

CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP

Libo Qin¹, Minheng Ni¹, Yue Zhang^{2,3}, Wanxiang Che¹

¹Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China

²School of Engineering, Westlake University, China

³Institute of Advanced Technology, Westlake Institute for Advanced Study
{lbqin, mhni, car}@ir.hit.edu.cn, yue.zhang@wias.org.cn

Abstract

Multi-lingual contextualized embeddings, such as multilingual-BERT (mBERT), have shown success in a variety of zero-shot cross-lingual tasks. However, these models are limited by having inconsistent contextualized representations of subwords across different languages. Existing work addresses this issue by bilingual projection and fine-tuning technique. We propose a data augmentation framework to generate multi-lingual code-switching data to fine-tune mBERT, which encourages model to align representations from source and multiple target languages once by mixing their context information. Compared with the existing work, our method does not rely on bilingual sentences for training, and requires only one training process for multiple target languages. Experimental results on five tasks with 19 languages show that our method leads to significantly improved performances for all the tasks compared with mBERT.

1 Introduction

Neural network models for NLP rely on the availability of labeled data for effective training [Yin *et al.*, 2019]. For languages such as English and Chinese, there exist manually labeled datasets on a variety of tasks, trained over which neural models for NLP can rival human performance. However, for most of the languages, manually labeled data can be scarce. As a result, cross-lingual transfer learning stands as a useful research task [Ruder *et al.*, 2017]. The main idea is to make use of knowledge learned from a resource-rich language to enhance model performance on a low-resource language. In particular, zero-shot cross-lingual learning has attracted much research attention [Wang *et al.*, 2019], which requires no labeled data for a target language. In this paper, we consider this transfer setting.

Recent state-of-the-art results have been achieved by methods based on cross-lingual contextualized embeddings [Conneau and Lample, 2019; Huang *et al.*, 2019; Liu *et al.*, 2019b; Devlin *et al.*, 2019]. In particular, a common set of subwords are extracted across different languages, which are taken as the basis for training contextualized embeddings. For such

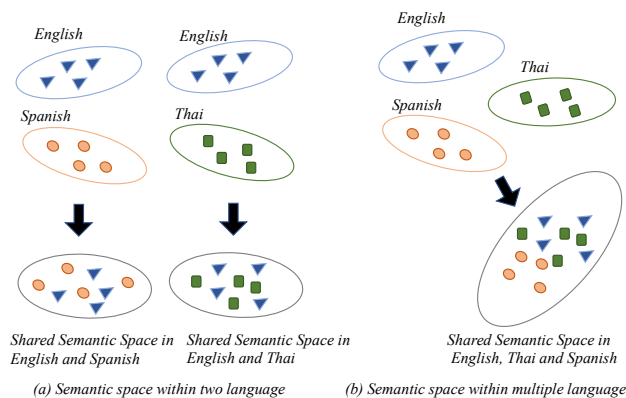


Figure 1: Prior work (a) vs. our method (b).

training, raw sentences from multiple languages are merged into a single training set, so that shared subword embeddings and other parameters are tuned across different languages. A representative model is mBERT [Devlin *et al.*, 2019], which is a multi-lingually trained version of BERT.

While the method above gives strong results for zero-shot cross-lingual adaptation through shared subwords and parameters, it has a salient limitation. The context for training cross-lingual embeddings is still mono-lingual, which can lead to inconsistent contextualized representations of subwords across different languages. To address this issue, several recent methods try to bridge the inconsistency of contextualized embeddings across languages. As shown in Figure 1(a), two main lines of methods are considered. One learns a mapping function from a source contextualized subword embedding to its target counterpart by using word alignment information [Wang *et al.*, 2019], and the other uses code mixing to construct training sentences that consist of both source and target phrases in order to fine-tune mBERT [Liu *et al.*, 2019b]. Unfortunately, both lines of work only consider a pair of source and target languages at a time, therefore resulting in a separate model for each target language.

