Saliency-based Multi-View Mixed Language Training for Zero-shot Cross-lingual Classification

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen*, Jinan Xu and Jian Liu
Beijing Jiaotong University, China
{20120374,18112023,20120370,chenyf,jaxu,jianliu}@bjtu.edu.cn

Abstract

Recent multilingual pre-trained models, like XLM-RoBERTa (XLM-R), have been demonstrated effective in many cross-lingual tasks. However, there are still gaps between the contextualized representations of similar words in different languages. To solve this problem, we propose a novel framework named **M**ulti-**V**iew **M**ixed **L**anguage **T**raining (MVMLT), which leverages code-switched data with multi-view learning to fine-tune XLM-R. MVMLT uses gradient-based saliency to extract keywords which are the most relevant to downstream tasks and replaces them with the corresponding words in the target language dynamically. Furthermore, MVMLT utilizes multi-view learning to encourage contextualized embeddings to align into a more refined language-invariant space. Extensive experiments with four languages show that our model achieves state-of-the-art results on zero-shot cross-lingual sentiment classification and dialogue state tracking tasks, demonstrating the effectiveness of our proposed model[1].

1 Introduction

Due to the availability of large labeled datasets and parallel corpus, neural network models have achieved remarkable performance on a variety of *natural language processing* (NLP) tasks. However, generally large-scale training data with high quality is only available in a few languages. Artificially collecting or translating training data for different languages could be time-consuming and expensive, which will inevitably create a massive performance gap between high-resource language models (e.g., English and French) and low-resource language models (e.g., Swahili and Urdu).

Cross-lingual transfer learning (CLTL) aims at bridging this gap by transferring the learned knowledge from a resource-rich language (source) to a resource-lean language (target) (David Yarowsky and Wicentowski, 2001). The main idea of CLTL is to learn a shared language-invariant feature space for both languages, so that a model trained on the source language could be applied to the target language directly. Recently, *Cross-Lingual Contextualized Embedding* methods such as multilingual BERT (mBERT) (Devlin et al., 2018), XLM (Conneau and Lample, 2019), and XLM-RoBERTa (XLM-R) (Conneau et al., 2019) have achieved state-of-the-art results on a variety of zero-shot cross-lingual tasks. However, those BERT-style transformer (Vaswani et al., 2017) architectures, training cross-lingual embeddings from self-supervised masked language modelling with monolingual corpus, may not well capture the semantic similarity of subwords across different languages.

In order to alleviate inconsistent contextualized representations within different languages, some supervised cross-lingual signals have been introduced in prior work (Kulshreshtha et al., 2020a), e.g., bilingual dictionaries and parallel corpora. Qin et al. (2020) propose a data augmentation framework called *Code-Switching* or *Mix Language Training*, which chooses a set of words randomly and replaces them with the corresponding words in a different language. For example, "I 喜欢 this 电影 so much"[2] is a code-switched sentence. They only use a bilingual dictionary to generate code-switched data to fine-tune mBERT, which encourage model to align representations between different languages. Nevertheless, there are two main problems in this method: (1) the importance of different words in a document is ignored, since they just replace words with the same probability randomly. Replacing some unimportant words will increase the burden of translation and even introduce noise that impairs the sentence semantic

*Yufeng Chen is the corresponding author.

[1]The code is publicly available at https://github.com/lisasiyu/MVMLT

[2]English: I love this movie so much!

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 599–610
November 7–11, 2021.

coherence; (2) they only use code-switched corpus to fine-tune mBERT, while the relation between original sentences and code-switched sentences is ignored completely, which may leads to the loss of some interactive information and hinder contextualized embeddings from further alignment.

To address the issues mentioned above, we propose a new framework named **M**ulti-**V**iew **M**ixed **L**anguage **T**raining (MVMLT), which leverages code-switched data with multi-view learning for zero-shot cross-lingual transfer. MVMLT first uses gradient-based saliency method to find keywords with high saliency scores in downstream tasks (Section 3.1). For example, in cross-lingual sentiment classification tasks, some words with sentiment information (e.g., "excellent", "interesting" and "boring") should have higher saliency scores than background words (e.g., "the", "a" and "what"). Relying on a bilingual dictionary, we replace these keywords with their corresponding words in the target language to generate code-switched data (Section 3.2). These code-switched keywords are the essential part for effective cross-lingual transfer, because they intersect with different languages and allow the shared encoder to learn some direct tying of meaning across different languages. Therefore, selecting the most task-related keywords by saliency detection facilitates cross-lingual performance for providing a strong tie across different languages.

