Effective Fine-tuning Methods for Cross-lingual Adaptation

Tao Yu[¶] and **Shafiq Joty**[¶†]
[¶]Nanyang Technological University, Singapore
[†]Salesforce Research
{tao003, srjoty}@ntu.edu.sg

Abstract

Large scale multilingual pre-trained language models have shown promising results in zero- and few-shot cross-lingual tasks. However, recent studies have shown their lack of generalizability when the languages are structurally dissimilar. In this work, we propose a novel fine-tuning method based on co-training that aims to learn more generalized semantic equivalences as complementary to multilingual language modeling using the unlabeled data in the target language. We also propose an adaption method based on contrastive learning to better capture the semantic relationship in the parallel data, when a few translation pairs are available. To show our method's effectiveness, we conduct extensive experiments on cross-lingual inference and review classification tasks across various languages. We report significant gains compared to directly fine-tuning multilingual pre-trained models and other semi-supervised alternatives.[1]

1 Introduction

Self-supervised pre-trained models (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019) have revolutionized natural language processing (NLP). Such pre-training with language modeling objectives provides a useful initial point for model parameters that adapt well to new tasks with supervised fine-tuning. Building on the success of monolingual pre-trained language models (LM) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), multilingual models like mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020) have pushed the state-of-the-art on cross-lingual tasks by pre-training large Transformer (Vaswani et al., 2017) models jointly on many languages.

The multilingual pre-trained LMs support zero-shot transfer from a source language to target languages, meaning that fine-tuning the pre-trained LM on the source-language labeled data such as English, could transfer well to other languages. Recent research (Wu and Dredze, 2019; K et al., 2020; Pires et al., 2019) has shown that the transfer capability of these multilingual LMs mainly relies on the structural similarity between the source and target languages. When the target language is structurally dissimilar to the source, the transfer ability is shown to be low in the zero-shot setting.

Since the multilingual LMs are generally trained with a self-supervised masked language modeling (MLM) objective without considering parallel information or semantic equivalences, they cannot capture well semantic similarity across languages as reflected by their low Tatoeba score (Phang et al., 2020). This could also potentially harm their zero-shot transfer performance on the tasks as Dufter and Schütze (2020) show that injecting cross-lingual signals by replacing masked tokens with semantically similar words from other languages improves mBERT's multilinguality and zero-shot cross-lingual inference (XNLI) results.

Concurrently, the multilingual embedding models such as LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2020) use parallel data to learn language invariant sentence representation by encoding texts from different languages into a shared embedding space. These models can capture semantic similarity well as often shown by their high Tatoeba scores and their success in tasks that involve cross-lingual similarity such as cross-lingual retrieval and bitext mining. However, it has been shown that these models generally lag behind the multilingual LMs on zero-shot cross-lingual classification tasks like XNLI (Wang, 2019). We hypothesize source information might be necessary to achieve better zero-shot transfer as shown empirically by Phang et al. (2020) with intermediate task fine-tuning in the source language (English).

In this work, we argue that the multilingual em-

[1]Code and models are available at https://github.com/tao-shiwu/co-training-xlu

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8492–8501
November 7–11, 2021.

bedding models and multilingual LMs are complementary to each other – the task adaptability of the multilingual LMs can be complemented by the semantic awareness of the sentence embedding models and vice versa. For this, we first propose a co-training (Blum and Mitchell, 1998) framework that facilitates the multilingual LMs and sentence embedding models to learn from each other by using unlabeled data from target languages.

Secondly, we aim to improve the multilingual LM fine-tuning when there are small amounts of parallel pairs within the training datasets for tasks. Compared to the existing *translate-train* methods (Singh et al., 2020; Conneau et al., 2018) that simply use translation as a data-augmentation method to generate labeled data in the target language, we propose a novel language adaptation approach based on *contrastive learning* that aligns the parallel data to model the semantic relationship between the translation pairs for effective fine-tuning.

We performed extensive experiments on XNLI (Conneau et al., 2018) and Multilingual Amazon Review Corpus (MARC) (Keung et al., 2020) datasets. The experimental results demonstrate that our approach outperforms previous methods for various classification tasks across different languages by a good margin. In particular, on XNLI, our proposed co-training method improves over the original mBERT and XLM-R by 2.3% and 1.7% on average for zero-shot cross-lingual transfer. On MARC, our approach gets on average 8% and 1.1% gains for mBERT and XLM-R, respectively.

2 Related Works

Since the introduction of the transformer network Vaswani et al. (2017), it has become a common model of choice for language representation learning. Pre-trained transformer-based models such as mBERT (Devlin et al., 2019) have proven effective in learning cross-lingual information. mBERT was pre-trained on raw Wikipedia texts in 104 languages using masked language modeling (MLM) and next sentence prediction (NSP) tasks with no explicit cross-lingual objective. XLM-R (Conneau et al., 2020) improves over mBERT by training longer with more data from CommonCrawl, and without the NSP objective.