We consider enhancing mBERT without creating multiple additional models, by constructing code-switched data in multiple languages dynamically for better fine-tuning. To this end, a set of English raw sentences and the bilingual dictionaries of MUSE [Lample *et al.*, 2018] are used as the ba-

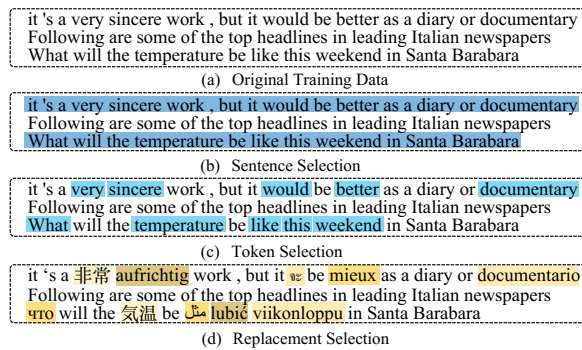


Figure 2: Augmentation process. The source language sentences (a), the sentence selection step (b), the token selection step (c) and the replacement selection step (d) (different shades yellow colors in (d) represent different languages translation).

sis. As shown in Figure 2, three data augmentation steps are taken. First, a set of sentences is randomly selected for code mixing. Second, a set of words is randomly chosen in each sentence for being replaced with the translation words in a different language. Third, for each word to translate, a target language is randomly selected. The above procedure is dynamically executed on a batch level, for fine-tuning mBERT. The intuition is to help the model automatically and implicitly align the replaced word vectors in the source and all target languages by mixing their context information.

Compared with existing methods, our method has the following advantages. First, the resulting model is as simple to use as mBERT, without the need to know the test language before hand. In addition, one training process is used for all different target languages. Second, unlike most existing methods, our method does not rely on parallel sentences, which is especially practical for low-resource languages. Third, the method is dynamic in the sense that a different set of code-switched sentences is constructed in each batch during training, therefore increasing the distribution of data instances [Liu *et al.*, 2019a]. Finally, contextualized embeddings for all the languages are aligned into the same space while prior work can only align representation in source and one target language for each language. This advantage is demonstrated in Figure 1 (b).

We conduct experiments on five zero-shot cross-lingual tasks: natural language inference, sentiment classification, document classification, natural language understanding and dialogue state tracking. Results show that our method leads to significantly improved performances for all the tasks compared with mBERT. In addition, our method also outperforms the existing enhancement methods over multi-lingual contextualized representation methods mentioned earlier. Finally, we find that our method is particularly helpful for small datasets. For some tasks, our model gives the best results with only 1/10 English training data. All codes are publicly available at: <https://github.com/kodenii/CoSDA-ML>.

2 Background

In this section, we will describe the background of mBERT as well as how to apply mBERT for cross-lingual classification

tasks and sequence labeling tasks.

2.1 mBERT

mBERT follows the same model architecture and training procedure as BERT [Devlin *et al.*, 2019]. It adopts a 12 layer Transformer, but instead of training only on monolingual English data, it is trained on the Wikipedia pages of 104 languages with a shared word piece vocabulary, which allows the model to share embeddings across languages.

2.2 Fine-tuning mBERT for Classification

Given an input utterance $\mathbf{s} = (s_1, s_2, \dots, s_n)$ from a source language (i.e., English), we first construct the input sequence by adding specific tokens $\mathbf{s} = ([CLS], s_1, s_2, \dots, s_n, [SEP])$, where $[CLS]$ is the special symbol for representing the whole sequence, and $[SEP]$ is the special symbol to separate non-consecutive token sequences [Devlin *et al.*, 2019]. mBERT takes the constructed input sequence of no more than 512 tokens and outputs the representation of the sequence $h = (h_{CLS}, h_1, \dots, h_n, h_{SEP})$.

For classification tasks, mBERT takes h_{CLS} into a classification layer to find the label c :

$$c = \text{softmax}(\mathbf{W}h_{CLS} + \mathbf{b}), \quad (1)$$

where \mathbf{W} is a task-specific parameter matrix. We fine-tune all the parameters of mBERT as well as \mathbf{W} jointly by maximizing the log-probability of the correct label.

2.3 Fine-tuning mBERT for Sequence Labeling

For sequence labeling tasks, we feed the final hidden states of the input tokens into a softmax layer to classify the tokens. Note that BERT produces embeddings in the wordpiece-level with WordPiece tokenization. We use the hidden state corresponding to the first sub-token as input to classify a word.

$$\mathbf{y}_n = \text{softmax}(\mathbf{W}^s \mathbf{h}_n + \mathbf{b}^s), \quad (2)$$

where \mathbf{h}_n is the first sub-token representation of word x_n .