Furthermore, MVMLT acquires comprehensive cross-lingual information from different perspectives and explores the consistency of multiple views by means of multi-view learning (Xu et al., 2013). Specifically, MVMLT constructs two views from the multilingual pre-trained model, i.e., XLM-R: (1) the encoded feature representation of the original sentence; (2) the encoded feature representation of the corresponding code-switched sentence. The key of cross-lingual transfer is to learn a language-invariant feature space, so these two feature representations should be as similar as possible. Therefore, we utilize multi-view learning to enforce a consensus between two views, which encourages similar words in different languages to align into a shared latent space (Section 3.3).

In summary, our main contributions are as follows:

- We propose a saliency-based mixed language training (MLT) framework, which utilizes gradient-based saliency to select task-related words for code-switching. Focusing on these keywords allows model to transfer cross-lingual signals more efficiently.
- We leverage multi-view (MV) learning to constrain the representation of original sentence and code-switched sentence consistently, and build a refined language-invariant space that is more robust to language shift compared to previous zero-shot cross-lingual transfer work (Liu et al., 2020; Fei and Li, 2020; Qin et al., 2020).
- Our MVMLT model is extensively evaluated in four languages on cross-lingual sentiment classification and dialogue state tracking tasks in zero-shot setting, and achieves state-of-the-art results in 10/11 tasks, demonstrating the effectiveness of MVMLT.

2 Related Work

2.1 Cross-Lingual Transfer Learning

Cross-lingual transfer learning aims at leveraging the learned knowledge of the source language to cope with the related task of the target language. Learning *Cross-Lingual Word Embeddings* (CLWE) (Mikolov et al., 2013) is a successful method for CLTL, which uses a bilingual dictionary to project words that have the same meaning close to each other. Recently, *Cross-lingual Contextualized Embeddings* use some form of language modeling to pre-train multilingual representations, which are then fine-tuned on the relevant tasks and transferred to different languages directly. Multilingual pre-trained models such as multilingual BERT (Devlin et al., 2018), XLM (Conneau and Lample, 2019), and XLM-RoBERTa (Conneau et al., 2019) have been successfully used for zero-shot cross-lingual transfer on various tasks (Wu and Dredze, 2019; Pires et al., 2019), i.e., Document Classification, Named Entity Recognition and Dependency Parsing. In addition, these multilingual pre-trained models can be further improved by different alignment methods (Kulshreshtha et al., 2020b; Cao et al., 2020), like rotation-based alignment and fine-tuning alignment. Our work is inspired by Qin et al. (2020), which propose a data augmentation framework and use task-related parallel word pairs to generate code-switched sentences for fine-tuning mBERT. The difference is that we use saliency detection to choose keywords rather than select-

ing words randomly. Moreover, we leverage code-switched data with multi-view learning to further align representations of multiple languages.

2.2 Multi-View Learning

Multi-view learning, aiming at learning from different views which contains complementary information and exploiting the consistency from multiple views (Li et al., 2019), has been widely used in many NLP tasks. Clark et al. (2018) proposed Cross-View Training (CVT), a novel self-training algorithm that works well for neural sequence models. Zhang et al. (2019) unified multiple views of entities to learn better embedding representations for entity alignment. Fei and Li (2020) proposed multi-view encoder-classifier (MVEC) for sentiment classification, which enforced a consensus between multiple-views (i.e., the encoded sentences in the source languages and the encoded back-translations of the source sentences from the target language) generated by encoder-decoder framework. Unlike MVEC, our model employs multi-view training to restrain the encoded representation of original sentence and code-switched sentence consistent without using parallel corpus.

2.3 Saliency Detection

Since attention mechanisms (Bahdanau et al., 2014) boosted performance on many current NLP tasks, using attention weight as explanation of model predictions is a general approach for many models (Wang et al., 2016; Lin et al., 2017; Ghaeini et al., 2018). However, some recent work (Serrano and Smith, 2019; Jain and Wallace, 2019) casts doubt on attention's interpretability. Besides, Bastings and Filippova (2020) claimed that saliency methods are more applicable for model explanations. There are three saliency methods for NLP as alternatives to attention (Arras et al., 2019): gradient-based (Denil et al., 2014), propagation-based (Bach et al., 2015), and occlusion-based (Zeiler and Fergus, 2014) methods. In our work, the gradient-based saliency method is adopted for selecting important words to be code-switched.

3 Methodology

Suppose we have two monolingual datasets $\{\mathcal{D}_{src}$ $\mathcal{D}_{tgt}\}$, where $\mathcal{D}_{src} = \{(x_i^S, y_i)\}_{i=1}^N$ is the labeled data only available in the source language L_S, and $\mathcal{D}_{tgt} = \{(x_i^T)\}_{i=1}^M$ is the unlabeled data in the target language L_T. We aim at using $\mathcal{D}_{src}$ to train an universal classification model and predicting the corresponding label when given an unseen language data $\mathcal{D}_{tgt}$.

