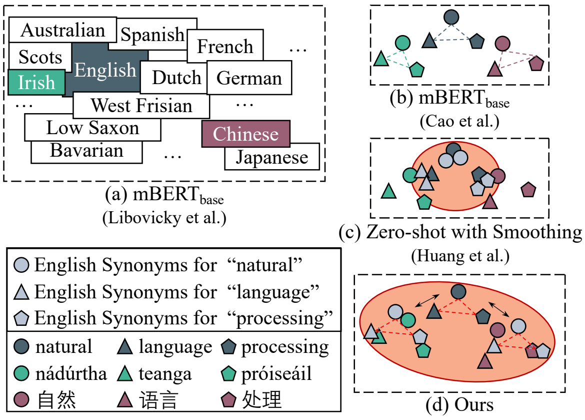
A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning

Kunbo Ding[1] [*], Weijie Liu[1,2] [†], Yuejian Fang[1], Weiquan Mao[2], Zhe Zhao[2]
Tao Zhu[2], Haoyan Liu[2], Rong Tian[2], Yiren Chen[2]
[1]Peking University, Beijing, China [2]Tencent Research, Beijing, China
kunbo_ding@stu.pku.edu.cn, dataliu@pku.edu.cn, fangyj@ss.pku.edu.cn
{weiquanmao, nlpzhezhao, mardozhu, haoyanliu, rometian, yirenchen}@tencent.com

Abstract

Existing zero-shot cross-lingual transfer methods rely on parallel corpora or bilingual dictionaries, which are expensive and impractical for low-resource languages. To disengage from these dependencies, researchers have explored training multilingual models on English-only resources and transferring them to low-resource languages. However, its effect is limited by the gap between embedding clusters of different languages. To address this issue, we propose Embedding-Push, Attention-Pull, and Robust targets to transfer English embeddings to virtual multilingual embeddings without semantic loss, thereby improving cross-lingual transferability. Experimental results on mBERT and XLM-R demonstrate that our method significantly outperforms previous works on the zero-shot cross-lingual text classification task and can obtain a better multilingual alignment.



(a) mBERT$_{base}$ (Libovicky et al.)

(b) mBERT$_{base}$ (Cao et al.)

(c) Zero-shot with Smoothing (Huang et al.)

(d) Ours

Figure 1: (a) Different languages clusters in mBERT. (b) The relative positions of "nature", "language" and "processing" are similar in English, Chinese and Irish (Cao et al., 2020). (c) Using synonym augmentation to train a robust region covering words in other languages. (d) We align different languages and construct a suitable robust region by pushing the embeddings away and pulling the relative distance among words.

1 Introduction

In recent years, advances in multilingual models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), etc., after being fine-tuned with annotated data, have enabled significant improvements in many cross-lingual tasks. However, due to the lack of annotated data, some tasks in low-resource languages have not enjoyed this technological advancement. To solve this issue, the academic and industrial community began to focus on zero-shot cross-lingual transfer learning (Huang et al., 2019; Artetxe et al., 2020), which aims to fine-tune multilingual models with annotated data in high-resource languages and obtain a nice performance in low-resource language tasks.

Some works aligned word embeddings between high- and low-resource languages through additional parallel sentence pairs (Artetxe and Schwenk, 2019; Wei et al., 2021; Chi et al., 2021; Pan et al., 2021) or bilingual dictionaries (Cao et al., 2020; Qin et al., 2020; Liu et al., 2020), so that high-resource fine-tuned models can be transferred to low-resource languages. Although this approach has achieved excellent results in many languages, parallel corpora and bilingual dictionaries are still prohibitively expensive, rendering it impracticable in some minority languages.