Meanwhile, several studies examine what makes these pre-trained language models multilingual, and why it works well for cross-lingual transfer. Pires et al. (2019) hypothesize that the cross-lingual capability of mBERT benefits from having a shared (sub-word) vocabulary for all languages, which helps to bind the languages by mapping the token representations into a shared space. K et al. (2020) point out that the contribution from shared subwords is minimal. On the other hand, the structural similarity (*e.g.,* word order, word frequency, etc.) is more important for effective cross-lingual transfer.

Another line of work on multilingual pre-training focuses on generating multilingual sentence embeddings such that semantically similar sentences across different languages will be closer in a shared vector space. LASER (Artetxe and Schwenk, 2019) uses an encoder-decoder architecture (Sutskever et al., 2014). It trains on large parallel data to learn multilingual fixed-length sentence embedding for 93 languages on a translation task. Multilingual Universal Encoder (mUSE) (Chidambaram et al., 2019) uses a dual-encoder architecture that is trained on one billion crawled question-answering pair with a translation ranking task: given a sentence from the source language and a group of candidate text from target languages, the model needs to recognize the corresponding translation of the source-language text from the candidates. LaBSE (Feng et al., 2020) is based on the BERT architecture using the same translation ranking task with mUSE but is trained on a much larger dataset of six billion translation pairs.

Some researchers also tried to introduce cross-lingual alignment from parallel data as an auxiliary objective of the original MLM in the pre-training. Cao et al. (2020) align mBERT embeddings in a post-hoc manner. They first apply a statistical word aligner to align tokens in the parallel sentences. Then, mBERT is tuned via minimizing the mean-squared error between the embedding of the English words and the corresponding words in other languages. Chi et al. (2020) tried to minimize the vector-space distance between a source language sentence and its translation during the pre-training. The problem with these kinds of methods is that they either need to pre-train a new model from scratch (Chi et al., 2020; Lample and Conneau, 2019), or need to do a second round of pre-training on top of the original multilingual LM (Cao et al., 2020). From both the computation and data perspective, the cost is very high.

In another line of work, researchers use data augmentation to solve language adaptation problem in cross-lingual tasks. For example, Bari et al. (2021)

use XLM-R's mask language model to augment the data with vicinal samples. Liu et al. (2021) proposed a labeled sequence translation method to translate source-language NER training data to target languages and train a generation-based multilingual data augmentation method. These methods are orthogonal to our methods.

Summary Current work on multilingual pre-training either does not consider sentence-level cross-lingual alignment in the pre-training (making them sacrifice transfer capability to structurally dissimilar language), or they only consider alignment signals, which makes them expensive to train. In contrast, in our work we focus on utilizing the parallel information in the *fine-tuning* phase with minimum costs to avoid building a pre-trained model from scratch or using large amounts of parallel data. Our approach can also naturally co-operate with different multilingual models based on different pre-training objectives.

3 Methods

In this section, we describe our co-training framework for cross-lingual transfer (§3.1) followed by the contrastive language adaptation (§3.2).

3.1 Co-training based Model Transfer

The goal of our approach is to make use of the cross-lingual semantic information from the multilingual embedding model to improve the zero-shot classification performance on downstream cross-lingual tasks when fine-tuning the multilingual LM with source language data.

Background Co-training (Blum and Mitchell, 1998) is one of the widely used semi-supervised methods, where two complementary classifiers utilize unlabeled data to bootstrap the performance of each other iteratively. Within the co-training framework, each classifier is trained on a unique view of the data. In each iteration, the algorithm selects high confidence data using each of the classifiers to form a pseudo-labeled dataset. The intuition is that if one classifier can confidently predict the class of an example that is very similar to some of the labeled ones, it can provide one more training data for the other classifier. If this data appears easy to be classified by the first classifier, it does not mean that it will be easy for the second classifier. So, the second classifier will get useful information to improve itself and vice versa. Co-training also avoids the *confirmation bias* issue (Tarvainen and Valpola, 2017) with single model self-training, where the model accumulates its own errors.