The architecture of our model is illustrated in Figure1, which consists of three components: (1) **Gradient-based keyword selection:** selecting keywords in the training set and building a code-switched dictionary; (2) **Dynamic code-switching:** code-switching the input sentence dynamically; (3) **Multi-view training:** training the encoder based on multi-view learning. We will elaborate each part in this section.

3.1 Gradient-based Keyword Selection

Intuitively, the influence of each word in a sentence is different when training a classification model. We call those words that have a greater impact on model as *keywords*. Different tasks or domains usually have different keywords, e.g., for News Classification task, keywords set should include words like "military", "salary" and "sport", and for Sentiment Classification task, keywords set should include words like "interesting", "fascinating" and "unworthy". Suppose we have a vocabulary set $\mathcal{V}$ contains v words in a dataset, we need to find a salient subset of keywords $\mathcal{K} \subseteq \mathcal{V}$ for code-switching, which would improve downstream tasks greatly. So we utilize saliency scores for selecting keywords. Gradient-based saliency computes the gradient of the loss $\mathcal{L}$ with respect to each token in the input text, and the magnitude of the gradient serves as a feature importance score (Arras et al., 2019).

Formally, let $x_i^S = (w_i^1, w_i^2, \cdots, w_i^n)$ denotes the i-th sentence with n words from $\mathcal{D}_{src}$, $\mathcal{L}_{\hat{y}}$ is the loss between model's prediction $\hat{y}_i$ and the ground truth y_i. For each token $w_i \in x_i^S$, we define the saliency score as:

$$S_x(w_i) = -\nabla_{e(w_i)}\mathcal{L}_{\hat{y}} \cdot e(w_i), \qquad (1)$$

where $e(w_i)$ is the embedding of w_i. Thus, the saliency value is a dot product between prediction function gradient and word embedding, which is referred as *Gradient* $\times$ *Input* (Shrikumar et al., 2017). The *Gradient* shows how much one word embedding contributes to the final decision, and the *Input* leverages the sign and magnitude of the input. Note that multi-lingual pre-trained models tokenize words into subwords, so we average the subword saliency scores of each word as the final result.

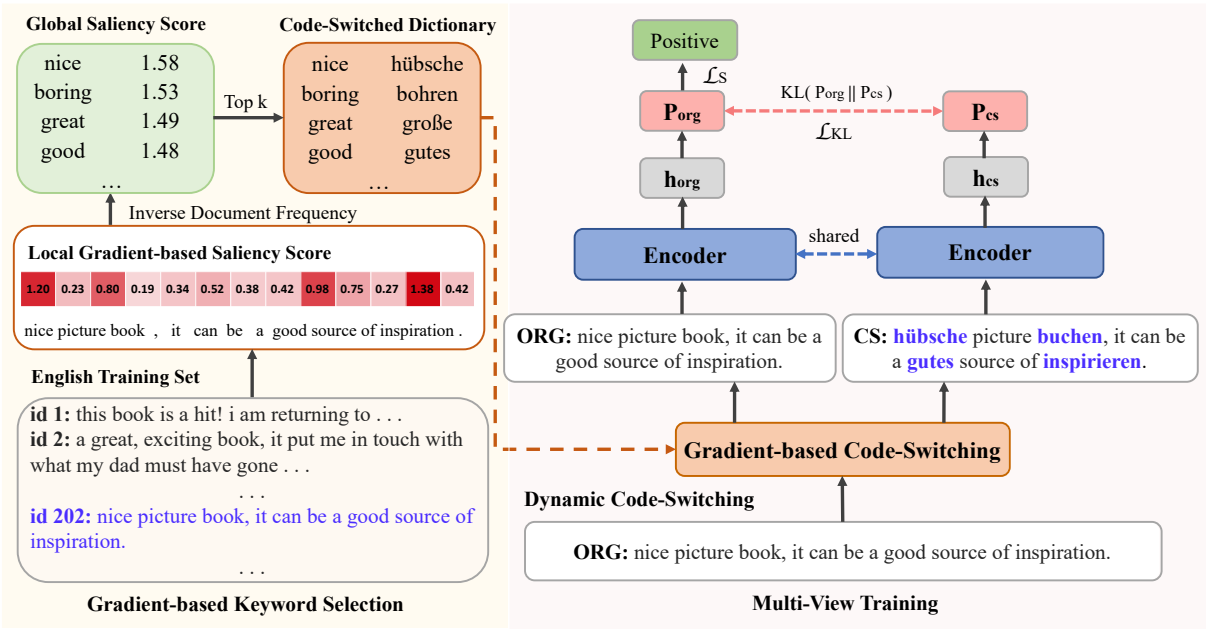


Figure 1: The overview of MVMLT architecture. **ORG** denotes the original English sentence and **CS** denotes the corresponding code-switched sentence. Left: the process of gradient-based keyword selection. Right: after dynamic code-switching, multi-view training jointly optimize cross-entropy loss $\mathcal{L}_S$ and KL loss $\mathcal{L}_{KL}$.

