Unsupervised Domain Adaptation of a Pretrained Cross-Lingual Language Model

Juntao Li[1,2*] , **Ruidan He**[2] , **Hai Ye**[2] , **Ng Hwee Tou**[2] , **Lidong Bing**[3] , **Rui Yan**[1]
[1]Center for Data Science, Academy for Advanced Interdisciplinary Studies, Peking University
[2]School of Computing, National University of Singapore
[3]DAMO Academy, Alibaba Group
{lijuntao, ruiyan}@pku.edu.cn, {dcsyeh, nght}@comp.nus.edu.sg
{ruidan.he, l.bing}@alibaba-inc.com

Abstract

Recent research indicates that pretraining cross-lingual language models on large-scale unlabeled texts yield significant performance improvements over various cross-lingual and low-resource tasks. Through training on one hundred languages and terabytes of texts, cross-lingual language models have proven to be effective in leveraging high-resource languages to enhance low-resource language processing and outperform monolingual models on low-resource languages. In this paper, we further investigate the cross-lingual and cross-domain setting when a pretrained cross-lingual language model needs to adapt to new domains. Specifically, we propose a novel unsupervised feature decomposition method that can automatically extract domain-specific features and domain-invariant features from the entangled pretrained cross-lingual representations given unlabeled raw texts in the source language. Our proposed model leverages mutual information estimation to decompose the representations computed by a cross-lingual model into domain-invariant and domain-specific parts. Experimental results show that our proposed method achieves significant performance improvements over the state-of-the-art pretrained cross-lingual language model.

1 Introduction

Recent progress in deep learning benefits a variety of NLP tasks and leads to significant performance improvements when large scale annotated datasets are available. For high-resource languages, e.g., English, it is feasible for many tasks to collect sufficient labeled data to build up deep neural models. While for many languages, there might not exist enough data in most cases to make full use of the advances of deep neural models. In view of this situation, various *cross-lingual transfer learning* methods are proposed to utilize labeled data from high-resource languages to construct deep models in low-resource languages [Kim *et al.*, 2019; Lin *et al.*, 2019; He *et al.*, 2019; Vulić *et al.*, 2019], among which, most cross-lingual transfer learning researches focus on mitigating the discrimination of languages while leaving the domain gap less explored. *In this study, we concentrate on a more challenging setting, i.e., cross-lingual and cross-domain (CLCD) transfer, where in-domain labeled data in the source language is not available.*

Conventionally, cross-lingual methods mainly rely on extracting language-invariant features from data to transfer knowledge learned from the source language to the target language. One straightforward method is weight sharing, which directly reuses the model parameters trained on the source language to the target language by mapping input text to a shared embedding space beforehand. However, previous research [Chen *et al.*, 2018] revealed that weight sharing is not sufficient for extracting language-invariant features that can generalize well across languages, as a result of which a language-adversarial training strategy was proposed to extract invariant features across languages, using non-parallel unlabeled texts from each language. Such strategy performs well for the bilingual transfer setting but is not suitable for extracting language-invariant features from multiple languages insomuch as features shared by all source languages might be too sparse to retain useful information.

Recently, cross-lingual language model pretraining methods at scale, e.g., multilingual BERT [Devlin *et al.*, 2019], XLM [Lample and Conneau, 2019; Conneau *et al.*, 2019], show very competitive performance over various cross-lingual tasks and even outperform pretrained monolingual models on low-resource languages. Through employing parallel texts (unlabeled for any specific task) and shared subword vocabulary over all languages, these pretrained cross-lingual models can effectively encode input text from multiple languages to one single representation space, which also refers to a feature space shared by multiple languages (more than one hundred). While generalizing well for extracting language-invariant features, cross-lingual pretraining methods have no specific strategy for extracting domain-invariant features. In our CLCD setting, both domain-invariant and language-invariant features are need to be extracted.

To address the aforementioned limitation of cross-lingual pretrained models [Conneau *et al.*, 2019] in CLCD scenarios, we propose an unsupervised feature decomposition (UFD) method, which only leverages unlabeled data in the source

*This work is done at National Univerity of Singapore.

language. Specifically, our proposed method is built on top of the recently proposed unsupervised representation learning method [Hjelm *et al.*, 2019] and can simultaneously extract domain-invariant features and domain-specific features by combing mutual information maximization and minimization. Compared with previous cross-lingual transferring methods, our proposed model maintains the merits of cross-lingual pretraining models, i.e., generalize well for over a hundred languages, and only needs unlabeled data in the source language for domain adaptation, which is suitable for more cross-lingual transfer scenarios.

