lingual information based on n-gram level is crucial to model the initialization of unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018b), which is absent in the current pre-training method. Secondly, BPE sharing is limited to languages that share much of their alphabet. For language pairs that are not the case, the above pre-training method may not provide much useful cross-lingual information.

In this paper, by incorporating explicit cross-lingual training signals, we propose a novel cross-lingual pre-training method based on BERT (Devlin et al., 2018) for unsupervised machine translation. Our method starts from unsupervised cross-lingual n-gram embeddings, from which we infer n-gram translation pairs. Then, we propose a new pre-training objective called Cross-lingual Masked Language Model (CMLM), which masks the input n-grams randomly and predicts their corresponding n-gram translation candidates inferred above. To solve the mismatch between different lengths of the masked source and predicted target n-grams, IBM models are introduced (Brown et al., 1993) to derive the training loss at each

*Contribution during internship at MSRA.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 770–779, Hong Kong, China, November 3–7, 2019.

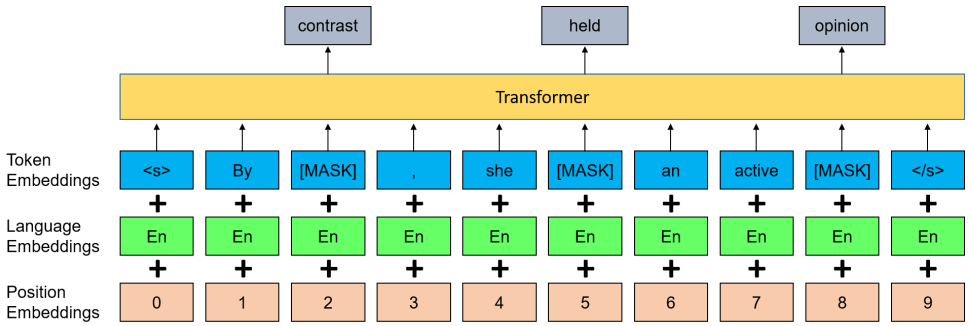

Figure 1: Masked Language Model (MLM) for BERT training. For a given sentence, this task is to predict randomly masked tokens, i.e., "contrast", "held" and "opinion". In practice, it is implemented based on BPE.

time step. In this way, we can guide the model with more explicit and strong cross-lingual training signals, and meanwhile, leverage the potential of BERT to model context information. We then use two pre-trained cross-lingual language models as the encoder and decoder respectively to build desired machine translation models. Our method can be iteratively performed with the n-gram translation table updated by downstream tasks. Experiments show that our method can produce better cross-lingual representations and significantly improve the performance of unsupervised machine translation. Our contributions are listed as follows.

- We propose a novel cross-lingual pre-training method to incorporate explicit cross-lingual information into pre-trained models, which significantly improves the performance of unsupervised machine translation.
- We introduce IBM models to calculate the step-wise training loss for CMLM, which breaks the limitation that masked n-grams and predicted ones have to be the same length during BERT training.
- We produce strong context-aware cross-lingual representations with our pre-training method, which helps in word alignment and cross-lingual classification tasks.

2 Background

2.1 BERT

BERT (Devlin et al., 2018), short for Bidirectional Encoder Representations from Transformers, is a powerful pre-training method for natural language processing and breaks records of many NLP tasks after corresponding fine-tuning. The core idea of BERT is pre-training a deep bidirectional Transfomer encoder (Vaswani et al., 2017) with two training tasks. The first one is Masked Language Model (MLM) referring to the *Cloze* task (Taylor, 1953), which takes a straightforward approach of masking some percentage of the input tokens at random, and then predicting them with the corresponding Transformer hidden states, as shown in Figure 1. The second one is Next Sentence Prediction, which means to predict whether two sentences are adjacent or not. This task is designed for some tasks that need modeling the relationship between two sentences such as Question Answering (QA) and Natural Language Inference (NLI).