Equation 1 computes the local contribution of a token in one sentence, but we aim to build a global keyword set $\mathcal{K}$ in $\mathcal{D}_{src}$. Following Yuan et al. (2019), we add all saliency scores for token w occurred in $\mathcal{D}_{src}$ and multiply them with the inverse document frequency (IDF) of w:

$$S(w) = log\frac{N}{|\{x \in X : w \in x\}|} \cdot \sum_{x \in X : w \in x} S_x(w), \tag{2}$$

where N is the total number of words in $\mathcal{D}_{src}$. The IDF term balances word frequency and saliency scores by assigning words with high document frequency a lower weight and vice versa. It is necessary because some irrelevant *stop words* (e.g., "of" and "a") have high total saliency scores, for they appear in the document many times.

Top-k salient words are chosen to compose the keyword set $\mathcal{K}$, and a bilingual dictionary MUSE (Conneau et al., 2017) is adopted to build a code-switched dictionary $\mathcal{D} = ((s_1, t_1), \cdots, (s_k, t_k))$, where s and t represent the source and target language words, respectively. k is the number of keywords, and the influence of k value on model performance will be discussed in Section 5.3. The process of constructing code-switched dictionary is illustrated in the left part of Figure 1.

3.2 Dynamic Code-Switching

Given a source language sentence $x_{org} = (w_1, w_2, \cdots, w_n)$, we replace the words in x_{org} with their corresponding translation with a certain probability if they appear in $\mathcal{D}$. After this code-switching process, we get a code-switched sentence $x_{cs} = (w'_1, w'_2, \cdots, w'_n)$. Because the replaced words in source language could have multiple translations in the target language, we randomly choose one for replacement. In addition, we reset the replacement after each epoch, namely we replace different words at different epochs, which could be referred as a data augmentation method.

3.3 Multi-View Training

We train our MVMLT based on XLM-R architecture with multi-view learning. We first feed original sentence x_{org} and code-switched sentence x_{cs} into a shared XLM-R model separately:

$$\begin{aligned} \boldsymbol{h}_{org} &= Encoder(x_{org}), \\ \boldsymbol{h}_{cs} &= Encoder(x_{cs}), \end{aligned} \tag{3}$$

where $\boldsymbol{h}_{org}$ and $\boldsymbol{h}_{cs}$ are the aggregated sentence representation for the original sentence and the code-switched sentence, respectively.

For classification tasks, we input $\boldsymbol{h}_{org}$ and $\boldsymbol{h}_{cs}$

into a classification layer:

$$\begin{aligned} \boldsymbol{p}_{org} &= Softmax(\boldsymbol{W}\boldsymbol{h}_{org} + \boldsymbol{b}), \\ \boldsymbol{p}_{cs} &= Softmax(\boldsymbol{W}\boldsymbol{h}_{cs} + \boldsymbol{b}), \end{aligned} \tag{4}$$

where $\boldsymbol{p}_{org}$ and $\boldsymbol{p}_{cs}$ are the task-specific probability for all candidates, $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters.

Our main learning objective is to train the classifier to match predicted labels with the ground truth, so we minimize the following cross-entropy loss between $\boldsymbol{p}_{org}$ and ground truth label $\boldsymbol{p}$:

$$\mathcal{L}_S = CrossEntropy(\boldsymbol{p}_{org}, \boldsymbol{p}). \tag{5}$$

On the other hand, we hope the output produced by the encoder is language-invariant. To achieve this goal, we leverage multi-view learning to exploit a more comprehensive representation from multiple views which usually contain complementary information. We take two views into consideration: (1) the original sentence feature representation $\boldsymbol{h}_{org}$; (2) the code-switched sentence feature representation $\boldsymbol{h}_{cs}$. The central assumption of MVMLT is that an ideal model for cross-lingual transfer should learn feature representations that perform well in the source language and are invariant to the shift in the target language. Therefore, we enforce a consensus between these two views, that is to say, predicted distributions on the two views should be as similar as possible:

$$\mathcal{L}_{KL} = KL(\boldsymbol{p}_{org} \parallel \boldsymbol{p}_{cs}), \tag{6}$$

where KL is Kullback-Leibler (KL) (Kullback and Leibler, 1951) divergence to measure the difference between two distributions.

The final objective, combining the cross-entropy loss (Equation 5) and the KL divergence loss (Equation 6), is written as follows:

$$\mathcal{L}_{ALL} = \mathcal{L}_S + \lambda_{kl} \times \mathcal{L}_{KL}, \tag{7}$$

λ_{kl} is a hyper-parameter to trade-off cross-entropy loss and KL divergence loss, preventing the latter from drifting too far. The process of multi-view learning is illustrated in the right part of Figure 1.