We evaluate our model on a benchmark cross-lingual sentiment classification dataset, i.e. Amazon Review [Prettenhofer and Stein, 2010], which involves multiple languages and domains. Experimental results indicate that, with the enhancement of the pretrained XLM cross-lingual language model, our proposed UFD model (trained on some unlabeled raw texts in the source language) along with a simple linear classifier (trained on a small labeled dataset in the source language and the source domain) outperforms the state-of-the-art models that have access to strong cross-lingual supervision (e.g., commercial MT systems) or labeled datasets in multiple source languages. Furthermore, through incorporating our proposed unsupervised feature decomposition strategy, a raw text dataset with 150k instances in the source language leads to continuous gains over the strong pretrained XLM model that is trained on one hundred languages and terabytes text. Extensive experiments further demonstrate that unsupervised feature decomposition upon pretrained cross-lingual language model outperforms pretrained domain-specific language model trained on over 100 million sentences.

2 Related Work

Cross-lingual transfer learning (CLTL) has long been investigated [Yarowsky *et al.*, 2001] and is still one of the frontiers of natural language processing [Chen *et al.*, 2019]. Through utilizing rich annotated data in high-resource languages, CLTL significantly alleviates the challenge of scarce training data in low-resource languages. Conventionally, CLTL mainly focuses on resources that are available for transferring, e.g., collecting parallel data between two languages to directly transfer model built in rich resource language to the low-resource one [Pham *et al.*, 2015], constructing annotated data in the target language by machine translation systems [Xu and Yang, 2017]. Later on, with the prosperity of deep learning, representation-based methods are proposed to model cross-lingual transfer learning in the feature space. Cross-lingual word embeddings learned the shared representation space at the fundamental level and can benefit various downstream tasks [Artetxe *et al.*, 2018; Conneau *et al.*, 2018a]. Later, the cross-lingual sentence representation learning method is also proposed for cross-lingual transferring [Conneau *et al.*, 2018b]. Chen et al. [2018] designed a language-adversarial training strategy to extract language-invariant features that can directly transfer to the target language.

Another direction for cross-lingual transfer learning is the recently proposed cross-lingual [Lample and Conneau, 2019] or multilingual language model pretraining methods [Devlin *et al.*, 2019]. Benefiting from the large scale training texts and model size, these pretraining methods have changed the phase of cross-lingual transfer learning. Empirical results demonstrate that representations space shared by one hundred languages can significantly outperform the language-specific pretrained models [Conneau *et al.*, 2019]. As language adversarial training will lead to sparse language-invariant representations when multiple languages are involved in cross-lingual transfer [Chen *et al.*, 2019], we follow the line of cross-lingual language models. Unlike previous cross-lingual language model pretraining methods, we focus on the domain adaptation of these pretrained models. To maintain the generalization ability of the cross-lingual pretrained model, we mainly consider the unsupervised domain adaptation setting. The most related work to us is proposed for unsupervised representation learning [Hjelm *et al.*, 2019], which is primarily for visual representation learning.

3 Model

In this section, we first define the problem discussed in this paper and then describe the proposed method in detail.

3.1 Problem Definition & Model Overview

In this paper, we consider a setting where we only have labeled set $D_{s,s}$ of a specific language and a specific domain which we call source language and source domain, and we want to train a classifier to be tested on set $D_{t,t}$ of a different language and a different domain which we call target language and target domain. We also assume access to some unlabeled raw data $D_{s,t}$ of the source language and the target domain at the training phase, which is usually feasible in practical applications. We call this setting unsupervised cross-lingual and cross domain (CLCD) adaptation.

As illustrated in Figure 1, the proposed method consists of three components: a pretrained multilingual embedding module which embeds the input document into a language-invariant representation, an unsupervised feature decomposition (UFD) module which extracts domain-invariant features and domain-specific features from the entangled language-invariant representation, and a task-specific module trained on the extracted domain-invariant and domain-specific features. We adopt XLM[1] [Lample and Conneau, 2019] as the multilingual embedding module in our method, which has been pretrained by large-scale parallel and monolingual data from various languages and is the current state-of-the-art cross-lingual language model. We describe the other two components and the training process in following subsections.