Algorithm 1: Co-training for cross-lingual task adaptation

Initialize threshold t;
Set $\mathcal{U}^{\text{emb}} = \mathcal{U}, \mathcal{U}^{\text{lm}} = \mathcal{U}$;
Set $\mathcal{D}^{\text{emb}} = \mathcal{D}, \mathcal{D}^{\text{lm}} = \mathcal{D}$;
for s *iterations* **do**
  fine-tune f_θ^{lm} on $\mathcal{D}^{\text{lm}}$;
  fine-tune f_θ^{emb} on $\mathcal{D}^{\text{emb}}$;
  for $x_u \in \mathcal{U}$ **do**
    $(y_u^{\text{lm}}, c_u^{\text{lm}}) = f_\theta^{\text{lm}}(x_u)$;
    $(y_u^{\text{emb}}, c_u^{\text{emb}}) = f^{\text{emb}}(x_u)$;
  end
  $\mathcal{C}_{\text{emb}} = \{x_u, y_u^{\text{emb}} | c_u^{\text{emb}} > t\}_{u=1}^{|\mathcal{U}|}$;
  $\mathcal{C}_{\text{lm}} = \{x_u, y_u^{\text{lm}} | c_u^{\text{lm}} > t\}_{u=1}^{|\mathcal{U}|}$;
  Random choose a subset $\mathcal{S}^{\text{emb}}$ from $\mathcal{C}^{\text{emb}}$;
  Random choose a subset $\mathcal{S}^{\text{lm}}$ from $\mathcal{C}^{\text{lm}}$;
  $\mathcal{U}^{\text{emb}} = \mathcal{U}^{\text{emb}} \setminus \mathcal{S}^{\text{emb}}$;
  $\mathcal{U}^{\text{lm}} = \mathcal{U}^{\text{lm}} \setminus \mathcal{S}^{\text{lm}}$;
  $\mathcal{D}^{\text{emb}} = \mathcal{D}^{\text{emb}} \cup \mathcal{S}^{\text{lm}}$;
  $\mathcal{D}^{\text{lm}} = \mathcal{D}^{\text{lm}} \cup \mathcal{S}^{\text{emb}}$;
end

Proposed Co-Training Framework In our proposed co-training framework (Fig. 1), we have two cross-lingual classifiers, which use two separate pre-trained multilingual models to get the cross lingual representation of a text. The first classifier is based on a multilingual LM like XLM-R (Conneau et al., 2020) that captures the structural similarity across languages by pre-training on MLM. We denote it as f_θ^{lm}. The second one is based on the multilingual sentence embedding model LaBSE (Feng et al., 2020) that is pre-trained on parallel texts to capture the semantic similarity across languages. We denote this model as f_θ^{emb}.

Every input text will get a cross-lingual representation $\boldsymbol{h} \in \mathbb{R}^d$ after being encoded by the pre-trained multilingual models f_θ^{lm} or f_θ^{emb}, where d is the dimension of the sentence representation. Subsequently, each model has a task-specific classification module for task fine-tuning, which consists of a dense layer followed by a Softmax that maps $\boldsymbol{h} \in \mathbb{R}^d$ to $\mathcal{Y}$, where $\mathcal{Y}$ is the set of target classes.

We first fine-tune the two classifiers based on two different pre-trained models for K epochs using the labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where n is the amount of labeled data, x_i is the text from the source language and y_i is the corresponding ground truth label. The next step is to make predictions

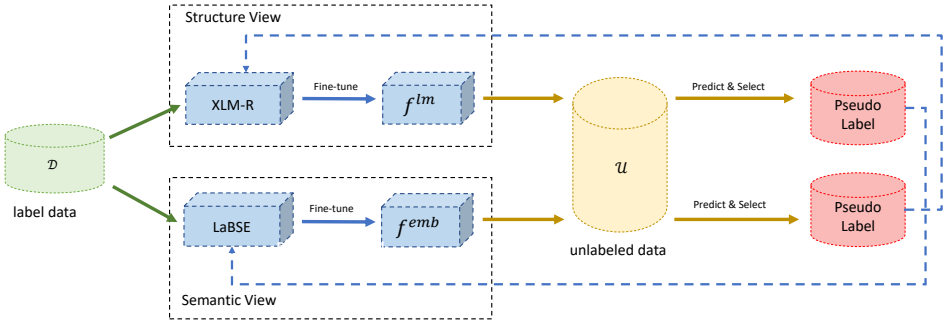

Figure 1: Block diagram of the proposed co-training framework for cross-lingual adaptation.

on the unlabeled data from the target languages: $\mathcal{U} = \{x_j\}_{j=1}^{m}$ with m being the number of unlabeled samples. Consequently, given an arbitrary unlabeled data x_u, the two classifiers f_θ^{lm} and f_θ^{emb} will yield pseudo labels y_u^{lm} and y_u^{emb} as well as confidence scores c_u^{lm} and c_u^{emb}, respectively.

We then set a threshold t and randomly select a subset of the pseudo-labeled data with a confidence score larger than t. The items selected by f_θ^{lm} is merged into the original labeled data $\mathcal{D}$ for f_θ^{emb} to re-train the model with. Similarly, the training dataset for f_θ^{lm} also gets updated with the pseudo labeled data selected by f_θ^{emb}. We perform this process iteratively. See Alg. 1 for a pseudocode.