2.2 XLM

Based on BERT, Lample and Conneau (2019) propose a cross-lingual version called XLM and reach the state-of-the-art performance on some cross-lingual NLP tasks including cross-lingual classification (Conneau et al., 2018), machine translation, and unsupervised cross-lingual word embedding. The basic points of XLM are mainly two folds. The first one is to use a shared vocabulary of BPE (Sennrich et al., 2016b) to provide potential cross-lingual information between two languages just as mentioned in Lample et al. (2018), in an inexplicit way though. The second point is a newly proposed training task called Translation Language Modeling (TLM), which extends MLM by concatenating parallel sentences into a single training text stream. In this way, the model can leverage the cross-lingual information provided by parallel sentences to predict the masked tokens. However, for unsupervised machine translation, TLM cannot be used due to the lack of parallel sentences. Different from them, we are motivated to give the model more explicit and strong cross-lingual information and propose a new pre-training method by (1) masking source n-grams and (2) predicting their corresponding translation candidates.

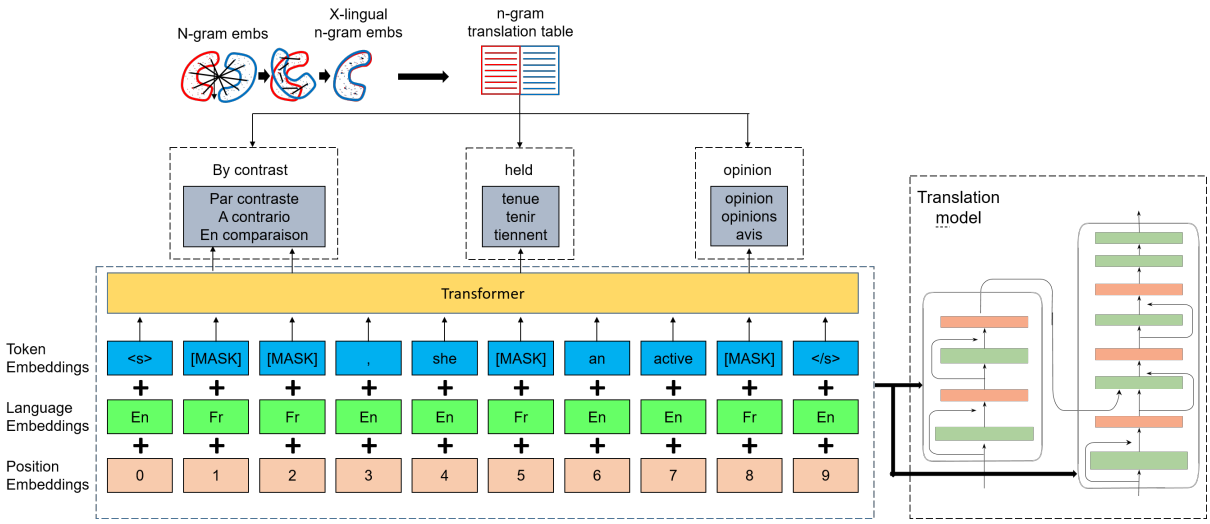

Figure 2: Method overview. Our method consists of three steps. The first one is the n-gram translation table inferring. The second one is pre-training with our proposed objective Cross-lingual Masked Language Model (CMLM) which is to predict the translation candidates of randomly masked n-grams. The last step is to leverage the pre-trained cross-lingual language models as the encoder and decoder of a neural machine translation model and fine-tune the translation model iteratively.

3 Method

3.1 Overview

Our method can be divided into three steps as shown in Figure 2. Given two languages X and Y, we first get unsupervised cross-lingual n-gram embeddings of them, from which we infer n-gram translation tables (source-to-target and target-to-source). The n-gram translation pairs inferred in this way have proven to be instructive for initial unsupervised machine translation models (Artetxe et al., 2018b; Lample et al., 2018; Marie and Fujita, 2018; Ren et al., 2019). Then, we pre-train cross-lingual BERT language models with our proposed Cross-lingual Masked Language Model (CMLM) objective. Specifically, we randomly choose n-grams in the monolingual sentences and predict corresponding translation candidates in the n-gram translation table inferred in the first step. In this way, we can guide the model with explicit and strong cross-lingual training signals. Finally, two pre-trained cross-lingual language models are used to initialize the encoder and decoder respectively, based on which, denoising auto-encoder and iterative back-translation are leveraged to fine-tune the unsupervised machine translation models.

The above process is repeated by updating the n-gram table with the n-gram translation pairs extracted from the pseudo data generated by the translation models. In the following subsections, we will give details of each step.