4 Experiments

We evaluate the effectiveness of our proposed method on zero-shot cross-lingual dialog state tracking and sentiment classification tasks in four languages. In details, English is the source language, and the target languages are German, Italian, French and Japanese, respectively.

4.1 Datasets

Sentiment Classification (SC) For the sentiment classification task, we use the multilingual multi-domain Amazon review dataset (Prettenhofer and Stein, 2010) which contains three domains: book, DVD and music. Each domain contains the reviews in four different languages: English, German, French and Japanese, which provides us 9 tasks in total. There are 1000 positive and 1000 negative reviews for each domain in each language. We use English as the source language, and the others as the target language. Following Fei and Li (2020), we combine the English training and test sets and randomly sample 20% (800) documents as the validation set for selecting model, and use the rest 3200 samples for training.

Dialogue State Tracking (DST) The DST data we use is Multilingual WOZ 2.0 (Mrkšić et al., 2017), a restaurant domain dataset, which is expanded from WOZ 2.0 by including two more languages (German and Italian) besides English. Multilingual WOZ 2.0 contains 1200 dialogues for each language, where 600 dialogues are used for training, 200 for validation, and 400 for testing. The corpus contains three goal-tracking slot types: food, price range and area. It can be treated as a collection of binary classification problems by predicting the slot-value pair from a current utterance and the previous system acts. In the experiments, we do not have access to any training or validation dataset for German and Italian, we only use target language for testing.

4.2 Training Details

We leverage the XLM-R-base as Encoder in Equation 3, with 12 Transformer blocks, 768 hidden units, 12 self-attention heads. For DST task, we use Adam (Kingma and Ba, 2014) optimizer and set learning rate to 1e-5, λ_{kl} to 1, the number of batch size to 8, word replacement ratio to 0.5 and keyword ratio to 0.1. For SC task, the learning rate is 1e-6, λ_{kl} is 5, batchsize is 12, replacement ratio is 0.7, keywords ratio is 0.4 for German and French, 0.5 for Japanese. Our approach is implemented with Pytorch[3] and all experiments are conducted on an NVIDIA Tesla P100. All experiment results are the average score over 5 runs with random seeds.

[3] https://pytorch.org

Approach	German				French				Japanese			
	books	DVD	music	avg	books	DVD	music	avg	books	DVD	music	avg
BWE†	76.00	76.30	81.32	81.41	80.27	80.27	79.41	79.98	71.23	72.55	75.38	73.05
CLDFA	83.95	83.14	79.02	82.04	83.37	82.56	83.31	83.08	77.36	80.52	76.64	78.11
Transformer based												
mBERT†	84.35	82.85	83.85	83.68	84.55	85.85	83.65	84.68	73.35	74.80	76.10	74.75
XLM†	86.85	84.20	85.90	85.65	88.10	86.95	86.20	87.08	80.95	79.20	78.02	79.39
XLM-R	88.10	86.60	87.95	87.55	88.55	88.30	87.00	87.95	81.15	**83.95**	83.50	82.87
CoSDA‡	88.90	86.05	87.20	87.38	86.00	87.60	86.70	86.80	80.65	77.80	80.90	79.78
MVEC	88.41	87.32	89.97	88.61	89.08	88.28	88.50	88.62	79.15	77.15	79.70	78.67
MVMLT(Ours)	**91.48**	**90.15**	**90.61**	**90.75**	**91.38**	**90.73**	**88.68**	**90.26**	**82.53**	83.49	**84.40**	**83.47**

Table 1: Prediction accuracy of binary classification on the Amazon Reviews dataset, and the highest performance is in bold. '‡' denotes the cross-lingual version of CoSDA (Qin et al., 2020) fine-tuned on XLM-R. '†' denotes results from Fei and Li (2020).

4.3 Comparison Methods

We compare MVMLT with the following strong baselines.

BWE: Zou et al. (2013) used Bilingual Word Embeddings (BWEs) to transfer source word embeddings to target word embeddings.

CLDFA: Xu and Yang (2017) utilized adversarial feature adaptation technique to distill discriminative knowledge across languages on parallel corpus.

XL-NBT: Chen et al. (2018) distilled and transferred teacher's knowledge in the source language to student state tracker in the target languages.

MLT: Liu et al. (2020) used attention to generate code-switched sentence, and the replacement is static in each epoch.

Multilingual Pre-training Models: mBERT (Devlin et al., 2018), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2019) directly fine-tuned a single layer classifier based on pre-training language model.

CoSDA: Qin et al. (2020) leveraged multi-lingual code-switched data by replacing words randomly to fine-tune mBERT, achieving the current best result in multi-lingual transfer.