3.2 Contrastive Language Adaptation

Our co-training method in §3.1 exploits unlabeled target data. In practical scenarios, it is easy to acquire some translations of the source texts either via machine translation (MT) or human translators even for low-resource languages. Let T represent a system (or human) that can translate a source text x_i into arbitrary target languages. With this, we can create a new dataset $\mathcal{D}^{\text{bi}} = \{(x_i^{(t)}, y_i)\}$, where $x_i^{(t)} = T(x_i)$ is the translation of x_i into language t using T. In our case, we translate only a small portion of $\mathcal{D}$ to build $\mathcal{D}^{\text{bi}}$ to make the setup realistic, as getting good translations could be expensive.

One straightforward way to fine-tune f_θ^{lm} using $\mathcal{D}$ and $\mathcal{D}^{\text{bi}}$ is to optimize the following cross-entropy loss, where $p_{y_i}^{\text{lm}}(x_i)$ is the predicted probability from f_θ^{lm} for the ground truth label y_i.

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{|\mathcal{D}\cup\mathcal{D}^{\text{bi}}|} y_i \log(p_{y_i}^{\text{lm}}(x_i)) \qquad (1)$$

However, such method does not fully benefit from the parallel information in $\mathcal{D}^{\text{bi}}$. So, we further develop a language adaptation approach to effectively use the commonality between the source and the target language data in terms of their label space and semantic relationship to deal with the limited size of the parallel training data when fine-tuning the multilingual LM (f_θ^{lm}), as we present below.

Label Alignment (LA) We encourage data with the same class label to be nearby in the embedding space. We utilize the supervised contrastive method (Gunel et al., 2020) to capture the similarities between examples of the same class and contrast them with the examples from the other classes. Specifically, given a batch of training data B from $\mathcal{D} \cup \mathcal{D}^{\text{bi}}$ containing examples of various classes, we optimize f_θ^{lm} using the following loss.

$$\mathcal{L}_{\text{LA}}(\theta) = \sum_{i=1}^{|B|} \frac{1}{N_{y_i} - 1} \sum_{j=1}^{|B|} \mathbb{1}_{i\neq j} \mathbb{1}_{y_i=y_j} \mathcal{L}_i$$
$$\text{where} \quad \mathcal{L}_i = -\log \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_j/\tau)}{\sum_{k=1}^{B} \mathbb{1}_{i\neq k} \exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_k/\tau)} \qquad (2)$$

where $\boldsymbol{h}_k$ and $\boldsymbol{h}_j$ respectively indicate the L_2 normalized `[CLS]` representations (encoded by f_θ^{lm}) of x_k and x_j drawn from the same batch B as x_i, N_{y_i} denotes the amount of data in B that have the same class of y_i, and τ is a temperature parameter.

Combined with the cross-entropy loss (Eq. 1), the final loss function for fine-tuning f_θ^{lm} is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda_{\text{LA}} \mathcal{L}_{\text{LA}}(\theta) \qquad (3)$$

Where λ_{LA} is the hyper-parameter for tuning the importance of the label alignment loss. Note that the label alignment loss can be applied to only $\mathcal{D}$ or $\mathcal{D}^{\text{bi}}$ or both. In our model, we apply it to both (*i.e.,* $\mathcal{D} \cup \mathcal{D}^{\text{bi}}$) to effectively use the commonality between the source and target languages data in their label space.

Semantic Alignment (SA) When fine-tuning with the parallel data, we encourage the source

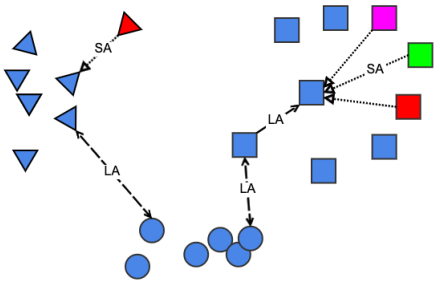

Figure 2: Illustration of the proposed language adaptation method. Different shape indicates different class and different color with same shape indicates different translations of the same source. Label alignment (LA) encourages data with the same label to be nearby in the embedding space, while Semantic Alignment (SA) encourages translation pairs to be nearby in the embedding space.

text and their corresponding translations in different languages to be nearby in the embedding space (they have the same label). Concurrently, for the texts with different labels from different languages, we encourage their embeddings to be far apart.

For each source language labeled instance (x_i, y_i) in $\mathcal{D}^{\text{bi}}$, we construct a batch B of size $|B| = 2b$. The first half of the batch contains parallel texts of x_i from different languages: $\{(x_i, x_i^{(t)})\}_{t=1}^{b}$; they have the same label y_i and are considered as positive pairs. The second half of the batch $\{(x_i, x_j)\}_{j=b+1}^{2b}$ are constructed by the source language texts with a different label (*i.e.*, $y_j \neq y_i$), and considered as negative pairs. The contrastive loss for one data point (x_i, y_i) in $\mathcal{D}^{\text{bi}}$ is:

$$\mathcal{L}_{\text{SA}}^{i} = -\sum_{t=1}^{b} \log \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_i^{(t)}/\tau)}{\sum_{k=1}^{B} \mathbb{1}_{i \neq k} \exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_k/\tau)} \quad (4)$$

Note that the negative pairs $\{(x_i, x_j)\}_{j=b+1}^{2b}$ can be sampled from only $\mathcal{D}$ or only $\mathcal{D}^{\text{bi}}$ or both. In our model we only sample negative pairs from $\mathcal{D}^{\text{bi}}$.