3.2 N-gram Translation Table Inferring

Following previous work of unsupervised machine translation (Artetxe et al., 2018b; Lample et al., 2018; Ren et al., 2019), given two languages X and Y, we build our n-gram translation tables as follows. First, we obtain monolingual n-gram embeddings using fastText (Bojanowski et al., 2017) and then get cross-lingual n-gram embeddings using vecmap (Artetxe et al., 2018a) in a fully unsupervised way. Based on that, we calculate the similarity score of n-grams x and y in two languages respectively with the marginal-based scoring method (Conneau et al., 2017; Artetxe and Schwenk, 2018). Specifically, given the cross-lingual embeddings of x and y, denoted as e_x and e_y, the similarity score is calculated as:

$$\text{sim}(x, y) = \text{margin}(\cos(e_x, e_y), \sum_{z \in \text{NN}_n(x)} \frac{\cos(e_x, e_z)}{2n} + \sum_{z \in \text{NN}_n(y)} \frac{\cos(e_y, e_z)}{2n}) \quad (1)$$

where $\text{margin}(a, b)$ is a marginal scoring function and $\text{NN}_n(x)$ denotes x's k-nearest neighbors in the other language. In our experiments, n is 5 and $\text{margin}(a, b) = \frac{a}{b}$.

We take the above similarity scores as the translation probabilities between x and y in the n-gram table. For each top frequent n-gram in the source language, we retrieve top-k n-gram translation candidates in the target language.

3.3 Cross-lingual Masked Language Model

In this section, we introduce our proposed method for pre-training cross-lingual language models based on BERT. Unlike the masked language model (MLM) described in Section 2.2 which masks several tokens in the input stream and predict those tokens themselves, we randomly select some percentage of n-grams in the input source sentence, and predict their translation candidates retrieved from Section 3.2. We call our proposed pre-training objective "Cross-lingual Masked Language Model" (CMLM) as shown in Figure 2. The difficulty for BERT to predict target phrases during training is that the lengths of the translation candidates are sometimes different from the source n-grams. To deal with this problem, we turn to IBM Model 2 (Brown et al., 1993) to calculate the training loss at each time step. Our proposed method breaks the limitation that masked n-grams and predicted ones have to be the same length during BERT training.

Specifically, according to IBM Model 2, given a source n-gram x_1^l and a target one y_1^m, where l and m are the numbers of tokens in the source and target n-grams respectively, the translation probability from x_1^l to y_1^m is calculated as:

$$\mathbf{Pr}(y_1^m|x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i|j,l,m) p(y_j|x_i) \quad (2)$$

where $\epsilon = p(m|x_1^l)$, i.e. the probability that the translation of x_1^l consists of m tokens; $a(j|i,l,m)$ is the probability that the i^{th} source token is aligned with the j^{th} target token conditioned on the lengths l and m, while $p(y_j|x_i)$ is the translation probability from the source token x_i to the target token y_i. Based on the IBM Model 2, the loss function of our CMLM is defined as

$$\begin{aligned} \mathcal{L}_{cmlm} &= -\log \mathbf{Pr}(y_1^m|x_1^l) \\ &= -\log \epsilon - \sum_{j=1}^{m} \log \sum_{i=0}^{l} a(i|j,l,m) p(y_j|x_i) \end{aligned} \quad (3)$$

The derived gradient w.r.t model parameters θ at each time step can be calculated as follows:

$$\begin{aligned} &\nabla_\theta \mathcal{L}_{cmlm} \\ &= -\sum_{j=1}^{m} \frac{a(i|j,l,m) p(y_j|x_i)}{\sum_{i=0}^{l} a(i|j,l,m) p(y_j|x_i)} \nabla_\theta \log p(y_j|x_i) \end{aligned} \quad (4)$$

Since the target n-gram y_1^m is predicted with our modified BERT, in practice, the source word x_i in Eq.(4) is replaced with its context-sensitive embedding $\mathrm{C}(x_1^l)$, which is the corresponding hidden state of the top Transformer layer. The alignment probability $a(i|j,l,m)$ cannot be learned during training because of the absence of bilingual corpus. Therefore, cross-lingual BPE embeddings are leveraged to calculate the normalized $\mathrm{sim}(x_i, y_j)$ to approximate $a(i|j,m,l)$. $p(y_j|x_i)$ is the model prediction in Softmax outputs. For each source n-gram, all of the retrieved k translation candidates are used to calculate the cross entropy loss, which are weighted with their translation probabilities in the n-gram table.