MVEC: Fei and Li (2020) leveraged an unsupervised machine translation system to construct an encoder-decoder framework with a language discriminator.

Approach	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
XL-NBT	55.00	30.80	71.00	41.20
MLT	69.50	32.20	69.50	31.40
Transformer based				
mBERT	57.61	14.95	53.34	12.88
XLM†	58.04	16.34	-	-
XLM-R	74.63	42.04	88.42	69.44
CoSDA	83.00	63.20	82.20	61.30
CoSDA(XLM-R)‡	84.77	59.60	85.86	61.00
MVMLT(Ours)	**88.88**	**70.84**	**93.44**	**81.41**

Table 2: Results on Multilingual WOZ 2.0. The slot accuracy individually compares each slot-value pair to its ground truth label. The joint goal accuracy compares the predicted dialogue states to the ground truth at each dialogue turn. '†' denotes results from (Liu et al., 2020). '‡' denotes our re-implemented results for this method based on XLM-R.

5 Results & Discussion

5.1 Overall Performance

Results of SC and DST are illustrated in Table 1 and Table 2, respectively. We can see that the fine-tuned multilingual pre-trained models like mBERT, XLM and XLM-R outperform all previous methods by a large margin, which indicates multilingual pre-trained models have a strong ability of cross-lingual transfer in zero-shot setting. Besides, compared with these strong baselines, our model MVMLT leads to significant improvements and achieves state-of-the art performance on 10/11 tasks. Particularly, in SC task, compared with CoSDA (Qin et al., 2020), our method improves 3.37, 3.46 and 3.69 on average for *de*, *fr* and *jp*, respectively. For DST task, MVMLT also achieves notable gains in both languages, especially for joint goal accuracy. All these results well demonstrate the effective-

Approach	German				French				Japanese			
	books	DVD	music	avg	books	DVD	music	avg	books	DVD	music	avg
Full model	**91.48**	**90.15**	**90.61**	**90.75**	**91.38**	**90.73**	**88.68**	**90.26**	**82.53**	**83.49**	**84.40**	**83.47**
w/o saliency	90.90	89.55	87.90	89.45	90.95	89.85	86.16	88.99	81.90	82.85	84.30	83.02
w/o multi-view	88.30	88.05	85.25	87.20	88.70	88.50	83.70	86.90	81.20	81.80	81.85	81.62

Table 3: Ablation study on Amazon reviews dataset for three languages.

ness of the proposed MVMLT, which is mainly attributed to leverage code-switched data with multi-view learning for cross-lingual transfer.

We also find MVMLT greatly improves XLM-R when the target language is more similar to the source language. For example, MVMLT improves a lot when transfer to German, French and Italian, but has limited improvement in Japanese. We hypothesize that English and Japanese belong to different language families and have completely different linguistic structures. In the process of code-switching, word-to-word replacement will disrupt the linguistic structure, especially for distant languages. Therefore, we can not simply map the English and Japanese sentence representations into the same space.

5.2 Ablation Study

We conduct an ablation study to explore the effect of saliency detection and multi-view learning on the overall performance. The results are reported in Table 3.

w/o saliency: selecting keywords randomly rather than extracting keywords based on saliency leads to approximately 1% degradation, which indicates that saliency has a strong ability to pick out the most important words in different downstream task documents.

w/o multi-view: the performance is also significantly degraded when the multi-view learning is substituted by just mixing original and code-switched sentences together, and feed them to the encoder independently. Without multi-view learning, the interactive information between original sentences and code-switched sentences is ignored completely, so that the distribution of the latent representations are discrepant between source and target languages, which leads to a 2% performance degradation.

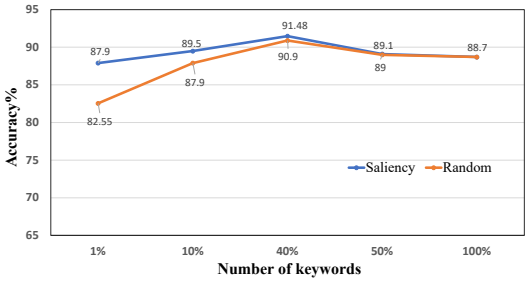


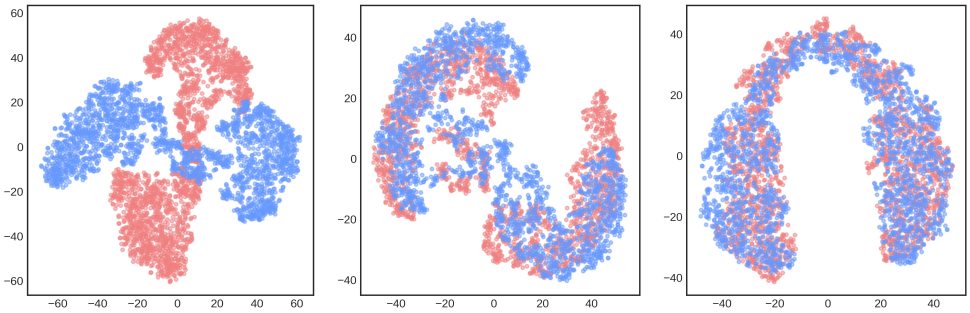
Figure 2: Test accuracy on German book domain as a function of the replacement rate $\frac{k}{v}$. **Random** denotes results from selecting keywords randomly. **Saliency** denotes results from selecting keywords by saliency detection.