To disengage from the dependence on parallel corpora or bilingual dictionaries (Wu and Dredze, 2019; Hu et al., 2020), some studies have found that syntactic features in high-resource languages can improve zero-shot cross-lingual transfer learning (Meng et al., 2019; Subburathinam et al., 2019; Ahmad et al., 2021a,b). Libovický et al. (2020) found that the embeddings of different languages are clustered according to their language families, as shown in Figure 1a and 1b, which demonstrated that different languages are not aligned perfectly

[*] Contribution during internship at Tencent Inc.
[†] Corresponding author: Weijie Liu.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4372–4380
October 12–17, 2022.

in mBERT (Deshpande et al., 2021). Huang et al. (2021) tried adversarial training and randomized smoothing with English synonym augmentation to build robust regions for embeddings in the multilingual models, as illustrated in Figure 1c. In this way, models can output similar predictions for different language embeddings in the same robust region even they are not well aligned. However, the transferability of English synonym augmentation is limited because its robust region remains close to the English cluster, as shown in Figure 1c.

In this work, we select English as a high-resource language and follow the studies that do not require additional parallel corpora or bilingual dictionaries to improve cross-lingual transfer learning performance with minimal cost. For this purpose, three strategies are proposed to enlarge the robust region of English embeddings. The first strategy is called *Embedding-Push*, which pushes the embedding of English to other language clusters. The second is *Attention-Pull*, which constrains the relative position of the word embeddings to prevent the meaning from straying. The last strategy, named *Robust target*, introduces a Virtual Multilingual Embedding (VME) to help the model build a suitable robust region, as shown in Figure 1d.

Experimental results on mBERT and XLM-R demonstrate that our method effectively improves the zero-shot cross-lingual transfer on classification tasks and outperforms a series of previous works. In addition, case studies show that our method improves the model through multilingual word alignment. Compared with existing works, our method has the following advantages. First, our method only needs English resources, which is suitable for low-resource languages. Second, our method can induce alignments in many languages without specifying the target language. Finally, our method is simple to implement and achieves effective experimental results. Our code is publicly available[1].

2 Method

Given an English training batch $\mathcal{B}$, for a specific $\boldsymbol{x} \in \mathcal{B}$ consisting of words (x_1, x_2, x_3), we first follow Huang et al. (2021) to generate an augmented example $\boldsymbol{x^a} = (x_1^a, x_2^a, x_3^a)$ by randomly replacing x_i with x_i^a from the pre-defined English synonym set (Alzantot et al., 2018). Then, we introduce three objective functions to get the Virtual Multilingual Embedding (VME) that provides a suitable robust region for zero-shot cross-lingual classification task as shown in Figure 2. We describe the details in the following subsections.

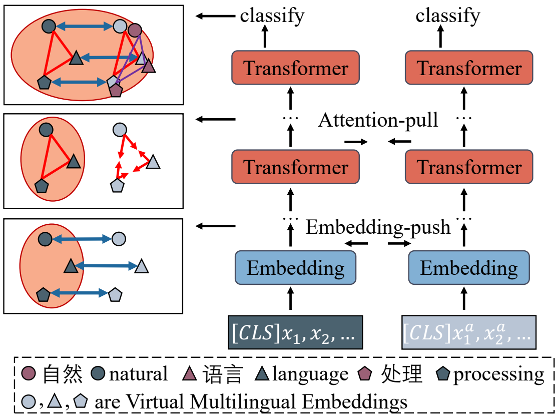


Figure 2: The two networks have tied weights. VMEs expand robust regions (orange circle) by aligning semantic-similar words in other languages. Note that VMEs do not specify the target language but improve multilingual performance, as shown in section 3.3.

2.1 Embedding-push target

The Embedding-Push target aims to make English embeddings leave their original cluster and robust region by pushing away $(\boldsymbol{x}, \boldsymbol{x^a})$ in the embedding space. The pushed embedding can be viewed as the VME. The loss function is (1).

$$\ell_{EPT} = -\frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} (M(E_{\boldsymbol{x}}) - M(E_{\boldsymbol{x^a}}))^2 \quad (1)$$

where $E_{\boldsymbol{x}}, E_{\boldsymbol{x^a}}$ denote the embedding output of $\boldsymbol{x}$ and $\boldsymbol{x^a}$, M is the mean-pooling method.