Given a language pair $X - Y$, we process both languages with the same shared BPE vocabulary using their monolingual sentences together during pre-training. Following Devlin et al. (2018); Lample and Conneau (2019), in our CMLM objective, we randomly sample 15% of the BPE n-grams from the text streams, and replace them by [MASK] tokens 70% of the time. During pre-training, in each iteration, a batch is composed of sentences sampled from the same language, and we alternate between MLM and CMLM objectives. Different from the original MLM in BERT, in the half of the MLM time, we randomly choose some source n-grams in the input text stream, and replace them with their translation candidates to construct code-switching sentences. Our final pre-training loss is defined as

$$\mathcal{L}_{pre} = \mathcal{L}_{cmlm} + \mathcal{L}_{mlm} \quad (5)$$

3.4 Unsupervised Machine Translation

Taking two cross-lingual language models pre-trained with the above method as the encoder and decoder, we build an initial unsupervised neural machine translation model. Then, we train the model with monolingual data until convergence via denoising auto-encoder and iterative back-translation, as described in Artetxe et al. (2017); Lample et al. (2017, 2018). Different from them, we step further and make another iteration with updated n-gram translation tables. Specifically, we translate the monolingual sentences with our latest translation model and run GIZA++ (Och and Ney, 2003) on the generated pseudo parallel data to extract updated n-gram translation pairs, which are used to tune the encoder as Section 3.3, together with the back-translation within a multi-task learn-

ing framework. Experimental results show that running another iteration can further improve the translation performance.

It is also interesting to explore the usage of pre-trained decoders in the translation model. It seems that pre-training decoders has a smaller effect on the final performance than pre-training encoders (Lample and Conneau, 2019), one reason for which could be that the encoder-to-decoder attention is not pre-trained. Therefore, the parameters of the decoder need to be re-adjusted substantially in the following tuning process for MT task. In our experiments, we explore some other usage of pre-trained decoders, i.e., we use the pre-trained decoder as the feature extractor and feed the outputs into a new decoder consisting of several Transformer layers with the attention to the encoder. We find this method improves the performance of some language translation directions.

4 Experiments

In this section, we conduct experiments to evaluate our proposed pre-training method. In Section 4.1, we will introduce the setup of our experiments, followed by the overall results of the final unsupervised MT models. Then, in Section 4.3, we will discuss another usage of pre-trained decoders for translation models. To evaluate the cross-lingual modeling capacity of our pre-trained encoders, in Section 4.4, we conduct experiments on word alignment and cross-lingual classification tasks. Finally, we do the ablation study to check the performance contribution of each component in our proposed method.

4.1 Setup

Data and Preprocess

In our experiments, we consider three language pairs, English-French (en-fr), English-German (en-de) and English-Romanian (en-ro). For each language, we use all the available sentences in NewsCrawl till 2018, monolingual datasets from WMT. The NewsCrawl data are used in both pre-training and the following unsupervised NMT iteration process. Our CMLM is optimized based on the pre-trained models released by Lample and Conneau (2019)[1], which are trained with Wikipedia dumps. For fair comparison, we use *newstest* 2014 as the test set for en-fr, and *newstest* 2016 for en-de and en-ro.

We use Moses scripts for tokenization, and use fastBPE[2] to split words into subword units with their released BPE codes[1]. The number of shared BPE codes for each language pair is 60,000.

Implementation Details

Our implementation is based on the released code of XLM[1] (Paszke et al., 2017). Specifically, we use a Transformer architecture with 1024 hidden units, 8 heads, GELU activations (Hendrycks and Gimpel, 2016), with a dropout rate of 0.1. The models are trained with the Adam optimizer (Kingma and Ba, 2014), a linear warmup (Vaswani et al., 2017) and the learning rates varying from 10^4 to 5×10^4.

For both of the MLM and CMLM objectives, we use streams of 256 tokens and mini-batches of size 64. We use the averaged perplexity over languages as a stopping criterion for training. For machine translation, we use 6 Transformer layers, and we create mini-batches of 2000 tokens.