5.3 Effectiveness of Saliency Detection

Figure 2 shows the influence of different strategies (i.e., selecting keywords by gradient-based saliency and selecting keywords randomly) with respect to different keyword sizes $\frac{k}{v}$.

The performance of selecting keywords randomly significantly declines when $\frac{k}{v}$ drops, while saliency-based method performs still well even with just 1% keywords (about 200 words). This is because gradient-based saliency helps MVMLT prioritize the most indicative keywords for code-switching. These keywords serve as powerful *anchor points* (i.e,. identical strings that appear in both languages in the training corpus) (Wu et al., 2019) for cross-lingual transfer, and provide sufficient cross-lingual information for aligning different languages representations into a shared space. As $\frac{k}{v}$ increases, the additional keywords are less indicative, so they have a minor or even negative effect on model performance.

It demonstrates that MVMLT remains effective under a minimal translation budget by leveraging gradient-based saliency to detect the most task-related keywords. Appendix A.1 shows the top 10 extracted keywords and their translations to German, French and Japanese in SC corpus.



(a) (b) (c)

Figure 3: t-SNE visualization of sentence vector space from XLM-R (a), CoSDA based on XLM-R (b), and our MVMLT method (c). Blue dots denote English sentence representations and pink dots denote German sentence representations.

	German	French	Japanese
KL	**91.48**	**91.38**	**82.53**
DIS	87.95	89.95	81.75
SIM	87.65	90.90	81.05

Table 4: Accuracy for three languages in book domain. **KL** denotes multi-view learning by calculating KL divergence. **DIS** denotes the distance-based alignment. **SIM** denotes the similarity-based alignment.

5.4 Effectiveness of Multi-View Learning

5.4.1 Visualization

We visualize the encoder's output of different methods for 2000 sampled parallel corpus in English and German provided by Amazon Reviews dataset with t-SNE (Van der Maaten and Hinton, 2008)

The XLM-R results in Figure 3(a) show that there is almost no overlap between the two language representations. CoSDA in Figure 3(b) further reduces the distance of representations by introducing code-switched sentences, but there are still some mismatching parts in the space. By leveraging multi-view learning, MVMLT in Figure 3(c) significantly decreases the distributional discrepancies between English and German instances.

It demonstrates that MVMLT effectively learns the language-invariant representations of different languages by multi-view training.

5.4.2 Compared with other alignments

Furthermore, we also try two other strategies to align multilingual embeddings directly.

Distance-based alignment minimizes the distance between the two contextual representations:

$$\mathcal{L}_{ALL} = \mathcal{L}_S + \lambda_{dis} \times \parallel \boldsymbol{h}_{org} - \boldsymbol{h}_{cs} \parallel . \quad (8)$$

Similarity-based alignment minimizes the similarity between the two contextual representations, and we use cosine similarity here:

$$\mathcal{L}_{ALL} = \mathcal{L}_S + \lambda_{sim} \times sim(\boldsymbol{h}_{org}, \boldsymbol{h}_{cs}). \quad (9)$$

As results shown in Table 4, we can conclude that minimizing the KL divergence between two probability distributions by multi-view learning is better than aligning contextual embeddings directly. Due to the different semantic structures and translation biases across different languages, forcing the encoded features to be exactly identical is harmful for its representation ability. While multi-view learning encourages two predicted distributions as close as possible, which gives model a softer way to learn language invariant representations.

5.4.3 MVMLT with Translate-Train

In this section, we add the third view called *Translate-Train*, which is the translation of the source language sentences by a Machine Translation system[4] trained on Europarl[5] corpus. The objective is written as follows:

$$\begin{aligned}\mathcal{L}_{ALL} = \mathcal{L}_S &+ \lambda_{kl1} \times KL(\boldsymbol{p}_{org} \parallel \boldsymbol{p}_{cs}) \\ &+ \lambda_{kl2} \times KL(\boldsymbol{p}_{org} \parallel \boldsymbol{p}_{trans}),\end{aligned} \quad (10)$$

where $\boldsymbol{p}_{trans}$ is predicted probability of *translate-train*, λ_{kl1} and λ_{kl2} are set to 1.