2.2 Attention-pull target

The self-attention matrices contain rich linguistic information (Clark et al., 2019) and can be regarded as a 1-hop graph attention between the hidden states of words (Vaswani et al., 2017; Veličković et al., 2018). The attention matrix represents the information transfer score between each pair of words, we regard it as the pulling force, so the attention matrix determines the relative linguistic positions of words in a sentence. We introduce the Attention-Pull target to encourage the relative linguistic position among (x_1^a, x_2^a, x_3^a) to be similar to (x_1, x_2, x_3) by fitting the middle layer multi-head attention matrices, as (2).

$$\ell_{APT} = \frac{1}{|\mathcal{B}|H} \sum_{\boldsymbol{x} \in \mathcal{B}} \sum_{i}^{H} (A_{\boldsymbol{x}}^i - A_{\boldsymbol{x^a}}^i)^2 \quad (2)$$

[1] https://github.com/KB-Ding/EAR

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
mBERT†	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
+ADV†	81.9	64.9	68.3	71.7	66.5	74.4	74.5	59.6	68.8	48.8	50.6	61.7	59.2	70.0	69.4	66.0
+RS-RP†	82.6	65.4	68.7	70.5	67.2	75.0	74.1	59.8	69.5	48.4	50.5	59.7	57.9	70.5	69.7	66.0
+RS-DA†	81.0	66.4	69.9	71.8	68.0	74.7	74.2	62.7	70.6	51.1	55.7	62.9	60.9	71.8	71.4	67.6
+Syntax‡	81.6	65.4	69.3	70.7	66.5	74.1	73.2	60.5	68.8	-	-	62.4	58.7	69.9	69.3	-
+ Ours	**83.2**	**67.4**	**71.0**	**72.9**	**68.3**	**75.7**	**75.2**	**64.0**	**71.6**	**51.3**	**56.7**	**63.6**	**61.4**	**72.4**	**71.5**	**68.4**
XLM-R*	84.0	72.6	**78.9**	77.0	76.5	78.6	78.2	70.3	76.4	65.0	72.4	73.4	67.6	75.5	75.0	74.8
+RS-DA*	83.5	73.2	78.2	77.1	76.9	79.2	79.0	72.3	**76.9**	66.5	73.2	73.1	68.2	**76.4**	75.1	75.3
+ Ours	**84.6**	**74.5**	78.8	**77.5**	**77.0**	**79.4**	**79.5**	**72.6**	76.8	**66.7**	**73.9**	**74.7**	**68.7**	**76.4**	**75.8**	**75.8**

Table 1: Zero-shot cross-lingual transfer results on the XNLI. We bold the highest accuracy scores (%). "†" and "‡" are taken from (Huang et al., 2021) and (Ahmad et al., 2021a), respectively. "$*$" is the result of our reimplementation.

Model	en	de	es	fr	ja	ko	zh	avg.
mBERT†	94.0	85.7	87.4	87.0	73.0	69.6	77.0	82.0
+ADV†	93.7	86.5	88.5	87.8	76.1	75.3	80.4	84.0
+RS-RP†	**94.5**	87.4	90.0	89.5	77.9	77.5	82.0	85.5
+RS-DA†	93.5	87.8	88.8	88.8	79.3	78.3	81.5	85.4
+Syntax‡	94.0	85.9	89.1	88.2	75.8	76.3	80.7	84.3
+Ours	94.2	**87.9**	**90.3**	**89.7**	**79.9**	**79.2**	**82.4**	**86.2**
XLM-R*	94.4	88.9	89.8	89.2	78.2	78.4	81.4	85.7
+RS-DA*	94.7	88.8	89.7	90.0	78.7	80.2	82.3	86.3
+Ours	**95.1**	**89.0**	**90.3**	**90.1**	**80.5**	**81.7**	**83.1**	**87.1**

Table 2: Experimental results on the PAWS-X across 7 languages. "†" and "‡" are taken from (Huang et al., 2021) and (Ahmad et al., 2021a), respectively. "$*$" is the result of our reimplementation.

where H is the number of attention head. Let L denote the sequence length, $A^i \in \mathbb{R}^{L\times L}$ is the attention matrix corresponding to the i-th head. ℓ_{APT} alleviates the semantic loss of the VME.