Baselines

Our method is compared with six baselines of unsupervised MT systems listed in the upper part of Table 1. The first two baselines (Artetxe et al., 2017; Lample et al., 2017) use a shared encoder and different decoders for two languages with the training methods of denoising auto-encoder and iterative back-translation. The third baseline (Artetxe et al., 2018b) is an unsupervised PBSMT model, which uses the initial PBSMT models built with language models and n-gram translation tables inferred from cross-lingual embeddings, followed with the iterative back-translation. The fourth baseline (Lample et al., 2018) is a hybrid method of unsupervised NMT and PBSMT by combining the pseudo data generated by PBSMT models into the final iteration of NMT. The fifth baseline (Ren et al., 2019) is also a hybrid method of NMT and PBSMT but different from Lample et al. (2018), they leverage PBSMT as posterior regularization in each NMT iteration. The last baseline is XLM described in Section 2.2.

4.2 Overall Results

The overall comparison results of unsupervised machine translation are shown in Table 1. From the table, we find that our proposed method significantly outperforms previous methods on all language pairs by the average BLEU score of 1.7,

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

[2]https://github.com/glample/fastBPE

	Method	fr2en	en2fr	de2en	en2de	ro2en	en2ro
Baselines	(Artetxe et al., 2017)	15.6	15.1	-	-	-	-
	(Lample et al., 2017)	14.3	15.1	13.3	9.6	-	-
	(Artetxe et al., 2018b)	25.9	26.2	23.1	18.2	-	-
	(Lample et al., 2018)	27.7	28.1	25.2	20.2	23.9	25.1
	(Ren et al., 2019)	28.9	29.5	26.3	21.7	-	-
	(Lample and Conneau, 2019)	33.3	33.4	34.3	26.4	31.8	33.3
Ours	Iter 1	34.8	34.9	35.5	**27.9**	33.6	34.7
	Iter 2	**34.9**	**35.4**	**35.6**	27.7	**34.1**	**34.9**

Table 1: Comparison of the final unsupervised MT performance (BLEU). In this table, "Iter 2" means we do the whole process with another iteration as described in Section 3.4.

and both the improvements of en2fr and ro2en are over 2 BLEU points. The results indicate that the explicit cross-lingual information incorporated by our proposed CMLM is beneficial to the unsupervised machine translation task. Notice that by doing another iteration ("Iter 2") with updated n-gram tables as described in Section 3.4, we further improve the performance a bit for most translation directions with the improvements of en2fr and ro2en bigger than 0.5 BLEU point, which confirms the potential that fine-tuned machine translation models contain more beneficial cross-lingual information than the initial n-gram translation tables, which can be used to enhance the pre-trained model iteratively.

The improvement made by Lample and Conneau (2019) compared with the first five baselines shows that cross-lingual pre-training can be necessary for unsupervised MT. However, the cross-lingual information learned with this method during pre-training is mostly from the shared subword space, which is inexplicit and not strong enough. Our proposed method can give the model more explicit and strong cross-lingual training signals so that the pre-trained model contains much beneficial cross-lingual information for unsupervised machine translation. As a result, we can further improve the translation performance significantly, compared with Lample and Conneau (2019) (with the significance level of $p<0.01$).

4.3 Usage of Pre-trained Decoder

As mentioned in Section 3.4, it is interesting to explore the different usage of pre-trained decoders in the MT task. According to our intuition, directly using the pre-trained model as the decoder may not work well because parameters of the decoder need substantial adjustment due to the attention part between the encoder and the decoder. Therefore, we treat the pre-trained decoder as the feature extractor and add several Transformer layers with the encoder-to-decoder attention on top of it. We also try to fix the pre-trained decoder and just fine-tune the encoder and the added decoder part. The experiments are conducted based on "Iter 1" with the results reported in Table 2.

From this table, we can see that directly using the pre-trained model as the decoder may be the best choice for most of the time, with the exceptions of en2fr and ro2en. By adding 6 Transformer layers on top of the original pre-trained decoder can achieve higher performance for en2fr and ro2en, but not significant. The reason could be that it is difficult to train the additional Transformer layers from scratch in the unsupervised scenario. There is also an interesting phenomenon that if we fix the pre-trained part of the decoder and only tune the added Transformer layers, the final performance will drop drastically, which indicates that the representation space of the decoder requires substantial adaptation, even though the pre-trained models already contain cross-lingual information. We think that further deep research on the decoder initialization could be a necessary and interesting topic in the future.