2.4 Zero-Shot Cross-Lingual Adaption

The baseline mBERT models, trained on the source language classification and sequence labeling tasks, perform *zero-shot* cross-lingual transfer tasks by directly being used for the target language. We assume that there are labeled training data for each task in English, and transfer the trained model to a target language without labeled training data.

3 Method

Our method enhances mBERT in §2.1. In this section, we first describe the overall training process (§3.1). Then, we explain our augmentation algorithm in detail (§3.2). Finally, we introduce the conducted tasks and their input construction for mBERT (§3.3).

3.1 Training and Adaptation

Our framework performs the cross-lingual tasks in two steps: Fine-tuning mBERT with augmented multi-lingual code-switch data and applying it for zero-shot testing, which is illustrated in Figure 3. Given a batch of training data $\mathbf{S} =$

Algorithm 1: Multi-lingual code-switching data augmentation framework.

Input: source language training data: $\mathbf{S} = \{\mathbf{s}^{(n)}\}_{n=1}^N$;
 a set of bilingual dictionaries: $dict$; {sentence, token} replacement ratio: $\{\alpha, \beta\}$; target language sets: LAN

Output: multi-lingual code-switching training data:

$$\mathbf{T} = \{\mathbf{t}^{(n)}\}_{n=1}^N.$$

```

1 for  $n \leftarrow 1 \dots N$  do
2   if  $random() < \alpha$  then
3      $i \leftarrow 0$ ;
4      $\mathbf{t}^{(n)} \leftarrow \emptyset$ ;
5     while  $s_i^n \neq [SEP]$  do
6       if  $random() < \beta$  then
7          $tgt \leftarrow random(LAN)$ ;
8          $s_i^n \leftarrow dict_{src}^{tgt}[s_i^n]$ ;
9       else
10         $s_i^n \leftarrow s_i^n$ ;
11      end
12       $\mathbf{t}^{(n)} \leftarrow \mathbf{t}^{(n)} \cup \{s_i^n\}$ ;
13       $i \leftarrow i + 1$ ;
14    end
15  else
16     $\mathbf{t}^{(n)} \leftarrow \mathbf{s}^{(n)}$ ;
17  end
18 end
    
```

$\{\mathbf{s}^{(n)}\}_{n=1}^N$ from a source language, the dynamic augmentation generator adopts Algorithm 1 to generate code-switched training data for fine-tuning mBERT. Formally, the procedure can be written as:

$$\mathbf{T} = \text{Generator}(\mathbf{S}), \quad (3)$$

$$out = \text{Fine-tune}(\text{mBERT}, \mathbf{T}), \quad (4)$$

where \mathbf{T} represents the generated code-switched data, out denotes the output of all tasks. In zero-shot test, the fine-tuned mBERT is used directly for target languages.

3.2 Data Augmentation Algorithm

The augmentation method consists of three steps, including sentence selection, word selection and replacement selection.

- (i) **Sentence Selection:** Given a batch of training data \mathbf{S} , we randomly select sentences for generating code-switched sentences. The unselected sentences keep in the original language. Take the sentences in Figure 2(b) for example, we randomly select the first and the third sentence while leaving the second sentence unchanged;
- (ii) **Token Selection:** For each selected sentences in the sentence selection step, we randomly choose words to translate. Take the example in Figure 2(c). The word “very” in first sentence and “What” in third sentence are chosen;
- (iii) **Replacement Selection:** After obtaining the selected word, we randomly choose a target language according to a bilingual-dictionary. As shown in Figure 2 (d), different target languages can be mixed in the code-switch

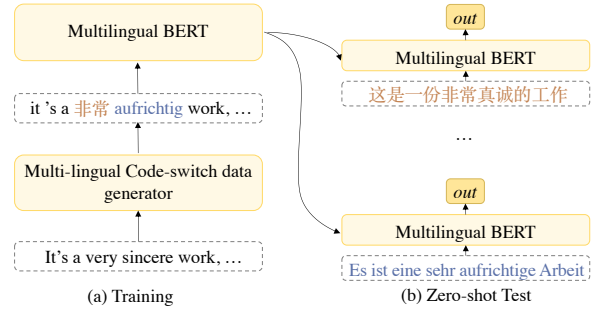


Figure 3: Illustration of our training and zero-shot test process with multi languages. Red color denotes Chinese and blue color denotes German. Better viewed in color.

generated data. It is worth noticing that words in the source language can have multiple translations in the target language. In this case, we randomly choose any of the multiple translations as the replacement target language word. Though we cannot guarantee that this is the correct word-to-word translation in the context, we can consider it as one of the data augmented strategy for our tasks.