2.3 Robust target

The robust target aims to build a robust region with the VME for the classification task. The hidden state of [CLS] in the last layer is taken to classify, as (3). The model is trained by (4).

$$P_n = \mathrm{softmax}(\boldsymbol{W} h_n^{[\mathrm{CLS}]} + \boldsymbol{b}) \tag{3}$$

$$\ell_{CE} = -\frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{x}\in\mathcal{B}} (y \log P_{\boldsymbol{x}} + y \log P_{\boldsymbol{x}^a}) \tag{4}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters. P_n is the prediction for n. y denotes the gold label for each $\boldsymbol{x} \in \mathcal{B}$. The final training objective is to minimize three targets as (5):

$$\ell = \ell_{CE} + \alpha\ell_{EPT} + \beta\ell_{APT} \tag{5}$$

where α and β are hyperparameters.

3 Experiment

3.1 Dataset and setup

We use mBERT$_{\text{base}}$ and XLM-R$_{\text{base}}$ to evaluate our method on XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) tasks, covering 17 languages. We consider English as the source language and other languages in test sets as low-resource target languages. More training details are in Appendix A. We set α=1, β=0.1 and apply the Attention-Pull target at the 6-th layer. The analysis of hyperparameters is in Appendix B. We measure results with accuracy.

3.2 Baseline methods

For XLM-R, we consider **RS-DA** as a strong baseline because it achieves the best performance. For mBERT, we consider all the following baselines.

Adv: Huang et al. (2021) uses adversarial training to build a robust region for cross-lingual transfer. They consider the most effective perturbation in each iteration.

RS-RP: Huang et al. (2021) perturbs sentence embeddings with randomly sampled δ to smooth the classifier and build robust regions.

RS-DA: Huang et al. (2021) augments training data with English synonym replacement to train a smooth classifier and build robust regions.

Syntax: Ahmad et al. (2021a) provides syntax features to mBERT by graph attention networks, which helps cross-lingual transfer.

3.3 Main results

As illustrated in Table 1 and Table 2. We can observe that: 1) Our method achieves up to 4.2% and 1.4% improvement on mBERT and XLM-R, respectively, outperforming existing works and demonstrating the effectiveness of our method. 2) Multiple low-resource languages benefit from our method. Based on mBERT, our method improves not only English-like languages such as **es** and **de** but also English-dissimilar (Littell et al., 2017) languages such as **tr** and **ko**. This result indicates that the VME we proposed helps align different languages in semantic space. 3) We avoid training

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
Ours	83.2	67.4	71.0	72.9	68.3	75.7	75.2	64.0	71.6	51.3	56.7	63.6	61.4	72.4	71.5	**68.4**
w/o EPT	82.8	67.0	71.2	72.7	67.6	75.5	75.1	63.4	71.2	51.1	56.0	63.4	60.7	72.5	71.4	68.1
w/o APT	82.4	66.5	70.8	72.8	68.5	76.0	75.1	63.4	70.9	50.0	57.0	62.8	61.9	72.0	72.3	68.2
w/o both	82.1	66.4	70.0	72.3	67.7	75.1	74.9	62.8	70.8	48.3	54.6	61.0	61.0	71.5	71.5	67.4

Table 3: Ablation experimental results of our method on the XNLI task. Experiments are based on mBERT.

each target language separately and achieves the best results in one epoch using the English-trained VME.

3.4 Ablation study

As shown in Table 3, we perform ablation studies on Embedding-Push Target (**EPT**) and Attention-Pull Target (**APT**). We find that both EPT and APT are effective, but they can not perform well alone. Besides, removing the APT causes improvement in some languages, such as **zh** and **ur**. We attribute this to the fact that the EPT-guided VME is unstable without the APT, which improves performance in some languages but drops in more languages such as **en**, **ar**, **ru**, etc., resulting in poor average performance. Thus EPT and APT need to be combined for better performance.

4 Analysis

4.1 Case study

To study the effects of VME, we do the T-SNE visualization for the word embeddings of parallel sentences, as shown in Figure 3. Compared with the RS-DA, our fine-tuned model aligns better across languages, and words are closer to their translations, leading to correct predictions. This observation shows that the VME can effectively help cross-lingual word alignment and improve the performance of the model. We choose Arabic for the case study because it can represent a class of languages far apart from English.

