Bilingual Alignment Pre-Training for Zero-Shot Cross-Lingual Transfer

Ziqing Yang[1]*, **Wentao Ma**[1]*, **Yiming Cui**[2,1], **Jiani Ye**[1], **Wanxiang Che**[2], **Shijin Wang**[3,4]
[1]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, China
[2] Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
[4]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[1,3,4]{zqyang5,wtma,ymcui,jnye,sjwang3}@iflytek.com
[2]{ymcui,car}@ir.hit.edu.cn

Abstract

Multilingual pre-trained models have achieved remarkable performance on cross-lingual transfer learning. Some multilingual models such as mBERT, have been pre-trained on unlabeled corpora, therefore the embeddings of different languages in the models may not be aligned very well. In this paper, we aim to improve the zero-shot cross-lingual transfer performance by proposing a pre-training task named Word-Exchange Aligning Model (WEAM), which uses the statistical alignment information as the prior knowledge to guide cross-lingual word prediction. We evaluate our model on multilingual machine reading comprehension task MLQA and natural language interface task XNLI. The results show that WEAM can significantly improve the zero-shot performance.

1 Introduction

Large-scale multilingual pre-trained language models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020) have shown significant effectiveness in transfer learning on various cross-lingual tasks. The pre-training methods of the multilingual language models can be divided into two groups: unsupervised pre-training like Multilingual Masked Language Model (MMLM) (Devlin et al., 2019; Conneau et al., 2020), and supervised pre-training like Translation Language Model (TLM) (Conneau and Lample, 2019). In the MMLM, the model predicts the masked tokens with the monolingual context; in the TLM, the model can attend to both the contexts in the source language and target language. Variations of TLM model can be found in Huang et al. (2019); Chi et al. (2021); Ouyang et al. (2020).

While it is possible for the model to learn the alignment knowledge by itself, some works have investigated injecting prior knowledge to the model to help it to align better. Cao et al. (2020) proposed a bilingual pre-training model for mBERT, where it identifies matched word pairs in parallel bilingual corpora using unsupervised standard techniques such as FastAlign (Dyer et al., 2013), and aligns the contextual representations between the matched words with a similarity loss function.

The previous works focus on aligning the contextual representations of the pre-trained models. In this paper, we propose a new cross-lingual pre-trained model called Word-Exchange Aligning Model (WEAM). Different from previous works, we align the static embeddings and the contextual representations of different languages in the multilingual pre-trained models.

Specifically, in the pre-training stage, we first use FastAlign to identify bilingual word pairs in parallel bilingual sentence pairs as our prior knowledge. Then we randomly mask some tokens in the bilingual sentence pairs. For each masked token, WEAM performs two kinds of predictions: a multilingual prediction and a cross-lingual prediction. The multilingual prediction task predicts the original masked word in the standard way. while the cross-lingual task predicts the corresponding word from the representations in the other language. For example, if the words *apple* and *Apfel* (German for *apple*) appear in the the English–German parallel sentence and *apple* is masked in the sentence, WEAM takes the representation of the masked *apple* and *Apfel* for multilingual prediction and cross-lingual prediction respectively to recover the original word *apple*.

Through the two ways of prediction, both the contextual representations from the last transformer layer and the static embeddings from the embedding layer can be aligned. We evaluated our method on the word-level machine reading comprehension task MLQA (Lewis et al., 2019) and the sentence-level classification task XNLI (Conneau et al.,

*Equal contribution.