3.2 Unsupervised Feature Decomposition

Mutual Information Estimation

Before elaborating on the proposed unsupervised feature decomposition module, we first present some preliminary

[1]The latest version XLM-R is adopted, which is trained on over one hundred languages and 2.5 terabytes text.

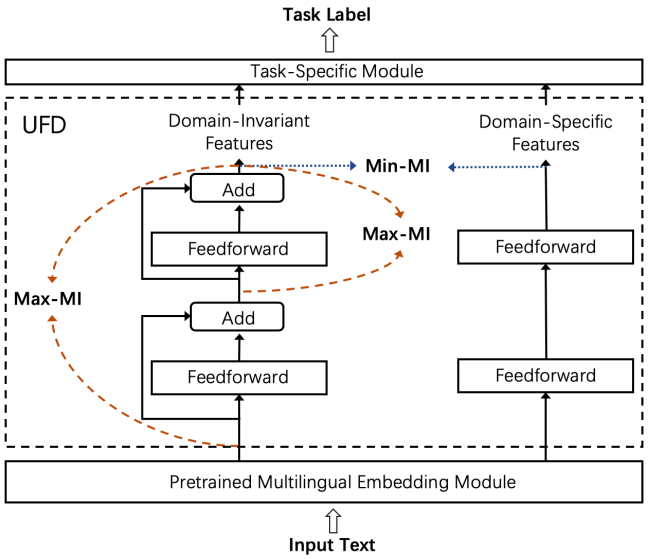


Figure 1: Our unsupervised domain adaptation model, where Min-MI and Max-MI refer to MI maximization and minimization. The middle-left part is the feature extractor $\mathcal{F}_s$ and the right is $\mathcal{F}_p$.

knowledge on mutual information estimation, which is employed in the training objectives of UFD. Mutual information (MI) is growing in popularity as an objective function in unsupervised representation learning. It measures how informative one variable is of another variable. In the context of unsupervised representation learning, MI maximization is usually adopted such that the encoded representation maximally encodes information of the original data. MI is extremely difficult to compute, particularly in continuous and high-dimensional settings, and therefore various estimation approaches have been proposed.

In our method, we adopt a recently proposed neural estimation approach [Belghazi *et al.*, 2018], which estimates MI of two continuous random variables X and Y by training a network to distinguish between samples coming from their joint distribution, $\mathbb{J}$, and the product of their marginal distributions, $\mathbb{M}$. This estimation utilizes a lower-bound of MI based on the Donsker-Varadhan representation (DV) of KL-divergence [Donsker and Varadhan, 1983],

$$\begin{aligned} \mathcal{I}(X;Y) &:= \mathcal{D}_{KL}(\mathbb{J}||\mathbb{M}) \geq \widehat{\mathcal{I}}^{DV}(X;Y) \\ &:= E_{\mathbb{J}}[T_\omega(x,y)] - \log E_{\mathbb{M}}[e^{T_\omega(x,y)}] \end{aligned} \tag{1}$$

where T_ω is a discrimination function parameterized by a neural network with learnable parameters ω. It maps a sample from space $X \times Y$ to a real value in $\mathbb{R}$. Through maximizing $\widehat{\mathcal{I}}^{DV}$, T_ω is encouraged to distinguish between samples drawn from $\mathbb{J}$ and $\mathbb{M}$ by assigning the former large values while the latter small ones.

Proposed Method

Let $X \in \mathbb{R}^d$ denote the language-invariant representations generated by the pretrained multilingual embedding module. It is then feeded into the proposed UFD module as input. As shown in Figure 1, we introduce two feature extractors: the domain-specific extractor $\mathcal{F}_p$ (i.e. the two-layer feedforward network with ReLU activation on the right), and the domain-invariant extractor $\mathcal{F}_s$ (i.e. the two-layer network on the left). We denote the extracted features as $\mathcal{F}_p(X)$ and $\mathcal{F}_s(X)$ respectively. Note that for $\mathcal{F}_s$, we add residual connections to better maintain domain-invariant attributes from X.

Specifically, $\mathcal{F}_s$ aims to extract domain-invariant features from the language-invariant representation in an unsupervised manner. Since the multilingual embedding module is pretrained on open domain datasets from over one hundred languages, presumably, the generated language-invariant representations should contain certain attributes that can be generalized across domains and this part of knowledge should be maximally maintained in the extracted domain-invariant features. Therefore, the language-invariant representations can serve as feedback signals for training $\mathcal{F}_s$ by maximizing MI between its inputs and outputs. In this way, $\mathcal{F}_s$ is forced to pass useful information from the language-invariant representations X to the domain-invariant features $\mathcal{F}_s(X)$.