The results are shown in Table 5. We can see that translate-train further improves the performance of MVMLT by offering an additional view. On the one hand, translate-train compensates for the shortcomings of code-switching that sometimes

[4] https://github.com/facebookresearch/fairseq

[5] https://statmt.org/europarl/

Approach	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
ORG	74.63	42.04	88.42	69.44
CS	84.77	59.60	85.86	61.00
TRANS	81.86	53.90	86.67	64.34
Multi-View training				
ORG + CS (MVMLT)	88.88	70.84	93.44	81.41
ORG + TRANS	86.80	65.86	91.82	78.31
ORG + CS + TRANS	**91.27**	**76.61**	**94.61**	**85.36**

Table 5: Accuracy on DST. **ORG** denotes original sentences. **CS** denotes code-switched sentences. **TRANS** denotes translate-train sentences.

breaks the semantic coherence. On the other hand, code-switching offers more target-related information compared to translate-train. Therefore, model could learn more robust cross-lingual representations from these complementary views.

However, it is an overkill to introduce a more complex translation system because large parallel data may not be available in every language. Overall, our MVMLT is still a simple yet efficient framework that can achieve promising scores, which is more suitable for rare-language and limited-budget scenarios.

6 Conclusion

In this paper, we propose **M**ulti-**V**iew **M**ixed **L**anguage **T**raining (MVMLT), a novel zero-shot cross-lingual transfer framework. Our approach utilizes gradient-based saliency to replace a few task-related words with target language, which is used for fine-tuning on downstream tasks. Besides, we introduce multi-view learning to construct a language-invariant feature space. Experiments show that our model achieves state-of-the-art results on cross-lingual sentiment classification and dialogue state tracking tasks. In the future, we will investigate the effectiveness of our approach in multi-lingual setting and apply our model to more tasks.

Acknowledgements

The research work descried in this paper has been supported by the National Nature Science Foundation of China (No. 61976016, 61976015, and 61876198) and the National Key R&D Program of China (2020AAA0108001). The authors also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*.

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

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. *arXiv preprint arXiv:1808.06244*.

K. Clark, M. T. Luong, C. D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

GN David Yarowsky and R Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. of the 1st International Conference on Human Language Technology Research (HLT)*, pages 161–168.

Misha Denil, Alban Demiraj, and Nando De Freitas.

2014. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Reza Ghaeini, Xiaoli Z Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *arXiv preprint arXiv:1808.03894*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

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

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang. 2020a. Cross-lingual alignment methods for multilingual bert: A comparative study. *arXiv preprint arXiv:2009.14304*.

Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang. 2020b. Cross-lingual alignment methods for multilingual bert: A comparative study. *arXiv preprint arXiv:2009.14304*.

Yingming Li, Ming Yang, and Zhongfei Zhang. 2019. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. *arXiv preprint arXiv:1705.02073*.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2019. Interactive refinement of cross-lingual word embeddings. *arXiv preprint arXiv:1911.03070*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view

knowledge graph embedding for entity alignment. pages 5429–5435.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

A Appendix

A.1 Examples of Keywords

ENGLISH —>	GERMAN	FRENCH	JAPANESE
1. book	buchen	livre	書物
2. read	lesen	lecture	読み
3. informative	informativen	instructif	有益
4. great	große	génial	グレート
5. good	guten	bien	良い
6. interesting	interessanten	intéressante	興味深い
7. disappointed	enttäuschen	déçu	がっかり
8. novel	novelle	roman	小説
9. better	besser	meilleures	ベター
10. witty	witzige	witzig	ユーモア

(a) book

ENGLISH —>	GERMAN	FRENCH	JAPANESE
1. great	große	génial	グレート
2. movie	kino	film	映画
3. good	guten	bien	良い
4. excellent	ausgezeichnet	excellent	優れた
5. bad	böse	méchant	バッド
6. worst	schlimmste	pire	最悪
7. classic	klassisch	classique	クラシック
8. poor	schlecht	pauvre	かわいそう
9. love	liebe	amour	ラヴ
10. funny	lustige	drôle	面白い

(b) DVD

ENGLISH —>	GERMAN	FRENCH	JAPANESE
1. great	große	génial	グレート
2. good	guten	bien	良い
3. best	beste	meilleur	最高
4. disappointed	enttäuschen	déçu	がっかり
5. bad	böse	méchant	バッド
6. music	musik	musique	音楽
7. excellent	ausgezeichnet	excellent	優れた
8. awesome	geil	impressionnant	すごい
9. terrible	schrecklich	horrible	ひどい
10. like	wie	aimez	ライク

(c) music

Figure 4: Top 10 selected keywords by gradient-based saliency detection for the “book”, “DVD” and “music” domain, and their translations to German, French and Japanese by MUSE dictionary.