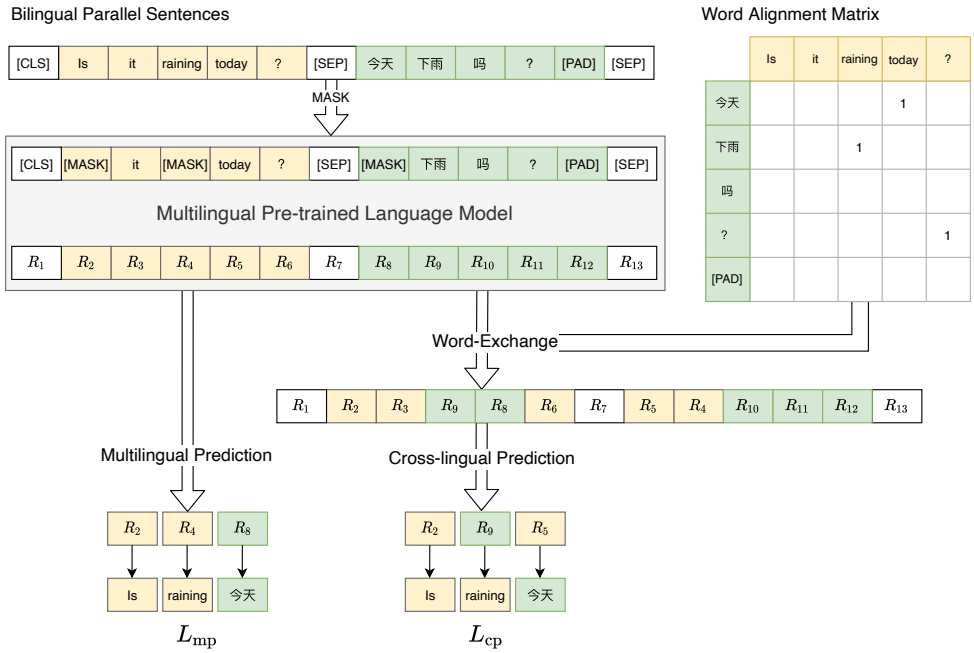

Figure 1: An overview of the Word-Exchange Aligning Model (WEAM). For each language pair, There are two tasks. The multilingual prediction task predicts the masked tokens. The cross-lingual prediction task utilizes a word alignment matrix to swap the representations of aligned words in parallel sentences, then predicts the masked tokens in the swapped sentences.

2018). The results show that WEAM significantly improves the cross-lingual transfer performance.

2 Methodology

2.1 Translation Language Model

We first briefly describe the Translation Language Model (TLM) (Conneau and Lample, 2019). Like MMLM in (Devlin et al., 2019), TLM performs the masked word prediction task, where it randomly masks some words and predicts the original ones within a parallel sentence pair. For each masked word, the model can either attend to the surrounding words or the translated context in the other language, encouraging the model to align the words in different languages.

2.2 Word-Exchange Aligning Model

Our proposed method WEAM is based on the multilingual pre-trained model and consists of two tasks: the multilingual prediction task and the cross-lingual prediction task, as shown in Figure 1.

Multilingual Prediction. In the multilingual prediction, we randomly mask tokens in the bilingual parallel sentences and predict the original tokens with the outputs from the last transformer layer. unlike TLM, we did not reset the position embeddings or add the language embeddings, so the distinction between languages will be purely learned from the token embeddings. We construct the inputs and obtain the representations for a source-target sentence pair $\langle S, T \rangle$ as

$$X = \texttt{[CLS]}\, S\, \texttt{[SEP]}\, T\, \texttt{[SEP]} \quad (1)$$

$$H = \texttt{Encoder}(X) \quad (2)$$

where X is the token sequence and $H \in \mathbb{R}^{m \times h}$ is the output from the last transformer layer of the pre-trained model $\texttt{Encoder}$; m is the max sequence length and h is the hidden size. For a masked token X_i, we predict the original token w_i with the corresponding representation

$$H'_i = \delta(W_1 \cdot H_i + b_1) \quad (3)$$

$$p(X_i = w_i | H'_i) = \frac{\exp(\texttt{linear}(H'_i) \cdot e_i)}{\sum_{k=1}^{|\mathcal{V}|} \exp(\texttt{linear}(H'_i) \cdot e_k)} \quad (4)$$

where δ is the GELU activation (Hendrycks and Gimpel, 2016), $\texttt{linear}(\cdot)$ is a linear layer, H_i is the token representation for X_i, as given by Eq (2). $|\mathcal{V}|$ is the vocabulary size. e_i is the emebdding vector of token w_i.

Cross-lingual prediction. In the cross-lingual prediction, we predict the masked tokens with the representations from the other language. Specifically, we first use FastAlign to construct an

Model	en	es	de	zh	AVG(all)	AVG(zero-shot)
Translate-Train						
mBERT †	65.2/77.7	37.4/53.9	47.5/62.0	39.5/61.4	47.4/63.8	43.0/60.3
mBERT (ours)	67.3/80.3	48.4/67.1	49.1/63.5	42.8/63.6	51.9/68.6	48.1/65.7
Zero-Shot						
mBERT †	65.2/77.7	46.6/64.3	44.3/57.9	37.3/57.5	48.4/64.4	44.2/61.0
mBERT+TLM	**66.8/80.0**	47.7/65.7	48.4/63.1	40.1/62.0	50.7/67.7	46.7/64.6
mBERT+WEAM	66.7/79.7	**49.6/67.8**	**49.7/64.3**	**41.7/63.7**	**51.7/68.9**	**48.2/66.2**

Table 1: EM/F1 scores on the test set of MLQA dataset. The results with † are taken from Lewis et al. (2019). *AVG(all)* is the average scores on all languages. *AVG(zero-shot)* is the average scores on the languages excluding English.