We utilize the neural network-based estimator as presented in Equation (1) for computing MI. In our case, as $\mathcal{F}_s(X)$ is dependent on X, we can simplify the DV-based MI estimator to a Jensen-Shannon MI estimator as suggested in [Hjelm *et al.*, 2019]:

$$\begin{aligned} \widehat{\mathcal{I}}^{JSD}(X;\mathcal{F}_s(X)) := E_{\mathbb{P}}[-sp(-T_\omega(x,\mathcal{F}_s(x)))] \\ -E_{\mathbb{P}\times\widetilde{\mathbb{P}}}[sp(T_\omega(x^{'},\mathcal{F}_s(x)))] \end{aligned} \tag{2}$$

where x is an input embedding with empirical probability distribution $\mathbb{P}$. As $\mathcal{F}_s(x)$ is directly computed from x, $(x, \mathcal{F}_s(x))$ can be regarded as a sample drawn from the joint distribution of X and $\mathcal{F}_s(X)$. $x^{'}$ corresponds to an input embedding from $\widetilde{\mathbb{P}} = \mathbb{P}$, i.e., $x^{'}$ is computed from a random sample drawn from the same input distribution, such that $(x^{'}, \mathcal{F}_s(x))$ is drawn from the product of marginal distributions. $sp(z) = \log(1+e^z)$ is the softplus activation function. The training objective of $\mathcal{F}_s$ is to maximize the MI on X and $\mathcal{F}_s(X)$ and the loss is formulated as follows:

$$\mathcal{L}_s(\omega_s, \psi_s) = -\widehat{\mathcal{I}}^{JSD}(X, \mathcal{F}_s(X)) \tag{3}$$

where ω_s denotes the parameters of the discrimination network in the estimator and ψ_s denotes the parameters of $\mathcal{F}_s$. To facilitate domain-invariant features learning, we also propose to maximize the MI on $\mathcal{F}_s(X)$ and the corresponded intermediate representation (first layer output) $\mathcal{F}_s'(X)$, and the training loss is as follows:

$$\mathcal{L}_r(\omega_r, \psi_s) = -\widehat{\mathcal{I}}^{JSD}(\mathcal{F}_s'(X), \mathcal{F}_s(X)) \tag{4}$$

where ω_r denotes the parameters of the discriminator network in the estimator.

Recall that the objective of $\mathcal{F}_p$ is to extract domain-specific features, which is supposed to be exclusive and independent of domain-invariant features. We propose to minimize the MI between features extracted by $\mathcal{F}_s$ and $\mathcal{F}_p$, and the training loss is formulated as follows:

$$\mathcal{L}_p(\omega_p, \psi_s, \psi_p) = \widehat{\mathcal{I}}^{JSD}(\mathcal{F}_s(X), \mathcal{F}_p(X)) \tag{5}$$

where ψ_p denotes the parameters of $\mathcal{F}_p$. ω_p denotes the parameters of the discrimination network in MI estimator.

Datasets	English		
	Book	DVD	Music
#Documents	8,898,041	1,097,592	1,697,533
#Sentences	101,061,948	16,447,191	21,062,292
#Words	1,302,754,313	194,145,510	277,987,802
Avg Length	146.4	176.9	163.8

Table 1: Statistics of domain-specific raw texts.

The training objective of the proposed UFD component is thus to minimize the overall loss as follows:

$$\mathcal{L}_{UFD} = \alpha\mathcal{L}_s + \beta\mathcal{L}_r + \gamma\mathcal{L}_p \quad (6)$$

where α, β, and γ are hyper-parameters to balance the effects of sub-losses.

3.3 Task-specific Module

In the task-specific module, we first employ a linear layer that maps the concatenation of the domain-invariant and domain-specific features in $\mathbb{R}^{2d}$ into a vector representation in $\mathbb{R}^d$. A simple feedforward layer with softmax activation is then employed on this mapped vector representation to output the task label. We train this module on $D_{s,s}$ and the cross-entropy loss denoted as $\mathcal{L}_t$ is utilized as the training objective.