Algorithm 1 shows pseudocode for the multi-lingual code-switching code augmentation process, where lines 1-2 denote the sentence selection step, lines 3-6 denote the word selection and lines 7-11 denote the replacement selection step.

In addition, the augmentation steps are performed per batch dynamically and the model trains with different augmented data in each batch, which can increase the distribution of data instances [Liu *et al.*, 2019a]. Intuitively, training with augmented code-switched data can make model automatically align the replaced word in the target language and the original word in a source language into a similar vector space according to their similar context information.

3.3 Tasks

Natural Language Inference. We use XNLI [Conneau *et al.*, 2018], which covers 15 languages for natural language inference. We feed a pair of sentences directly into the mBERT encoder and a task-specific classification layer is used for classification. Models are evaluated by the classification accuracy (ACC).

Sentiment Classification. Following Barnes *et al.* [2018], we use the OpENER English and Spanish datasets, and the MultiBooked Catalan and Basque datasets. We directly provide the sentence to mBERT encoder and the specific [CLS] representation is fed into a linear layer for classification. Models are evaluated by the macro F1.

Document Classification. We use MLDoc [Schwenk and Li, 2018] for document classification, which includes a balanced subset size of the Reuters corpus covering 8 languages for document classification. Similar to sentiment classification, we also directly provide the document to mBERT encoder, and the specific [CLS] representation is fed into a linear layer for classification. Models are evaluated by classification accuracy (ACC).

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Average
Artetxe and Schwenk [2018]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM [Conneau and Lample, 2019]	84.1	77.1	78.0	75.0	74.1	75.1	72.4	70.0	70.6	71.5	68.3	73.2	66.7	67.5	62.2	72.4
+CoSDA-ML	84.4*	79.0*	79.2*	77.9*	76.8*	77.6*	75.7*	72.6*	73.4*	75.3*	72.6*	75.1*	71.2*	70.0*	68.3*	75.3*
mBERT from Wu and Dredze [2019]	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
+CoSDA-ML	82.9	76.7	76.9	74.1	70.9	72.7	73.2	63.9	68.0	73.6	59.8	73.8	65.5	51.0	62.3	69.7

Table 1: Natural Language Inference experiments.

Model	es		eu		ca	
	2-Class	4-Class	2-Class	4-Class	2-Class	4-Class
BLSE [Barnes <i>et al.</i> , 2018]	74.6	41.2	69.3	30.0	72.9	35.9
XLM [Conneau and Lample, 2019]	86.1	32.9	-	-	-	-
+CLCSA	91.3	46.8	-	-	-	-
mBERT [Devlin <i>et al.</i> , 2019]	93.1	51.0	73.1	35.1	83.5	52.3
+CLCSA	95.2*	57.9*	85.7*	55.6*	88.7*	64.3*

Table 2: Sentiment classification experiments.

Dialogue State Tracking (DST). Following prior work [Liu *et al.*, 2019b], we use the Multilingual WOZ 2.0 dataset [Mrkšić *et al.*, 2017], which includes German and Italian languages. DST aims to predict the slot-value pair given a current utterance and the previous system acts. It can be viewed as a collection of binary prediction problems by using a distinct estimator for each slot-value pair [Chen *et al.*, 2018]. We concatenate the current utterance and the previous system act for input into mBERT and obtain the [CLS] representation. We also feed each slot-value pair into mBERT and obtain another [CLS] representation. Finally, the two representations are provided to the classification layer to decide whether it should be selected. Similar to prior work, we use the turn-level request tracking accuracy, joint goal tracking accuracy and the slot tracking accuracy for evaluation.

Spoken Language Understanding. We follow Schuster *et al.* [2019b] and use the cross-lingual spoken language understanding dataset, which contains English, Spanish and Thai. We adopt a joint model which provides the utterance for mBERT and the [CLS] is used for intent detection. The token representations are used for slot prediction as local classifier task on each word, which can be treated as a sequence labeling task. Intent detection is evaluated by the classification accuracy (ACC) and slot filling is evaluated by F1 score.