alignment words set from parallel sentences $\langle S, T \rangle$. We denote the words set as $d(s,t) = \{(i_1, j_1), ..., (i_n, j_n)\}$, where i is the word index of source language in the input sequence, j is the word index of the target language. n is the number of word pairs in the sentence pair. Then we generate effectively code-mixed representations by exchanging the positions of each word pair in parallel sentences. We denote the exchange operation with an off-diagonal matrix $A \in \{0,1\}^{m \times m}$:

$$A(i,j) = \begin{cases} 1, & \text{if } \{(i,j) \textit{ or } (j,i)\} \in d \\ 0, & \text{otherwise} \end{cases}$$

We take A as the transformation matrix to construct the word-exchange representations $H^{'}$, which is calculated by

$$H' = A^{\mathrm{T}} \cdot H \tag{5}$$

$$\tilde{H} = W_2 \cdot H' + b_2 \tag{6}$$

We have applied another linear transformation on $H^{'}$ and obtained $\tilde{H}$. Lastly, we conduct the masked word predictions on $\tilde{H}$ similar to the multilingual prediction:

$$\tilde{H}'_i = \delta(W_3 \cdot \tilde{H}_i + b_3) \tag{7}$$

$$\tilde{p}(X_i = w_i | \tilde{H}'_i) = \frac{\exp(\texttt{linear}(\tilde{H}'_i) \cdot e_i)}{\sum_{k=1}^{|\mathcal{V}|} \exp(\texttt{linear}(\tilde{H}'_i) \cdot e_k)} \tag{8}$$

If the word w_i is paired with word w_j, what the cross-lingual prediction does is predicting w_i with the contextual representation of w_j. In this way we are align the embedding of w_i (e_i) with the contextual representation of w_j (H_j).

Pre-training Objective. Given a bilingual parallel corpus $\mathcal{D}$, we train the multilingual model with the cross-entropy loss. Based on the discussion above, the objective function of pre-training consists of multilingual part L_{mp} and cross-lingual prediction part L_{cp}. Let Θ denote the parameters of the model, then the objective function $L(\mathcal{D}, \Theta)$ can be formulated as

$$L_{\text{mp}} = -\sum_{i=1}^{M} \log(p(w_i)) \tag{9}$$

$$L_{\text{cp}} = -\sum_{i=1}^{M} \log(\tilde{p}(w_i)) \tag{10}$$

$$L(\mathcal{D}, \Theta) = L_{\text{mp}} + \lambda L_{\text{cp}} \tag{11}$$

where M is the number of masked tokens in the instance, $p(w_i)$ and $\tilde{p}(w_i)$, given by Eq.(6) and Eq.(8), are the predicted probability of the masked token w_i over the vocabulary size, λ is a hyper-parameter to balance L_{mp} and L_{cp}.

3 Experiments

3.1 Experiment Setup

We use three parallel corpora with the source language English and the target languages Chinese [1], German and Spanish [2] respectively. We initialize the mBERT model with the weights released by Google[3]. We pre-train three models for the three target languages separately to avoid alignment interference among different language pairs.

During the pre-training steps, we empirically set the masking probability as 0.3. Experimentally we find that 0.3 gives better performance. The other settings for masking are the same as the MLM (Devlin et al., 2019). The hyper-parameters of the three models are the same: we set the learning rate as 5e-5, the batch size as 32, the max sequence length as 128, and the number of pre-training epochs as 2. We set λ to 1.

[1] We use the corpus from Xu (2019).

[2] http://www.statmt.org/europarl

[3] https://github.com/google-research/bert

Model	en	es	de	zh	AVG(all)	AVG(zero-shot)
Translate-Train						
mBERT	82.1	77.8	75.9	75.7	77.9	76.5
Zero-Shot						
mBERT†	82.1	74.3	71.1	69.3	74.2	71.6
Word-aligned BERT†	80.1	75.5	73.1	-	-	-
mBERT+TLM	82.0	75.0	73.5	73.1	75.9	73.9
mBERT+WEAM	**82.6**	**76.4**	**74.5**	**74.4**	**77.0**	**75.1**

Table 2: Accuracy scores on XNLI dataset. The results with † are taken from Conneau et al. (2020).