3.4 Training

Note that the parameters of the multilingual embedding module are pretrained and set to be frozen in the entire training process. We first optimize the parameters of UFD, i.e., $\{\widehat{\omega}_s, \widehat{\omega}_r, \widehat{\omega}_p, \widehat{\psi}_s, \widehat{\psi}_p\}$ by minimizing $\mathcal{L}_{UFD}$ on the unlabeled set $D_{s,t}$. Once the UFD module is trained, we fix its parameters and train the task-specific module by minimizing $\mathcal{L}_t$ on the labeled set $D_{s,s}$.

4 Experimental Setting

4.1 Datasets

We conduct experiments on the multi-lingual and multi-domain Amazon review dataset [Prettenhofer and Stein, 2010], which serves as a benchmark in previous cross-lingual sentiment analysis researches and also supports cross-lingual and cross-domain evaluation. In details, this dataset comprises of four languages, i.e., English, German, French, and Japanese, and each language contains three domains, i.e., Book, DVD, and Music. There are a training set and a test set for each domain in each language and both consist of 1,000 positive reviews and 1,000 negative reviews.

In our CLCD evaluation, we treat English as the only source language and attempt to adapt to the other three languages respectively. As each language contains three domains, we can construct 3×2 CLCD source-target pairs between English and a specific target language. Therefore, we have 18 CLCD source-target pairs in total considering all three target languages. At training phase, we first utilize some unlabeled raw data from the source language and target domain for optimizing the proposed UFD. For example, if we are adapting from English-DVD to German-Book, unlabeled data in English-Book is utilized for training UFD. Then, the training set from the source language and source domain is used for training the task-specific module. At testing phase, the model is evaluated on the test set of the target language and target domain.

We draw samples from 3 larger unannotated datasets of Book, DVD, and Music domains released in [He and McAuley, 2016]. The statistics of the three datasets are given in Table 1. We randomly sample 50K documents from each domain as the unlabeled domain-specific set in the source language (i.e. English) to be utilized at the training phase. To encourage the domain-invariant extractor to learn domain-invariant features, we utilize domain-specific unlabeled sets from all domains (50K*3) in training UFD module. We also show the change of model performance when varying the number of unlabeled samples in Section 5.

4.2 Baselines

We denote our proposed model as **XLM-UFD**, and we compare it with the following baselines:

CL-RL [Xiao and Guo, 2013] is a cross-lingual word representation learning method, which learns the connection between two languages by sharing part of the word vectors.

Bi-PV [Pham *et al.*, 2015] attempts to learn paragraph vectors in a bilingual context setting by sharing the distributed representations of unannotated parallel data from different languages.

CLDFA [Xu and Yang, 2017] is a cross-lingual distillation method which leverages a parallel corpus of documents. An adversarial feature adaptation strategy is applied for reducing the mismatch between the labeled data and the unlabeled parallel document.

MAN-MOE [Chen *et al.*, 2019] addresses the multi-lingual transfer setting, i.e., there are multiple source languages with labeled data. Building upon a language-adversarial training module, this model utilizes a mixture-of-experts (MOE) module to dynamically combine private features of different languages.

The above four baselines were originally proposed for adaptations in a cross-lingual setting, e.g. adapting from English-Book to German-Book. We report their official results released in original papers, which can be regarded as upper bounds for their CLCD performances. Note that the setting of MAN-MOE is different, where N to 1 adaption is performed, i.e. from N source languages to one target language. Thus, its cross-lingual performance cannot be simply viewed as the upper bound of its CLCD performance. We re-train the model in CLCD setting as another baseline described later. For the baselines described below, they are all trained in the CLCD setting.

ADAN [Chen *et al.*, 2018] exploits adversarial training to reduce the representation discrepancy between the encoded source and target embeddings.

MAN-MOE-D is the version of MAN-MOE trained in a CLCD setting. As this specific model performs N to 1 adaptation, it can adapt from multiple source domains from the same source language to a specific target domain and target language. In our experiments, MAN-MOE-D utilizes two source domains from the same source language. For example, when the target language and domain is German-Book,