Combined with the cross-entropy loss, the final loss function for fine-tuning f_θ^{lm} on $\mathcal{D}^{\text{bi}}$ is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}} + \lambda_{\text{SA}} \mathcal{L}_{\text{SA}} \quad (5)$$

where λ_{SA} is the hyper-parameter for tuning the importance of the semantic alignment loss.

Overall, when fine-tuning f_θ^{lm} on $\mathcal{D}$ and $\mathcal{D}^{\text{bi}}$, we first train it with Eq. 3 on $\mathcal{D} \cup \mathcal{D}^{\text{bi}}$. Then, we train it on $\mathcal{D}^{\text{bi}}$ using Eq. 5 with a smaller learning rate. See Fig. 2 for an intuitive illustration of the proposed language adaptation method.

4 Experiments

To show the effectiveness of our model, we evaluate our proposed methods on two cross-lingual classification tasks as we describe below.

4.1 Evaluation Tasks & Datasets

Multilingual Amazon Review Corpus (MARC) MARC (Keung et al., 2020)[2] is a large-scale collection of Amazon reviews for multilingual text classification. The corpus contains reviews in English, Japanese, German, French, Spanish, and Chinese. The corpus is balanced across the five possible star ratings, so each rating constitutes 20% of the reviews in each language. We test our model on the binarized classification task from Keung et al. (2020), where we predict whether the reviewer gave a negative review (1-2 stars) or a positive review (4-5 stars). We drop the 3-star reviews in the training and evaluation data. We use only the review body for training and testing. The training data for target languages are used as unlabeled data.

XNLI XNLI (Conneau et al., 2018) is an evaluation benchmark for cross-lingual NLI that covers 15 languages. The dataset is created by translating (by human) the development and test sets of the English MultiNLI dataset (Williams et al., 2018). Given a sentence pair of premise and hypothesis, the task is to classify their relationship as entailment, contradiction, and neutral. On XNLI, we directly use the translation that comes with the dataset[3] as unlabeled dataset for co-training.

4.2 Experimental Setup

On both XNLI and MARC datasets, we experiment with three different setups: (*i*) We train the model using 1.2% percent sampled data from the original XNLI and MARC English training set which we denote as **1.2% zero-shot**. This is to investigate our co-training method's performance when there are only a few labeled data in the source language for zero-shot transfer. (*ii*) we add translations from English to target languages for some (200 samples) of the data in the **1.2% zero-shot** setting to show the effectiveness of our proposed language adaption method; we denote it as **1.2% few-shot**. (*iii*) we also report our co-training method's performance using full English training dataset from XNLI and

[2] https://registry.opendata.aws/amazon-reviews-ml/

[3] https://cims.nyu.edu/~sbowman/xnli/

Setting	$\mathcal{D}$ Size	$\mathcal{U}$ Size	$\mathcal{D}^{\text{bi}}$ Size
MARC **1.2% zero-shot**	2000	120,000	0
MARC **1.2% few-shot**	2000	120,000	200
MARC **100% zero-shot**	160,000	640,000	0
XNLI **1.2% zero-shot**	5000	420,000	0
XNLI **1.2% few-shot**	5000	420,000	200
XNLI **100% zero-shot**	392,702	1,400,000	0

Table 1: Detail numbers of labeled data size and unlabeled data size for different experiment settings.

MARC, which we denote as **100% zero-shot**. The details of the settings are shown in Table 1

Multilingual Pre-trained Model In the experiments, we consider two multilingual language models as f_θ^{lm}: mBERT[4] (Devlin et al., 2019) and XLM-R[5] (Conneau et al., 2020). We use base versions for both mBERT and XLM-R. We use LaBSE[6] (Feng et al., 2020) as the multilingual embedding model f_θ^{emb} in the co-training framework.

Baselines. We compare our co-training method with self-training (Dong and de Melo, 2019). Instead of using multilingual embedding model to generate pseudo label for f_θ^{lm}, self-training uses f_θ^{lm}'s own prediction on the unlabeled data to obtain training data with pseudo labels. We also compare our language adaption method with *translate-train* on mBERT and XLM-R.

Training Details We use the AdamW (Loshchilov and Hutter, 2019) optimizer with 0.00005 initial learning rate, 0.01 weight decay rate, and a linear learning rate scheduler for all our experiments. We use a batch size of 16 and a max sequence length of 128 when fine-tuning on both MARC and XNLI datasets. We fine-tune five epochs on XNLI, and we fine-tune for two epochs on MARC.