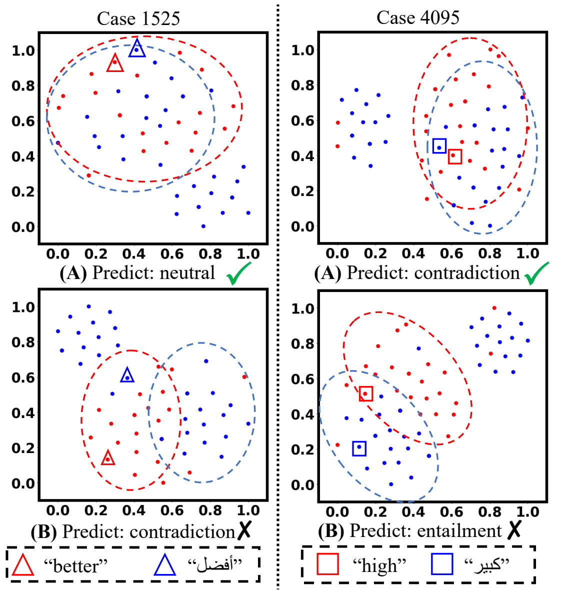


Figure 3: T-SNE visualization for word embeddings of English and Arabic translated sentences in XNLI test sets. Blue dots are Arabic words. Red dots are English words. (A) mBERT trained with our method. (B) mBERT trained with RS-DA.

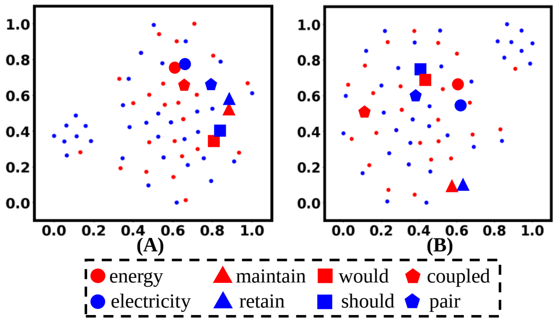


Figure 4: Visualization for English synonyms in the XNLI dataset using the embedding layer of mBERT. (A) Untrained. (B) Trained with our method.

Model	en	es	de	fr	bg	ru	el	th	sw	vi	ar	zh	hi	ur	tr	avg.
EPT + APT	83.2	75.7	72.9	75.2	71.0	71.6	68.3	56.7	51.3	72.4	67.4	71.5	64.0	61.4	63.6	**68.4**
NT + APT	82.7	75.6	72.5	75.2	70.6	71.0	67.9	55.9	50.6	72.3	66.9	71.8	63.4	61.0	62.9	68.0
EPT + SRPT	83.1	76.0	73.2	74.9	70.7	71.4	68.9	56.5	50.5	72.3	67.0	71.8	63.3	60.9	63.1	68.2

Table 4: Results on the XNLI task when replacing some targets, based on the mBERT. We sort languages according to their differences from English (Littell et al., 2017), from top left (small) to bottom right (big).

4.2 Effect of EPT

To study the impact of EPT, we do the T-SNE visualization using the embedding layer of mBERT. As shown in Figure 4, some synonyms such as "coupled / pair" and "energy / electricity" are pushed away in the embedding layer trained with EPT, and some synonyms are still close to their original words. It indicates that the EPT push away synonyms selectively. We also try to replace the EPT in (5) with the Noise Target (NT), which perturbs word embeddings with Gaussian noise (Cohen et al., 2019). As shown in Table 4, we find

Source Language	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	<u>avg.</u>
en	**83.2**	67.4	71.0	72.9	68.3	75.7	**75.2**	64.0	71.6	**51.3**	56.7	63.6	61.4	72.4	71.5	68.4
de	79.6	**68.7**	71.9	**77.7**	**68.8**	**76.2**	74.9	64.2	72.4	50.1	55.2	**64.0**	**62.8**	73.0	72.6	<u>**68.8**</u>
ru	78.5	68.2	**73.3**	73.1	**68.8**	74.8	73.9	**65.8**	**75.7**	49.3	**57.2**	**64.0**	62.4	**73.4**	**73.7**	<u>**68.8**</u>

Table 5: Results of our method on the XNLI task when training mBERT with three source languages.