Model	German				French				Japanese			
	Books	**DVD**	**Music**	**Avg**	**Books**	**DVD**	**Music**	**Avg**	**Books**	**DVD**	**Music**	**Avg**
CL-RL	79.9	77.1	77.3	78.1	78.3	74.8	78.7	77.3	71.1	73.1	74.4	72.9
Bi-PV	79.5	78.6	82.5	80.2	84.3	79.6	80.1	81.3	71.8	75.4	75.5	74.2
CLDFA	84.0	83.1	79.0	82.0	83.4	82.6	83.3	83.1	77.4	80.5	76.5	78.1
MAN-MOE	82.4	78.8	77.2	79.5	81.1	84.3	80.9	82.1	62.8	69.1	72.6	68.2
ADAN	82.7	77.1	79.2	79.6	75.9	75.2	73.8	74.9	72.5	72.3	74.3	73.0
MAN-MOE-D	82.8	80.1	81.6	81.5	83.0	85.5	82.0	83.5	70.5	76.0	70.8	72.4
Multi-BPE	51.0	53.4	53.0	52.5	50.5	51.4	51.1	51.0	50.0	49.8	50.0	49.9
DLM	52.1	53.7	53.3	53.0	57.4	51.5	55.2	54.7	52.8	51.5	50.8	51.7
XLM	80.4	84.9	79.3	81.5	86.4	86.3	83.2	85.3	81.7	81.6	84.1	82.5
XLM-UFD	**89.2**	**86.4**	**88.8**	**88.1**	**89.5**	**89.4**	**89.1**	**89.3**	**83.8**	**84.5**	**85.2**	**84.5**
XLM*	86.3	81.2	84.5	84.0	90.6	86.9	87.6	88.4	82.9	85.0	87.0	85.0

Table 2: Overall comparison of classification accuracy between our proposed model and baseline models. The upper part refers to the accuracy reported in previous studies in a cross-lingual setting while the middle part refers to our implemented models trained in a CLCD setting. XLM* denotes the XLM model trained on source language target domain labeled data. We report the average values of three runs.

MAN-MOE-D takes labeled set from both English-DVD and English-Music at training phase.

Multi-BPE combines the pretrained multilingual byte-pair embeddings in 275 languages [Heinzerling and Strube, 2018] [2] with the task-specific classifier used in our proposed model to perform CLCD adaptation. This model is used to calibrate the performance of the subword embeddings shared across multiple languages.

DLM is a pretrained domain-specific language model[3] implemented with the code of XLM [Lample and Conneau, 2019] [4]. It employs the pretrained multilingual byte-pair embeddsings as the initialized representations of input texts to mitigate the gap between the source language and the target languages. This model is presented for studying the effect of leveraging large scale domain-specific unlabeled text.

XLM refers to the model structure where we simply add a feedforward layer with softmax activation as the output layer on top of pretrained XLM [Conneau *et al.*, 2019].

4.3 Training Details

The hidden dimension of XLM is 1024. The input and output dimensions of the feedforward layers in both $\mathcal{F}_s$ and $\mathcal{F}_s$ are 1024. The discriminator of T_{ω_s}, T_{ω_r}, and T_{ω_p} share the same model structure as suggested in previous work [Hjelm *et al.*, 2019], i.e., the discriminator consists of two feedforward layers with ReLU activation. The input and output dimensions of the first feedforward layer in the discriminator are 2048 and 1024. The input and output dimensions of the second feedforward layer are 1024 and 1. The input dimension of the single-layer task-specific classifiers is 1024. All trainable parameters are initialised from an uniform distribution $[-0.1, 0.1]$.

We utilize 100 labeled data in the target language and target domain as the validation set, which is used for hyperparameter tuning and model selection during training. The hyperparameters are tuned on the validation set of a specific source-target pair, and are then fixed in all experiments of XLM-UFD. Specifically, both UFD and the task-specific module are optimized by Adam [Kingma and Ba, 2014] with a learning rate of 1×10^{-4}. The batch size of training UFD and the task-specific module are set to 16 and 8, respectively. The weights α, β, γ in Equation (6) are set to 1, 0.3, and 1, respectively. During training, the model that achieves the best performance (lowest loss) on the validation set is saved for evaluation purpose.

5 Results

Table 2 presents the model comparison results and Table 3 shows the results of different ablation tests on XLM-UFD. Classification accuracy is used as the evaluation metric.

5.1 Model Comparison

In Table 2, the top 4 models are trained in a cross-lingual setting, and the middle 6 models are trained in a CLCD setting. We repeat the experiment on each source-target pair for 3 times with different random seeds and record the average result on each pair. Each reported result for models trained in CLCD setting is the average result of the adaption performance from two source domains in English. For example, a result under German-Book is the average of adaption accuracies from English-DVD and English-Music.