Co-training We have a pre-defined threshold t. During each iteration of the co-training, we randomly choose $\frac{1}{3}n$ pseudo labels from qualified candidates (pseudo labels with confidence score larger than t), where n is the raw size of labeled data. If the amount qualified candidates amounts is less than $\frac{1}{3}n$, we then choose all the qualified candidates.

[4] https://github.com/google-research/bert/blob/master/multilingual.md

[5] https://github.com/pytorch/fairseq/tree/master/examples/xlmr

[6] https://tfhub.dev/google/LaBSE/1

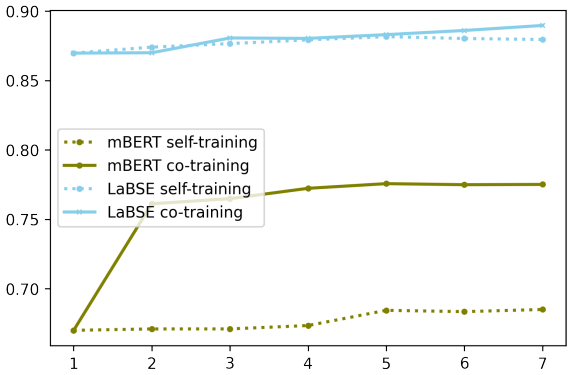

Figure 3: Accuracy after each iteration during self-training and co-training for LaBSE and mBERT.

4.3 Results and Analysis

For evaluation, we report the results on the entire test sets of MARC and XNLI.

4.3.1 MARC

Table 2 shows the results on MARC measured by accuracy. We can see that both self-training and co-training methods can use unlabeled data to improve the model's performance. Compared to self-training, co-training can further enhance the model's performance. We observe that under **1.2% zero-shot** setup, the improvement from self-training is minor. However, our co-training can still learn better knowledge from LaBSE. On average, co-training outperforms self-training by 12% and 2.4% for mBERT and XLM-R, respectively.

To show that the multilingual embedding model and multilingual LM are complementary to each other, we report the accuracy after each iteration during co-training and self-training for mBERT and LaBSE under the **1.2% zero-shot** setup in Fig. 3. Although LaBSE outperforms mBERT by a large margin (18%), it can still learn more useful information from mBERT than from itself in self-training, as it gets about 1% gains in co-training compared to self-training. On the **full (100%)** setup, our co-training method on average gives 8% gain for mBERT and 1.1% for XLM-R compared to their respective fine-tuning versions.

Finally, under the **1.2% few-shot** setup (*i.e.,* with translation data), our language adaptation method improves mBERT by 0.9% on average and is also 0.4% better than XLM-R. We also demonstrate that our language adaptation method and co-training framework can be combined to further improve mBERT's performance by 5.1% and improve XLM-R by 1.7% on average, respectively.

Model	*de*	*en*	*fr*	*ja*	*zh*	**avg**
Zero-shot Cross-lingual Transfer (1.2%)						
mBERT	0.712	0.816	0.540	0.738	0.578	0.670
+ self-training	0.756	0.846	0.566	0.749	0.506	0.685
+ co-training	0.827	0.849	0.750	0.847	0.752	0.805
XLM-R	0.886	0.897	0.809	0.874	0.880	0.869
+ self-training	0.866	**0.899**	0.809	0.871	0.867	0.863
+ co-training	**0.901**	0.899	**0.839**	**0.906**	**0.888**	**0.887**
Zero-shot Cross-lingual Transfer (100%)						
mBERT	0.861	0.922	0.741	0.876	0.697	0.819
+ self-training	0.832	0.927	0.762	0.858	0.773	0.830
+ co-training	0.922	0.929	0.854	0.921	0.870	0.899
XLM-R	0.933	0.938	0.861	0.929	0.912	0.915
+ self-training	0.941	**0.945**	0.846	0.936	0.918	0.917
+ co-training	**0.941**	0.943	**0.882**	**0.941**	**0.925**	**0.926**
Few-shot Cross-lingual Transfer (1.2%)						
mBERT	0.784	0.832	0.677	0.798	0.695	0.757
+ lang. adaptation	0.783	0.840	0.683	0.809	0.714	0.766
+ lang. adaptation & co-training	0.828	0.850	0.739	0.851	0.742	0.802
XLM-R	0.899	0.898	0.831	0.895	0.887	0.882
+ lang. adaptation	0.905	0.909	0.832	0.902	0.881	0.886
+ lang. adaptation & co-training	**0.918**	**0.915**	**0.849**	**0.913**	**0.897**	**0.899**

Table 2: MARC accuracy score for *English (en)*, *French (fr)*, *German (de)*, *Chinese (zh)*, *Japaneses (ja)*. We report the mean accuracy running on three different random seeds