4 Experiments

We evaluate the effectiveness of our proposed method across 19 languages on five tasks. In addition to mBERT, we also conduct all experiments on the recent strong pre-trained cross-lingual model (XLM) [Conneau and Lample, 2019]. XLM outperforms mBERT on XNLI tasks, but underperforms mBERT for some other tasks [Liu *et al.*, 2019b]. We choose it as a secondary baseline for verifying the generalizability of our augmentation method.

4.1 Experimental Settings

For all tasks, no preprocessing is performed except tokenization of words into subwords with WordPiece. Following Devlin *et al.* [2019], we use WordPiece embeddings with a 110k token vocabulary. We use the base case multilingual BERT

(mBERT), which has $N = 12$ attention heads and $M = 12$ Transformer blocks. In fine-tuning, we select the best hyperparameters by searching a combination of batch size, learning rate, the number of fine-tuning epochs and replacement ratio with the following range: learning rate $\{1 \times 10^{-6}, 2 \times 10^{-6}, 3 \times 10^{-6}, 4 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$; batch size $\{8, 16, 32\}$; number of epochs: $\{4, 10, 20, 40, 100\}$; token and sentence replacement ratio: $\{0.4, 0.5, 0.6, 0.8, 0.9, 1.0\}$. Note that the best model are saved by development performance in the *English*.

4.2 Baselines

We include the following state-of-the-art baselines:

Natural Language Inference. Artetxe and Schwenk [2018] use multilingual sentence representation, pre-trained with sequence-to-sequence model. This model requires bitext for training.

Sentiment Classification. BLSE [Barnes *et al.*, 2018] jointly represents sentiment information in a source and target language and achieves the state-of-the-art performance in zero-shot cross-lingual sentiment classification.

Document Classification. 1) Schwenk and Li [2018] use MultiCCA, multilingual word embeddings trained with a bilingual dictionary, and convolution neural networks. 2) Artetxe and Schwenk [2018] also obtain the promising performance and the detail has been described in Natural Language Inference paragraph.

Dialogue State Tracking (DST). 1) XL-NBT [Chen *et al.*, 2018] proposes a state tracker for the source language as a teacher and then distills and transfers its own knowledge to the student state tracker in target languages. 2) Attention-Informed Mixed Training: Liu *et al.* [2019b] use the generated attention-informed code-switch data for training and achieves the state-of-the-art performance.

Spoken Language Understanding. 1) Multi. CoVe: [Schuster *et al.*, 2019a] use Multilingual CoVe [Yu *et al.*, 2018] as the encoder and add an autoencoder objective to produce more general representations for semantically similar sentences across languages. 2) Attention-Informed Mixed

Model	en	de	zh	es	fr	it	ja	ru	Average
Schwenk and Li [2018]	92.2	81.2	74.7	72.5	72.4	69.4	67.6	60.8	73.9
Artetxe and Schwenk [2018]	89.9	84.8	71.9	77.3	78.0	69.4	60.3	67.8	74.9
XLM [Conneau and Lample, 2019]	94.2	76.8	46.2	64.0	70.5	-	-	61.5	68.9
+CLCSA	93.4	81.4	71.1	73.1	83.7	-	-	68.3	78.5
mBERT [Devlin <i>et al.</i> , 2019]	94.2	80.2	76.9	72.6	72.6	68.9	56.5	73.7	74.5
+CLCSA	95.2*	86.3*	85.5*	79.2*	86.7*	72.6*	73.7*	75.1*	81.8*

Table 3: Document classification experiments.

Model	German			Italian		
	slot acc.	joint goal acc.	request acc.	slot acc.	joint goal acc.	request acc.
XL-NBT [Chen <i>et al.</i> , 2018]	55.0	30.8	68.4	72.0	41.2	81.2
Attention-Informed Mixed Training [Liu <i>et al.</i> , 2019b]	69.5	32.2	86.3	69.5	31.4	85.2
XLM from Liu <i>et al.</i> [2019b]	58.0	16.3	75.7	-	-	-
+CLCSA	77.4	48.7	88.3	-	-	-
mBERT [Devlin <i>et al.</i> , 2019]	57.6	15.0	75.3	54.6	12.6	77.3
+CLCSA	83.0*	63.2*	94.0*	82.2*	61.3*	94.2*

Table 4: Dialog State Tracking experiments.