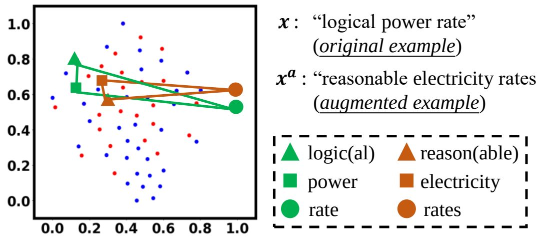


Figure 5: T-SNE visualization on the outputs of the mBERT trained with our method. The original words (x) and synonyms (x^a) are from the XNLI training sets.

that the EPT setting outperforms NT. One possible explanation could be that the noise in NT affects all English tokens and thus may hurt performance.

4.3 Effect of APT

To investigate the effects of APT, we replace the APT in (5) with the Sentence Representation Pull Target (SRPT). SRPT uses the mean squared error between sentence embeddings of $\boldsymbol{x}$ and $\boldsymbol{x^a}$ as the objective. Formally, $\ell_{SRPT} = \frac{1}{|\mathcal{B}|}\sum_i^{|\mathcal{B}|}(\text{Sent}(\boldsymbol{x}) - \text{Sent}(\boldsymbol{x^a}))^2$, where Sent(x) represents the mean-pooled sentence embeddings (Reimers and Gurevych, 2019) obtained by the middle layer of the model. Results in Table 4 show that: 1) The average performance of SPRT is lower than that of APT. 2) The SRPT mainly improves performance on English-like languages, such as **es**, **de**, and **el**, while drops that of most English-dissimilar languages, such as **tr**, **hi**, **sw**, **ur**, etc. This phenomenon shows that SRPT suffers heavily from English training resources, biasing the VME towards English-like languages, which hurts the overall zero-shot cross-lingual transferability.

We perform T-SNE visualization on the outputs of the mBERT trained with our method. As shown in Figure 5, the synonym is still in the same relative position as the original word, which proves the effectiveness of APT.

4.4 Effect of source language

In addition to **en**, both **de** and **ru** show preference as source languages in cross-lingual learning (Turc et al., 2021). We translate the training set into **de** and **ru** using OPUS-MT (Tiedemann and Thottingal, 2020) models, as shown in Table 5, the performance of our method can be further improved.

scale	size of dictionary	<u>XNLI result</u>
1.0	49975	**68.424**
0.75	37481	**68.392**
0.5	24987	**68.218**
0.25	12493	**68.080**

Table 6: Results on the XNLI task when using the scaled English synonym dictionaries for data augmentation.

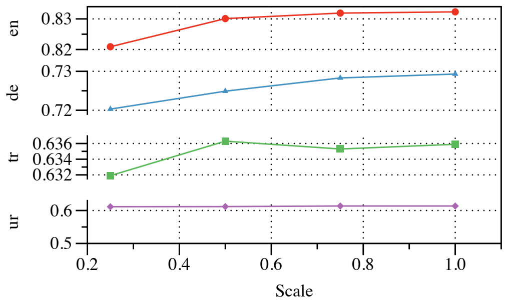


Figure 6: Results on XNLI test sets of four languages when using scaled synonym dictionaries in our method.

4.5 Effect of dictionary size

The data augmentation in our method relies on the size of pre-defined synonym dictionary. As shown in Table 6 and Figure 6, we can observe that: 1) The overall performance decreases as the dictionary size decreases. 2) Some languages are not sensitive to the dictionary size, such as **tr** and **ur**. 3) The performance of **en**, **de**, and **tr** degrades significantly when the dictionary size is scaled from 0.5 to 0.25. This phenomenon may be related to some important synonyms in the dictionary, which are effective for cross-lingual transfer learning.