Model	*en*	*ur*	*vi*	*es*	*hi*	*fr*	*sw*	*bg*	*de*	*ar*	*el*	*ru*	*th*	*zh*	*tr*	**avg**
Zero-shot Cross-lingual Transfer (1.2%)																
mBERT	0.653	0.455	0.501	0.545	0.463	0.541	0.424	0.507	0.525	0.482	0.495	0.509	0.342	0. 481	0.467	0.493
+ self-training	0.687	0.475	0.519	0.574	0.492	0.600	0.469	0.548	0.572	0.511	0.517	0.553	0.343	0.504	0.481	0.523
+ co-training	0.694	0.514	0.562	0.624	0.521	0.625	0.468	0.576	0.602	0.555	0.576	0.591	0.346	0.562	0.542	0.557
XLM-R	0.658	0.502	0.575	0.613	0.531	0.610	0.522	0.580	0.589	0.529	0.571	0.557	0.527	0.523	0.542	0.568
+ self-training	0.704	0.554	0.641	0.665	0.572	0.660	0.565	0.636	0.649	0.578	0.629	0.625	0.569	0.592	0.609	0.617
+ co-training	**0.713**	**0.568**	**0.652**	**0.672**	**0.582**	**0.677**	**0.581**	**0.656**	**0.653**	**0.614**	**0.640**	**0.646**	**0.615**	**0.615**	**0.630**	**0.634**
Zero-shot Cross-lingual Transfer (100%)																
mBERT	0.807	0.587	0.718	0.675	0.667	0.611	0.563	0.627	0.640	0.526	0.691	0.659	0.355	0.619	0.734	0.632
+ self-training	0.814	0.602	0.732	0.691	0.697	0.630	0.585	0.643	0.667	0.531	0.705	0.678	0.349	0.639	0.744	0.647
+ co-training	0.814	0.604	0.731	0.698	0.708	0.659	0.578	0.650	0.699	0.544	0.700	0.677	0.352	0.640	0.763	0.655
XLM-R	0.836	**0.651**	0.745	0.778	0.689	0.772	0.653	0.767	0.775	0.741	0.770	0.752	0.707	0.727	0.716	0.734
+ self-training	0.838	0.646	0.753	0.780	0.685	0.771	0.666	0.769	0.759	0.702	0.751	0.745	0.713	0.729	0.725	0.736
+ co-training	**0.843**	0.640	**0.751**	**0.798**	**0.701**	**0.786**	**0.683**	**0.787**	**0.775**	**0.741**	**0.770**	**0.768**	**0.724**	**0.753**	**0.730**	**0.751**
Few-shot Cross-lingual Transfer (1.2%)																
mBERT	0.623	0.504	0.551	0.535	0.540	0.568	0.495	0.511	0.525	0.427	0.552	0.532	0.338	0.491	0.570	0.518
+ lang. adaptation	0.646	0.511	0.564	0.544	0.552	0.569	0.509	0.517	0.533	0.451	0.561	0.543	0.342	0.514	0.577	0.527
+ lang. adaptation & co-training	0.668	0.525	0.598	0.625	0.546	0.619	0.527	0.600	0.604	0.556	0.586	0.587	0.547	0.570	0.578	0.582
XLM-R	0.643	0.527	0.589	0.606	0.554	0.607	0.523	0.601	0.602	0.578	0.580	0.595	0.523	0.562	0.577	0.578
+ lang. adaptation	0.521	0.580	0.619	0.559	0.614	0.537	0.597	0.597	0.597	0.574	0.589	0.593	**0.559**	0.586	0.574	0.582
+ lang. adaptation & co-training	**0.726**	**0.630**	**0.646**	**0.649**	**0.645**	**0.650**	**0.604**	**0.655**	**0.654**	**0.634**	**0.649**	**0.626**	0.537	**0.642**	**0.664**	**0.641**

Table 3: XNLI accuracy for *English (en)*, *French (fr)*, *Spanish (es)*, *German (de)*, *Greek (el)*, *Bulgarian (bg)*, *Russian (ru)*, *Turkish (tr)*, *Arabic (ar)*, *Vietnamese (vi)*, *Thai (th)*, *Chinese (zh)*, *Hindi (hi)*, *Swahili (sw)* and *Urdu (ur)*. We report the mean accuracy running on three different random seed.

4.3.2 XNLI

Table 3 shows the results in accuracy of the experiments on XNLI. Overall we observe that our model outperforms the baselines on almost all 15 test languages in the three experimental setups.