4.4 Evaluation of Cross-lingual Pre-trained Encoder

Word Alignment

To evaluate the cross-lingual modeling capacity of our pre-trained models, we first conduct experiments on the English-French (en-fr) dataset of the HLT/NAACL 2003 alignment shared task (Mihalcea and Pedersen, 2003). Given two parallel sentences in English and French respectively, we feed each sentence into the pre-trained cross-lingual encoder and get its respective outputs. Then, we calculate the similarities between the outputs of the two sentences and choose target words with max similarity scores as the alignments of corre-

Decoder type	fr2en	en2fr	de2en	en2de	ro2en	en2ro
Pre-trained	**34.8**	34.5	**35.5**	**27.9**	33.4	**34.7**
Pre-trained + 4 TF layers	34.2	33.9	34.9	26.8	32.5	33.2
Pre-trained + 6 TF layers	34.6	**34.9**	34.9	27.5	**33.6**	34.1
Pre-trained (fix) + 4 TF layers	26.8	22.0	23.9	19.2	24.7	25.9
Pre-trained (fix) + 6 TF layers	28.2	22.4	24.2	19.7	25.1	26.2

Table 2: Test BLEU scores with different usage of the pre-trained decoder.

sponding source words.

We compare the context-unaware method (i.e., directly calculating the similarity scores between unsupervised cross-lingual embeddings (Artetxe et al., 2018a) of source and target words), XLM (Lample and Conneau, 2019) and our proposed CMLM pre-training method in the Table 3. In this experiment, we leave out all the OOV words and those torn apart by the BPE operations.

Method	P	R	F	AER
Context-unaware	0.3860	0.1854	0.2505	0.6061
XLM	0.5480	0.3178	0.4023	0.4302
Ours	**0.5898**	**0.3497**	**0.4391**	**0.4016**

Table 3: Results of word alignment tasks using different cross-lingual word embeddings. In this table, "P" means "precision", "R" means recall", "F" means "F-measure" and "AER" means the "alignment error rate".

From this table, we find that, based on BERT, both XLM and our method can model cross-lingual context information, indicating that context information can greatly enhance the cross-lingual mapping between the source and target words. By leveraging the explicit cross-lingual information in the model training, our CMLM can outperform XLM significantly. This confirms that our CMLM does better to connect the source and target representation space, with which as pre-trained models, the performance of unsupervised NMT can be improved.

Cross-lingual Classification

We also conduct experiments on the cross-lingual classification task (Conneau et al., 2018) using the cross-lingual language inference (XNLI) dataset (Conneau et al., 2018). Specifically, we add a linear classification layer on top of the first hidden state of our pre-trained model and fine-tune its parameters on the English NLI dataset. Without using any labeled data for French (fr) and Germany (de) languages, we only report the zero-shot classification results for them as shown in Table 4. We can find that our method reaches a new record of the zero-shot cross-lingual classification task on languages of French (fr) and Germany (de), which confirms again that our CMLM works better on modeling cross-lingual information than previous methods in the unsupervised scenario.

Method	en	fr	de
(Conneau et al., 2018)	73.7	67.7	67.7
(Devlin et al., 2018)	81.4	-	70.5
(Lample and Conneau, 2019)	83.2	76.5	74.2
Ours	**83.4**	**77.1**	**74.7**

Table 4: Results of zero-shot cross-lingual classification (on XNLI test sets).

4.5 Ablation Study

In this section, we will discuss the different settings of our method. Firstly, the training loss of our pre-trained method contains two parts, i.e., CMLM and MLM, just as Eq.(5) shows. To study the respective influences of these two parts, we remove the MLM loss from it and compare the performance on en-fr and en-de translation tasks. Since our CMLM task differs from XLM in two aspects during pre-training. The first one is that we randomly choose n-grams to mask in the input text stream rather than BPE tokens, and the second one is that we predict the translation candidates of a source n-gram rather than predicting the source n-gram itself. Although the first one has proven to be beneficial during pre-training to some NLP tasks, we want to check how much its influence is to our final translation performance. Therefore, we disable those two modifications in CMLM one by one and report the translation results. Our experiments are conducted based on "Iter 1" with the results in Table 5.