Model	Spanish		Thai	
	Intent acc.	Slot F1	Intent acc.	Slot F1
Multi. CoVe [Yu <i>et al.</i> , 2018]	53.9	19.3	70.7	35.6
Attention-Informed Mixed Training [Liu <i>et al.</i> , 2019b]	86.5	74.4	70.6	28.5
XLM from Liu <i>et al.</i> [2019b]	62.3	42.3	31.6	7.9
+ CLCSA	90.3	69.0	86.7	34.9
mBERT [Devlin <i>et al.</i> , 2019]	73.7	51.7	28.2	10.6
+ CLCSA (Static)	92.8	75.2	74.8	28.1
+ CLCSA	94.8*	80.4*	76.8	37.3*

Table 5: Slot filling and Intent detection experiments.

Training: Liu *et al.* [2019b] use attention to generate code-switched sentences, achieving the current best result. The method translates only one word into each augmented sentence.

4.3 Results

We perform t-test for all experiments to measure whether the results from the proposed model are significantly better than the baselines. The numbers with asterisks indicate that the improvement is significant with $p < 0.01$. “-” represents the absence of languages in the XLM models and we cannot report the results. Five tasks results are shown in Table 1, 2, 3, 5 and 4, respectively. Across the tasks, we can observe that: 1) mBERT achieves strong performance on all zero-shot cross-lingual tasks, which demonstrates that mBERT is a surprisingly effective cross-lingual model for a wide range of NLP tasks. This is consistent with the observation of Wu and Dredze [2019]. Additionally, the XLM achieves much better performance than mBERT on XNLI and achieves the promising performance on four other tasks. 2) Our method outperforms mBERT and XLM by a large margin and achieves state-of-the-art performance on all the tasks, which demonstrates the effectiveness of our proposed method. Note that we have not reproduced the results on XNLI task of original paper because of lacking the exact best hyper-parameters, which is also mentioned on some issues on Github.¹ So we run their open-source code² to obtain the results and we ap-

¹<https://github.com/facebookresearch/XLM/issues/199>.

²<https://github.com/facebookresearch/XLM>

ply the CoSDA-ML to it with the same hyper-parameters. Besides, our method not only obtains 2.9% improvement on average score but also outperforms the reported results (Average 75.1 score) from Conneau and Lample [2019], which further demonstrates the effectiveness of our method. 3) Our method outperforms *Attention-Informed Mixed Training* in both DST and SLU tasks, which indicates that our dynamic sampling and multi-lingual code-switch data training technique are more effective for aligning representation between source and target languages than only translating one word to the target language.

4.4 Analysis

Robustness. To verify the robustness of CoSDA-ML, we conduct experiments with different token replacement ratios β during the fine-tuning process and keep the sentence replacement ratio α as 1. The results are shown in Figure 4(a) and 4(b). With all the values of β , our model consistently outperforms the state-of-the-art model (*Attention-Informed Mixed Training*) in slot filling and intent detection when $\beta > 0.7$, which verifies the robustness of our method.

Varying Amounts of Training Data. We study the effectiveness of CoSDA-ML by varying amounts of training data. Figure 4(c) and 4(d) report the results of adding varying amounts of training data between *Attention-Informed Mixed Training* and our model. We have two interesting observations: 1) Our augmentation method consistently outperforms the baseline with all training data sizes, which demonstrates consistency. 2) Using only 1/10 of the training data, our