Under the **1.2% zero-shot** setup, our co-training method gives a sizeable improvement of 6.4% and 6.6% for mBERT and XLM-R, respectively, compared to their fine-tuning versions. Our co-training method also gives an average increase of 3.4% and

Model	*de*	*en*	*fr*	*ja*	*zh*	**avg**
Few-shot Cross-lingual Transfer (1.2%)						
mBERT + lang. adaptation	0.795	0.838	0.687	0.806	0.712	**0.767**
- *SA*	0.790	0.838	0.676	0.811	0.697	0.762
- *LA* ($\mathcal{D}^{bi}$)	0.783	0.818	0.686	0.789	0.691	0.753
- *LA* ($\mathcal{D}$)	0.784	0.827	0.677	0.797	0.719	0.759
- *LA* ($\mathcal{D} \cup \mathcal{D}^{bi}$)	0.787	0.818	0.678	0.789	0.672	0.752
XLM-R + lang. adaptation	0.905	0.909	0.842	0.899	0.882	**0.887**
- *SA*	0.906	0.904	0.841	0.899	0.880	0.886
- *LA* ($\mathcal{D}^{bi}$)	0.901	0.905	0.826	0.899	0.884	0.883
- *LA* ($\mathcal{D}$)	0.898	0.903	0.828	0.898	0.878	0.881
- *LA* ($\mathcal{D} \cup \mathcal{D}^{bi}$)	0.898	0.900	0.826	0.8925	0.880	0.879

Table 4: Ablation study on the MARC dataset. **- *SA*** refers to the the model variant without the semantic alignment objective based on the translation data $\mathcal{D}^{bi}$. **- *LA***($\mathcal{D}^{bi}$) refers to the variant that uses SA and LA($\mathcal{D}$) (recall the full model uses $LA(\mathcal{D} \cup \mathcal{D}^{bi})$). **- *LA***($\mathcal{D}$) refers to the variant that uses SA and $LA(\mathcal{D}^{bi})$. **- *LA***($\mathcal{D} \cup \mathcal{D}^{bi}$) refers to the variant that uses only SA.

1.7% on self-training. Specifically, we observe over 2% gain for *ur*, *bg*, *ar*, *th*, *zh* and *tr* when we compare our co-training with self-training.

At the **full (100%)** setup, our co-training method yields an average gain of 2.3% for mBERT and 1.7% for XLM-R compared to their respective fine-tuning versions. This shows that even with a large amount of labeled data from the source language, the model can still benefit from multilingual embedding model through our co-training method.

We also show the importance of target language data in our experiment by adding a small number (200 in our case) of translation pairs in the **Few-shot Cross-lingual Transfer (1.2%)** setting. The traditional *translate-train* method can give mBERT and XLM-R 2.5% and 1% average gains over all languages, respectively. We further improve this gap to 3.4% and 1.4% by adopting language adaptation. Similar to the experimental results on MARC, combining language adaption and co-training method, we achieve the best performance in this setup.

4.4 Ablation

To better understand the contribution from different optimization objectives, we perform an ablation study on the MARC dataset by ablating one component at a time from the complete model.

From the results in Table 4, we observe that generally removing one of the objectives would reduce the performance on average, indicating that all objectives contribute to the overall performance.

When we remove the semantic alignment loss (**- *SA***), we observe an accuracy drop of 0.5% in mBERT compared to the full system. The accuracy drop for XLM-R is 0.1%.

Removing label alignment loss on the source language data (-*LA*($\mathcal{D}$)) leads to ∼0.7% accuracy drop across the board. Removing label alignment on translation data (-*LA*($\mathcal{D}^{bi}$) leads to 1.4% accuracy drop on mBERT and 0.4% accuracy drop on XLM-R. This observation shows positive effects of the label alignment loss on both source language data $\mathcal{D}$ and target language translation data $\mathcal{D}^{bi}$.

5 Conclusion & Future Work

In this paper, we have proposed an effective fine-tuning method to improve cross-lingual transfer capability of multilingual pre-trained LMs. In contrast to previous work, our proposed co-training framework can make multilingual pre-trained LMs learn cross-lingual semantic relationships from the multilingual embedding model. Moreover, we propose a novel language adaptation approach based on contrastive learning. When there exist translation pairs within the training dataset, our language adaption approach can better model the semantic relationship across languages on translation pairs for effective fine-tuning. Extensive experiments have been conducted on the XNLI and Amazon multilingual review dataset, which show that our method outperforms previous methods on both zero-shot transfer and few-shot transfer.

For future studies, we will investigate the data selection policies for the co-training methods. In some cases, the distribution of the labeled data could be different from that of the unlabeled data. It may yield a sampling bias in the training iterations of co-training that shifts towards the unlabeled set, thus hurting the model performance. A more robust

data selection policy could solve this problem.

References

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*, 7:597–610.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A Robust Unsupervised Data Augmentation Framework for Cross-Lingual NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21, pages xx—-xx, Bangkok, Thailand. ACL.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. 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. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. 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*, volume 32. Curran Associates, Inc.

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, Brussels, Belgium. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. 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 6306–6310, Hong Kong, China. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics, ACL'21, pages xx—-xx, Bangkok, Thailand. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2020. {XLDA}: Cross-lingual data augmentation for natural language inference and question answering.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204.

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.

Pei Wang. 2019. Transformer-based multi-lingual sentence embeddings. Master's thesis, École Polytechnique Fédérale de Lausanne.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.