We make the following observations from Table 2. (1) XLM-UFD achieves significantly better results over all baselines across all settings. It even substantially outperforms baselines trained in a cross-lingual setting with parallel text from source and target languages such as CLDFA, which is a much less challenging setting. (2) One interesting finding is that MAN-MOE-D performs better than MAN-MOE. One possible reason is that MAN-MOE involves multiple source language while invariant features shared by multiple languages might be too sparse to maintain enough information for extracting task-specific features. (3) Among the pretrained models, the multilingual byte pair embeddings (Multi-BPE) only achieves low performances. With the enhancement of large scale domain-specific unlabeled text, the domain-specific language model (DLM) taking the multilingual byte pair embeddings as inputs obtains observable performance gains but still has a large room for improvement.

[2] https://nlp.h-its.org/bpemb/multi/

[3] trained with the datasets presented in Table 1

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

	Settings	German				French				Japanese			
		Books	**DVD**	**Music**	**Avg**	**Books**	**DVD**	**Music**	**Avg**	**Books**	**DVD**	**Music**	**Avg**
Basic Model	XLM	80.4	84.9	79.3	81.5	86.4	86.3	83.2	85.3	81.7	81.6	84.1	82.5
Model Ablation	Max	84.5	82.0	81.9	82.8	81.1	83.2	81.8	82.0	81.8	80.6	81.5	81.3
	Max-Min	88.4	85.9	87.3	87.2	88.0	88.4	88.3	88.2	**84.4**	83.3	85.0	84.2
	2Max-Min	**89.2**	**86.4**	**88.8**	**88.1**	**89.5**	**89.4**	**89.1**	**89.3**	83.8	**84.5**	**85.2**	**84.5**
Unlabeled Data Sizes	1K*3	87.2	85.4	86.4	86.4	88.6	88.3	87.0	88.0	78.9	80.0	81.0	80.0
	2K*3	86.7	84.4	85.6	85.6	87.9	88.1	83.8	86.6	**84.2**	83.4	84.4	84.0
	5K*3	89.0	86.0	86.9	87.4	87.8	89.1	86.9	87.9	83.0	83.8	82.2	82.9
	10K*3	88.5	86.3	88.1	87.6	88.8	88.8	88.3	88.6	83.7	84.4	**85.5**	**84.5**
	50K*3	**89.2**	**86.4**	**88.8**	**88.1**	**89.5**	**89.4**	**89.1**	**89.3**	83.8	**84.5**	85.2	**84.5**

Table 3: Classification accuracy of ablations and using different sizes of unlabeled target domain data in the source language (i.e. English).

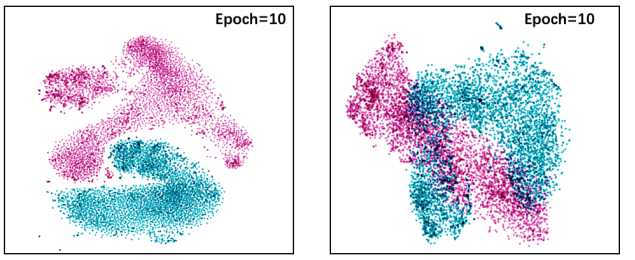
Benefited from the large scale training data and network size, XLM is able to perform better than the state-of-the-art task-specific models on French and Japanese such as CLDFA and MAN-MOE-D. When combined with the proposed UFD, significant performance gains are observed on XLM. This points out that domain adaptation is necessary for pretrained multilingual language models when applied to a specific task.

5.2 Ablation Study

To learn the effect of each module of XLM-UFD, we conduct a thorough model ablation. As presented in Table 3, we first examine the domain-invariant feature extractor along with the MI maximization between the language-invariant features from the multilingual embedding module and the extracted domain-invariant features, namely Max-MI. Classification accuracy shows that Max-MI with only domain-invariant features enhances the performance of XLM on German and leads to performance decreasing on French and Japanese. Through supplementing the domain-specific feature extractor and the Min-MI objective (i.e. $\mathcal{L}_p$), Max-Min-MI has a noticeable performance increase over Max-MI and outperforms XLM, which confirms that the unsupervised feature decomposition can support the dynamical domain-specific and domain-invariant feature combination and prompt the task performance. With the enhancement of the intermediate Max-MI objective (i.e. $\mathcal{L}_r$) between the intermediate features and output of domain-invariant feature extractor, 2Max-Min-MI achieves significant performance improvement over Max-MI, which is used as the main model for conducting other comparison and ablation. We also experiment with the influence of different sizes of unlabeled data of the source language. It can be seen from Table 3 that, 5K*3 unlabeled raw text already yields a very promising performance. Further increasing the unlabeled raw text will continuously improve the model performance on French and German. When the raw data size is larger than 10K*3, the performance improvement on Japanese becomes marginal and the performance on German and French still increase.