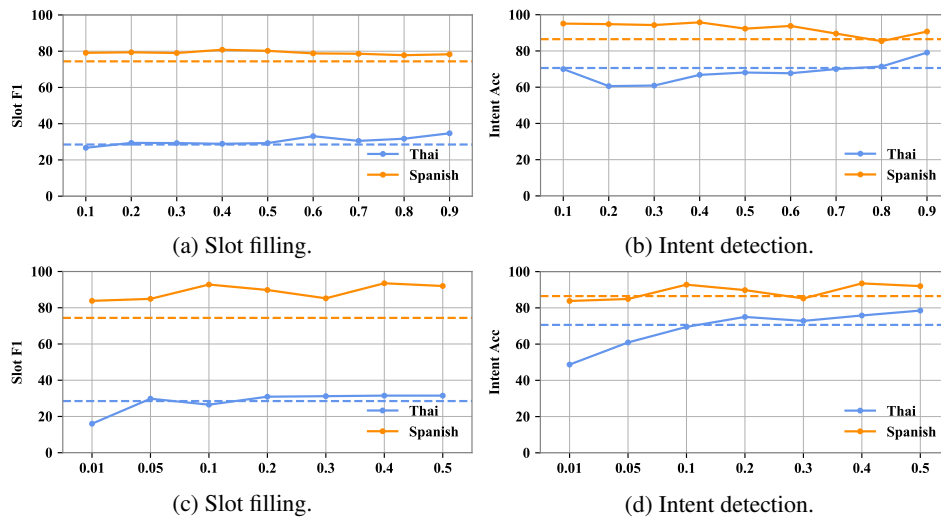


Figure 4: Comparison between our model (solid lines) and Attention-Informed Mixed Training (Att.) model (dashed lines). Results with different β in (a) and (b) and different subset size of training data in (c) and (d). In (c) and (d), it's worth that the dashline denotes Att. performance with 100% training data and the solid line represents our model performance by varying the proportion training data size.



Figure 5: t-SNE visualization of sentences vector space from mBERT (a) and with our CoSDA-ML method (b). The different color represents different languages and the dots in the same color denotes sentence representation with same intent.

approach performs better than the *Attention-Informed Mixed Training* using 100% of the training data, demonstrating that our approach is particularly useful when we only access to small amounts of training data.

Effectiveness of Dynamic Sampling. To verify the effectiveness of our proposed dynamic augmentation mechanism, we make comparison with *static* augmentation method, in which we adopt Algorithm 1 to obtain augmented multi-lingual code-switch training data once for all the batches. The results are shown in the *static* row of Table 5. We find that the dynamic method outperforms the *static* method in all the tasks. We attribute this to the fact that the dynamic mechanism can generate more varying code-switched multi-lingual data within the batch training process while *static* method can only augment one time of origin training data. Dynamic sampling allows the model to align more words representation closer in multiple languages.

Visualization. In order to see whether our framework aligns the representation between the source language and all the target languages, we select three intents with 100 sentences respectively and obtain their sentence vector [CLS] to visualize between our method with mBERT. The mBERT results are shown in Figure 5(a). We can see that there is

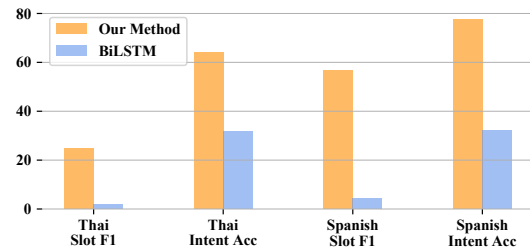


Figure 6: Evaluation result based on BiLSTM.

nearly no overlap between different languages, which shows that the distance of the representations of different languages with the same intent is distant. In contrast, the representations from our CoSDA-ML fine-tuned model in Figure 5 (b) in different languages become closer and overlap with each other, which further demonstrates that our method effectively and successfully aligns representations of different languages closer.

CoSDA-ML with BiLSTM. A natural question that arises is whether our augmentation method is effective for a general encoder in addition to Transformer. To investigate the question, we replace mBERT with BiLSTM and keep other components the same. BiLSTM does not include any information pre-trained over Wikipedia pages of multiple languages. We conduct experiments on top of BiLSTM to better verify whether our method strongly depends on the pre-trained model. The results are shown in Figure 6. We can see that our framework outperforms BiLSTM in all metrics in all languages, which further demonstrates that our augmentation method is not only effective on top of mBERT but also can work on a general encoder.