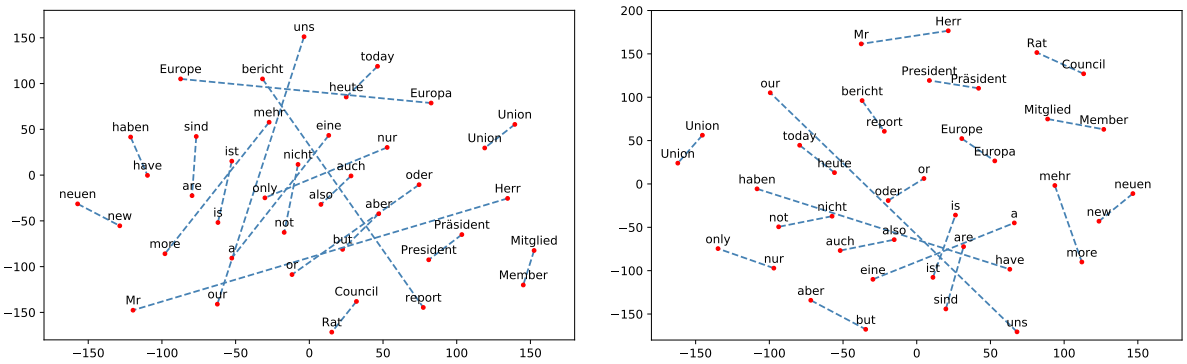

Figure 2: A visualization of the word embeddings from mBERT before and after WEAM pre-training. We select 20 English-German word alignment pairs that appear most frequently in the pre-training corpus. Each word alignment pair is connected by a blue dotted line. All the word pairs are identified by FastAlign (Dyer et al., 2013).

For the downstream evaluation, we fine-tune and test our pre-trained model along with several baselines on the MLQA and XNLI tasks respectively. The specific settings of baselines are described in the following section. Since in this work we mainly focus on evaluating the zero-shot performance, we fine-tune all the models in the zero-shot setting where only the English training set is available. We also fine-tune mBERT in the translate-train setting for comparison.

3.2 Baselines

We use mBERT (Devlin et al., 2019) as our main baseline, which consists of 12 transformer layers, with a hidden size of 768 and 12 attention heads. For a fair comparison, we also include a baseline mBERT+TLM with the same pre-training settings but uses TLM as the pre-training task. An additional baseline word-aligned mBERT from Cao et al. (2020) is included for the XNLI dataset.

3.3 Results on MLQA

Table 1 shows our results on MLQA. Note that the results on the target languages of the TLM and WEAM are from models of different language pairs as introduced in the experiment setup section. The results of TLM and WEAM on English are the average of the three models.

The mBERT+TLM model outperforms mBERT by a large margin in the zero-shot setting, but is not as good as the mBERT in the translate-train setting. Our model mBERT+WEAM improves the scores in the zero-shot setting and also outperforms mBERT in the translate-train setting. This result is promising, as it indicates that a properly aligned pre-training model can exceed the performance of translate-train even with zero-shot training.

3.4 Results on XNLI

Table 2 shows our results on XNLI. The mBERT+TLM and word-aligned mBERT achieved similar improvements on this task compared to mBERT, whereas mBERT+WEAM has significantly outperformed both of them. Because all of these models are pre-trained with the same parallel corpus, the differences in performance indicate the importance of considering both the word-level and contextual-level alignment. Compared with the translate-train result, the mBERT+WEAM result is slightly lower but is close. This is different from MLQA. This observation may indicate that the examples in XNLI have shorter input sequences and

thus have fewer translation noises.

4 Visualization

The effect of contextual alignment has been well studied in Cao et al. (2020), where the authors demonstrate that the contextual alignment is powerful in improving the transferability of mBERT. but the effect of the word-level information alignment is still unclear. To further explore this problem, we use t-SNE (Maaten and Hinton, 2008) to visualize the distances between embeddings of word alignment pairs with the highest frequencies (excluding stop words). The result is illustrated in Figure 2.

The left panel shows word pairs in the embedding layer of mBERT without WEAM pre-training, we can see that these word pairs are partly aligned. For example, the pairs *today-heute*, *Council-Rat* are aligned well, but *Beriche-report*, *Mr-Herr* are distant away. As a comparison, we show the word pairs from the embedding layer of mBERT with WEAM pre-training in the right panel, where most of the word pairs are aligned much better. There are also words that remained poorly aligned even with WEAM. For example, *our-uns*, which may be due to that they are not the exact translation pair (*us-uns* are more exact pairs in this case). In general, the embeddings are aligned much better after the WEAM pre-training procedure.

5 Conclusion

In this paper, we propose a new pre-training task named WEAM to align the contextual representations and static word embeddings from multilingual pre-trained models. WEAM consists of a multilingual prediction task and a cross-lingual prediction task. As a supplement to previous works MMLM or TLM, WEAM introduces the statistic alignment information as prior knowledge to guide the cross-lingual prediction. Through the experiments on MLQA and XNLI, we show that WEAM can improve the transfer performance significantly and align the word embeddings within the models much better. In the future, we plan to extend our method to other multilingual models like XLM-R.

References

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Eval-

uating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-M: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *CoRR*, abs/2012.15674.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.