5 Conclusion

To get rid of the dependence on parallel corpora, enable cross-lingual transfer to low-resource languages, we propose Embedding-Push, Attention-Pull, and Robust targets to combat the influence of language clusters in multilingual models. Experimental results demonstrate that our method outperforms previous works and obtains better-aligned embeddings when trained with only English.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021a. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554.

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021b. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12462–12470.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics (TACL)*, 7:597–610.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320. PMLR.

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 (ACL)*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. *CoRR*, abs/2110.14782.

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 (NAACL)*, pages 4171–4186.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 4411–4421.

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.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8433–8440.

Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target language-aware constrained inference for cross-lingual dependency parsing. 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 1117–1128.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. 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 313–325. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

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*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692.

A Implementation details

Dataset XNLI is the cross-lingual natural language inference task. PAWS-X is used to determine whether two sentences paraphrase each other. The augmentation datasets are obtained from Huang et al. (2021). They augmented 3 and 10 examples for each sentence in XNLI and PAWS-X by synonym replacement, respectively. The pre-defined English synonym set is from Alzantot et al. (2018). The scripts for splitting training, test, and validation sets are provided by XTREME (Hu et al., 2020).

Layer	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
3	83.23	67.17	71.44	73.33	68.06	75.99	74.89	63.27	70.94	51.00	56.61	63.23	61.04	72.30	71.66	68.28
6	83.05	67.01	70.88	72.63	67.98	76.05	74.91	62.99	71.82	51.28	56.81	63.53	61.44	72.48	71.48	**68.29**
9	82.87	67.56	71.22	73.05	68.36	75.81	74.63	63.65	71.14	50.96	56.75	62.97	61.00	72.55	71.68	68.28
12	83.05	67.05	70.56	72.81	68.22	75.55	75.35	63.35	71.48	50.82	56.71	63.11	60.30	72.48	71.68	68.17

Table A.1: Results of the XNLI task when we apply the Attention-Pull target at different layers of mBERT.

β	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
0.1	83.23	67.41	71.04	72.93	68.28	75.75	75.19	63.99	71.64	51.28	56.73	63.59	61.38	72.44	71.50	**68.42**
0.2	83.15	67.05	71.38	73.79	68.34	75.99	75.01	63.53	71.58	50	56.41	63.21	60.52	72.40	71.88	68.32
0.3	83.09	67.47	71.52	72.99	68.44	75.65	75.03	63.57	71.42	50.76	55.87	63.29	61.26	72.63	71.58	68.30
0.5	83.01	67.15	70.58	72.95	68.10	75.87	74.99	62.99	71.64	50.34	56.37	63.61	60.82	72.57	72.18	68.21
0.7	82.69	66.83	71.08	72.87	68.14	75.89	74.43	63.15	71.00	51.60	56.57	63.15	60.88	72.16	71.82	68.15
0.9	82.51	66.83	71.00	72.87	68.50	75.65	75.01	62.99	71.30	50.68	55.77	63.29	61.38	72.42	71.54	68.12

Table A.2: The experimental results of the XNLI task based on mBERT when β takes different values, where α=1.

α	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
0.6	82.85	67.35	71.64	73.03	68.32	75.65	74.57	63.37	71.56	50.56	56.39	63.67	61.16	72.44	72.02	68.30
0.8	82.87	67.23	70.96	73.17	68.68	75.23	74.87	63.53	71.26	50.86	56.45	63.21	61.30	72.55	72.02	68.28
1	83.23	67.41	71.04	72.93	68.28	75.75	75.19	63.99	71.64	51.28	56.73	63.59	61.38	72.44	71.50	**68.42**
1.2	83.19	67.03	71.08	72.97	67.86	75.87	74.75	63.49	71.50	51.60	56.37	63.21	60.88	72.59	71.28	68.24
1.4	83.09	67.03	71.44	73.35	68.78	75.79	74.51	63.23	71.50	51.14	56.35	63.45	60.72	72.75	71.60	68.32
1.6	83.19	67.05	71.50	73.23	68.18	76.25	74.77	63.43	71.06	51.10	56.43	62.95	60.52	72.38	71.98	68.27
1.8	83.29	67.09	71.44	73.51	68.54	75.85	74.79	63.83	71.36	50.78	56.47	63.23	60.86	72.59	71.98	68.37

Table A.3: The experimental results of the XNLI task based on mBERT when α takes different values, where β=0.1.