5 Related Work

Zero-shot Cross-lingual Transfer. The main strands of work focused on learning *cross-lingual word embeddings*. Ruder *et al.* [2017] surveyed methods [Klementiev *et al.*, 2012; Kočiský *et al.*, 2014; Guo *et al.*, 2016] for learning cross-lingual word embeddings by either joint training or post-training mappings of monolingual embeddings. Xing *et al.* [2015], Lample *et al.* [2018] and Chen and Cardie [2018] proposed to take pre-trained monolingual word embeddings of different languages as input, aligning them into a shared semantic space. Our work follows in the recent line of cross-lingual contextualized embedding methods [Huang *et al.*, 2019; Devlin *et al.*, 2019; Wu and Dredze, 2019; Conneau and Lample, 2019; Artetxe *et al.*, 2019], which are trained using masked language modeling or other auxiliary pre-training tasks to encourage representation in source and target language space closer, achieving state-of-the-art performance on a variety of zero-shot cross-lingual NLP tasks. We propose a data augmentation framework to dynamically construct multi-lingual code-switching data for training, which encourages model implicitly to align similar words in different languages into the same space.

Data Augmentation. Recently, some augmentation methods have been successfully applied in the cross-lingual setting. Liu *et al.* [2019b] proposed an attention mechanism to select the most important word to translate into the target language for training. In contrast, our framework can augment data dynamically in each epoch to encourage the model to align the representation in different languages, and can generate multiple languages code-switch data making training once and directly testing for all languages multiple times. Zhang *et al.* [2019] proposed using code-mixing to perform the syntactic transfer in dependency parsing. However, they need a high-accuracy translator to obtain multiple language data which can be difficult to train for low-resource language. In contrast, our method uses the existing bilingual dictionaries, which can be more practical and useful.

6 Conclusion

We proposed an augmentation framework to generate multi-lingual code-switching data to fine-tune mBERT for aligning representations from source and multiple target languages. Experiments on five tasks show that our method consistently and significantly outperforms mBERT and XLM baselines. In addition, our method is flexible and can be used to fine-tune all base encoder models. Future work includes the application of CoSDA-ML on the task of multi-lingual language modeling task, so that a more general version of the multi-lingual contextual embedding can be investigated.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61976072, 61632011 and 61772153. Besides, this work also faxed the support via Westlake-BrightDreams Robotics research grant. We thank Yijia Liu for the helpful discussion and anonymous reviewers for the insightful comments. Wanxiang Che and Yue Zhang are the corresponding author.

References

- [Artetxe and Schwenk, 2018] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*, 2018.
- [Artetxe *et al.*, 2019] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.
- [Barnes *et al.*, 2018] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proc. of ACL*, pages 2483–2493, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [Chen and Cardie, 2018] Xilun Chen and Claire Cardie. Unsupervised multilingual word embeddings. *arXiv preprint arXiv:1808.08933*, 2018.
- [Chen *et al.*, 2018] Wenhua Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. XL-NBT: A cross-lingual neural belief tracking framework. In *Proc. of EMNLP*, October–November 2018.
- [Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067, 2019.
- [Conneau *et al.*, 2018] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proc. of EMNLP*, 2018.
- [Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.
- [Guo *et al.*, 2016] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*, 2016.
- [Huang *et al.*, 2019] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proc. of EMNLP*, November 2019.
- [Klementiev *et al.*, 2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *Proc. of COLING*, 2012.
- [Kočiský *et al.*, 2014] Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. Learning bilingual word representations by marginalizing alignments. In *Proc. of ACL*, June 2014.
- [Lample *et al.*, 2018] Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

- [Liu *et al.*, 2019a] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [Liu *et al.*, 2019b] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems, 2019.
- [Mrkšić *et al.*, 2017] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017.
- [Ruder *et al.*, 2017] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*, 2017.
- [Schuster *et al.*, 2019a] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proc. of NAACL*, June 2019.
- [Schuster *et al.*, 2019b] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proc. of NAACL*, June 2019.
- [Schwenk and Li, 2018] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, May 2018.
- [Wang *et al.*, 2019] Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proc. of EMNLP*, November 2019.
- [Wu and Dredze, 2019] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proc. of EMNLP*, 2019.
- [Xing *et al.*, 2015] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. of NAACL*, 2015.
- [Yin *et al.*, 2019] Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Dialog state tracking with reinforced data augmentation. *arXiv preprint arXiv:1908.07795*, 2019.
- [Yu *et al.*, 2018] Katherine Yu, Haoran Li, and Barlas Oguz. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, July 2018.
- [Zhang *et al.*, 2019] Meishan Zhang, Yue Zhang, and Guohong Fu. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proc. of EMNLP*, pages 997–1006, Hong Kong, China, November 2019. Association for Computational Linguistics.