From Table 5, we can find that the combination of CMLM and MLM can improve the translation performance by about 0.6 to 0.7 BLEU compared with any one only. This confirms the monolingual context modeling capacity of the MLM, which is quite useful for unsupervised machine translation. By combining CMLM and MLM, we can enforce

	fr2en	en2fr	de2en	en2de
CMLM + MLM	**34.8**	**34.9**	**35.5**	**27.9**
CMLM	34.1	34.3	35.1	27.2
- translation prediction	33.7	33.9	34.8	26.6
- - n-gram mask	33.3	33.4	34.3	26.4

Table 5: Ablation study. "CMLM + MLM" means we use $\mathcal{L}_{pre}$ as the pre-training loss; "CMLM" means we only use $\mathcal{L}_{cmlm}$ as the pre-training loss; "- translation prediction" means we predict the masked n-grams themselves rather than their translation candidates during pre-training; "- - n-gram mask" means we randomly mask BPE tokens rather than n-grams based on "- translation prediction" during pre-training, which degrades our method to XLM.

our model to learn both monolingual and cross-lingual information during pre-training. Besides, we find the two modifications(translation prediction and n-gram mask) made by CMLM have nearly equal contributions to the translation performance, except for en2de, where the "n-gram mask" has little influence.

5 Related Work

Unsupervised machine translation dates back to Klementiev et al. (2012); Nuhn et al. (2012), but becomes a hot research topic in recent years. As the pioneering methods, Artetxe et al. (2017); Lample et al. (2017); Yang et al. (2018) are mainly the modifications of the encoder-decoder structure. The core idea is to constrain outputs of encoders of two languages into a same latent space with a weight sharing mechanism such as using a shared encoder. Denoising auto-encoder (Vincent et al., 2010) and adversarial training methods are also leveraged. Besides, they apply iterative back-translation to generated pseudo data for cross-lingual training. In addition to NMT methods for unsupervised machine translation, some following work shows that SMT methods and the hybrid of NMT and SMT can be more effective (Artetxe et al., 2018b; Lample et al., 2018; Marie and Fujita, 2018; Ren et al., 2019). They all build unsupervised PBSMT systems, and all of their models are initialized with language models and phrase tables inferred from cross-lingual word or n-gram embeddings and then use the initial PBSMT models to do iterative back-translation. Lample et al. (2018) also build a hybrid system by combining the best pseudo data that SMT models generate into the training of the NMT model while Ren et al. (2019) alternately train SMT and NMT models with the framework of posterior regularization.

More recently, Lample and Conneau (2019) reach new state-of-the-art performance on unsupervised en-fr and en-de translation tasks. They propose a cross-lingual language model pre-training method based on BERT (Devlin et al., 2018), and then treat two cross-lingual language models as the encoder and decoder to finish the translation. Leveraging much more monolingual data from Wikipedia, their work shows a big potential of pre-training for unsupervised machine translation. However, the cross-lingual information is obtained mostly from the shared BPE space during their pre-training method, which is inexplicit and limited. Therefore, we figure out a new pre-training method that gives the model more explicit and stronger cross-lingual information.

In the recent work of Song et al. (2019), they also mask several consecutive tokens in the source sentence, but jointly pre-train the encoder and decoder by making the decoder to predict those masked tokens in both the source and target sides. Their method is a good case of pre-training for seq-to-seq tasks but the cross-lingual information incorporated with their method is from BPE sharing, which is also implicit. Our proposed method can be combined with their method within a multi-task framework, which could be done in the future.

6 Conclusion

In this paper, we propose a novel cross-lingual pre-training method for unsupervised machine translation. In our method, we leverage Cross-lingual Masked Language Model (CMLM) to incorporate explicit and strong cross-lingual information into pre-trained models. Experimental results on en-fr, en-de, and en-ro language pairs demonstrate the effectiveness of our proposed method.

In the future, we may apply our pre-training method to other language pairs and delve into the performance of the pre-trained encoders on other NLP tasks, such as Name Entity Recognition.

Acknowledgments

This work is supported in part by National Key R&D Program of China 2018YFB1700403, and NSFC U1636210&61421003.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

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*, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

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.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. arXiv:1901.07291.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages 1–10.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 156–164. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *arXiv preprint arXiv:1901.04112*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, volume 1, pages 1715–1725.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

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(Dec):3371–3408.

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)*, volume 1, pages 46–55.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.