5.3 Visualization

To intuitively learn the process of domain-invariant feature and domain-specific feature extraction, we also give the t-SNE plots [Maaten and Hinton, 2008] of the UFD module at the tenth epoch. Specifically, we sample five thousand raw texts from the source domain and target language. Each raw text is processed by XLM, and the following domain-invariant feature extractor and domain-specific feature extractor, respectively. As presented in Figure 2, each data point in the plots represents an input text. We can observe from the left plot that the domain-invariant features and domain-specific features of input texts have a clear border that can be distinguished, which suggests that the mutual information minimization can force the two extractors exclusively extract domain-invariant and domain-specific features. The right plot in Figure 2 further demonstrates that domain-invariant features and language-invariant representations from XLM are partly entangled, which can be explained by the fact that maximizing mutual information between the language-invariant representations and the extracted domain-invariant features can force the domain-invariant feature extractor to pass the information that has certain shared attributes with the language-invariant representations.



(a) Domain-invariant & domain-specific (b) Domain-invariant & language-invariant

Figure 2: t-SNE plots, where the left figure refers to domain-invariant features and domain-specific features of input texts, and the right figure corresponds to domain-invariant features of input texts and language-invariant representations from XLM.

6 Conclusions and Future Work

In this paper, we propose a simple but effective unsupervised feature decomposition model to extend the cross-lingual model pretrained on over one hundred languages and terabytes texts to the cross-domain scenario. Through introducing the mutual information maximization and minimization objectives in representation learning, our proposed model can automatically extract domain-invariant and domain-specific features from the language-invariant cross-lingual space with only a small in-domain unlabeled dataset in the source language as the training data. Experimental results indicate that,

with the enhancement of our proposed model, the state-of-the-art cross-lingual language model XLM achieves continuous gains, which leads to the new SOTA on the Amazon review benchmark dataset. In the future, we will explore the effect of our proposed unsupervised feature decomposition model on other pretrained models and downstream tasks.

References

[Artetxe *et al.*, 2018] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, pages 789–798, 2018.

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *ICML*, pages 530–539, 2018.

[Chen *et al.*, 2018] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*, 6:557–570, 2018.

[Chen *et al.*, 2019] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. In *ACL*, 2019.

[Conneau *et al.*, 2018a] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *ICLR*, 2018.

[Conneau *et al.*, 2018b] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*, pages 2475–2485, 2018.

[Conneau *et al.*, 2019] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[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 *ACL, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[Donsker and Varadhan, 1983] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[He and McAuley, 2016] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

[He *et al.*, 2019] Junxian He, Zhisong Zhang, Taylor Berg-Kiripatrick, and Graham Neubig. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. *arXiv preprint arXiv:1906.02656*, 2019.

[Heinzerling and Strube, 2018] Benjamin Heinzerling and Michael Strube. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. In *LREC*, 2018.

[Hjelm *et al.*, 2019] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019.

[Kim *et al.*, 2019] Yunsu Kim, Yingbo Gao, and Hermann Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. *arXiv preprint arXiv:1905.05475*, 2019.

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

[Lample and Conneau, 2019] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *NeurIPS*, 2019.

[Lin *et al.*, 2019] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*, 2019.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

[Pham *et al.*, 2015] Hieu Pham, Thang Luong, and Christopher Manning. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94, 2015.

[Prettenhofer and Stein, 2010] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127, 2010.

[Vulić *et al.*, 2019] Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. Multilingual and cross-lingual graded lexical entailment. In *ACL*, pages 4963–4974, 2019.

[Xiao and Guo, 2013] Min Xiao and Yuhong Guo. Semi-supervised representation learning for cross-lingual text classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1465–1475, 2013.

[Xu and Yang, 2017] Ruochen Xu and Yiming Yang. Cross-lingual distillation for text classification. In *ACL*, pages 1415–1425, 2017.

[Yarowsky *et al.*, 2001] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics, 2001.