Setup The mBERT$_{\text{base}}$ and XLM-R$_{\text{base}}$ are obtained from Huggingface's *transformers* package (Wolf et al., 2020). The maximum sequence length is set as 128. The learning rate is set as 2e-5. Our method is trained for one epoch with the batch size of 32. other models are trained following Hu et al. (2020) and Huang et al. (2021).

Input construction Both XNLI and PAWS-X are sentence pair classification tasks. Taking mBERT as an example, for each s_1, s_2 and augmented $\mathrm{s}_1^{\mathrm{a}}$, $\mathrm{s}_2^{\mathrm{a}}$ in the training data, we set $\boldsymbol{x}$ as [CLS]s_1[SEP]s_2[SEP], $\boldsymbol{x}^{\boldsymbol{a}}$ as [CLS]$\mathrm{s}_1^{\mathrm{a}}$[SEP]$\mathrm{s}_2^{\mathrm{a}}$[SEP]. Then, we take $\boldsymbol{x}$ and $\boldsymbol{x}^{\boldsymbol{a}}$ as the input of our method in Figure 2, [CLS] token is used for classification.

B Hyperparameter analysis

There are three main hyperparameters in our method that need to be adjusted. 1) We need to determine which layer is most effective for applying Attention-Pull target. 2) We need to determine the weight of β in the final loss. 3) We need to determine the weight of α in the final loss. We conduct experiments on XNLI task based on mBERT.

For 1), we first set α=1 and β=1, then apply the Attention-Pull target on the {3, 6, 9, 12} layers respectively, and the results are shown in Table A.1. We find that applying the Attention-Pull target to all layers works well. The most significant improvement is achieved at the 6-th layer and the minimal improvement is achieved at the last layer, which may be related to the quality of sentence representation at different layers of the model (Carlsson et al., 2021; Merchant et al., 2020).

For 2), we apply the Attention-Pull target at the 6-th layer and set α=1, then select β from {0.1, 0.2, 0.3, 0.5, 0.7, 0.9}. The experimental results are shown in Table A.2. First, we find that model performance improved when using any of the above β values. Second, we also find that the improvement becomes significant as β decreases, we attribute this phenomenon to the fact that the Attention-Pull target should not over-focus on features of the English corpus but should help the VME capture features in other language clusters. Note that this result does not mean that the Attention-Pull target is unnecessary, as ablation experiments in section 3.4 show that the Attention-Pull target can improve the model. Finally, the best experimental result is obtained when β=0.1.

For 3), we apply the Attention-Pull target at the 6-th layer and set β=0.1, then select α from {0.6,

Model	en	es	de	fr	bg	ru	el	th	sw	vi	ar	zh	hi	ur	tr	avg.
EPT + APT	84.6	79.4	77.5	79.5	78.8	76.8	77.0	73.9	66.7	76.4	74.5	75.8	72.6	68.7	74.7	<u>**75.8**</u>
NT + APT	84.4	79.4	77.2	79.0	78.8	76.7	76.4	73.4	67.3	76.3	73.6	75.2	72.2	67.7	74.1	<u>75.5</u>
EPT + SRPT	84.4	80.0	77.8	79.2	78.5	76.8	77.1	74.2	65.2	76.6	73.5	75.8	72.5	68.7	74.4	<u>75.6</u>

Table A.4: Results on the XNLI task when replacing some targets, based on the XLM-R.

0.8, 1.0, 1.2, 1.4, 1.6, 1.8}. Results are shown as Table A.3. We find that the best performance is achieved when α is 1.0. The performance is also improved when using other α values, which shows that the Embedding-Push target can robustly improve the cross-lingual transferability of models. Therefore, in our main experiments, we set α=1.0, β=0.1 and apply the Attention-Pull target at the 6-th layer.

C Analysis on XLM-R

We perform analysis based on XLM-R, the results are shown in Table A.4.