Towards a Unified End-to-End Approach for Fully Unsupervised Cross-lingual Sentiment Analysis

Yanlin Feng and **Xiaojun Wan**
Wangxuan Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{fengyanlin,wanxiaojun}@pku.edu.cn

Abstract

Sentiment analysis in low-resource languages suffers from the lack of training data. Cross-lingual sentiment analysis (CLSA) aims to improve the performance on these languages by leveraging annotated data from other languages. Recent studies have shown that CLSA can be performed in a fully unsupervised manner, without exploiting either target language supervision or cross-lingual supervision. However, these methods rely heavily on unsupervised cross-lingual word embeddings (CLWE), which has been shown to have serious drawbacks on distant language pairs (e.g. English - Japanese). In this paper, we propose an end-to-end CLSA model by leveraging unlabeled data in multiple languages and multiple domains and eliminate the need for unsupervised CLWE. Our model applies to two CLSA settings: the traditional cross-lingual in-domain setting and the more challenging cross-lingual cross-domain setting. We empirically evaluate our approach on the multilingual multi-domain Amazon review dataset. Experimental results show that our model outperforms the baselines by a large margin despite its minimal resource requirement. [1]

1 Introduction

While English sentiment analysis has achieved great success with the help of large-scale annotated corpus, this is not the case for most of languages where only limited data is available. Cross-lingual sentiment analysis (CLSA) tackles this problem by adapting the sentiment resource in a source language to a poor-resource language (the target language).

Current state-of-the-art CLSA methods rely heavily on cross-lingual word embeddings (CLWE) to transfer sentiment information from the source language to the target language. CLWE encodes words from multiple languages in a common space, thus making it possible to share a classifier across languages. Recent studies have shown that CLWE can be obtained in an unsupervised way, i.e., without any cross-lingual resources (Zhang et al., 2017; Conneau et al., 2017; Artetxe et al., 2018). This motivates fully unsupervised CLSA approaches (Chen et al., 2018a) that do not rely on either target language supervision or cross-lingual supervision. These methods generally involve the following steps:

1. Train monolingual embeddings separately on multiple languages using monolingual unlabeled data.

2. Map the monolingual embeddings to a shared space using unsupervised CLWE methods, either adversarial training methods (Conneau et al., 2017) or non-adversarial methods (Artetxe et al., 2018; Xu et al., 2018).

3. Train a sentiment classifier using the annotated corpus in the source language.

However, it has been shown that the quality of unsupervised CLWE is highly sensitive to the choice of language pairs and the comparability of the monolingual data (Søgaard et al., 2018). Therefore, these methods fail when the source language and the target language have very different linguistic structures (e.g. English and Japanese) and require additional cross-lingual supervision (e.g. a small seed dictionary or shared identical strings) in such cases (Chen et al., 2018a).

In this paper, we propose a unified end-to-end framework to perform unsupervised CLSA, bypassing the complex multi-step process and the drawbacks of unsupervised CLWE methods. Instead of mapping monolingual embeddings to a

[1] The source code is available at https://github.com/Evan-Feng/UXSenti

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 1035–1044
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

shared continuous space, we propose to bridge the language gap by multilingual multi-domain language modeling (i.e., we model the probabilities of sentences from multiple language-domain pairs). The language modeling objective is jointly trained with a classification objective in an end-to-end fashion using the unlabeled data in multiple language-domain pairs and labeled data in a source language-domain pair. Our model applies to two CLSA settings: the traditional cross-lingual in-domain setting and the more challenging cross-lingual cross-domain setting.

The rationale for using unlabeled data in multiple domains is that there may not be a domain shared by all languages in low-resource scenarios. If we want to perform CLSA on two languages that only have resources in two different domains, it is natural to bridge the language gap with another language that have resources on both domains. Even in the case where resources in a specific domain are available for all languages, which is a common assumption made by most CLSA approaches, we show that exploiting unlabeled data in other domains significantly improves performance.

Our contributions are as follows:

1. We propose a unified end-to-end framework to perform CLSA. Our approach is fully unsupervised and does not rely on any form of cross-lingual supervision (even shared identical strings) or target language supervision.

2. We show that cross-lingual language modeling based methods are able to outperform CLWE based methods in the unsupervised setting.

3. Our model can be easily generalized to different CLSA settings. Experiments on the multilingual multi-domain Amazon review dataset show that our method achieves state of the art in both the cross-lingual in-domain setting and the cross-lingual cross-domain setting despite its minimal resource requirement.

2 Related Work

Cross-lingual Sentiment Analysis The most related topic to our work is cross-lingual sentiment analysis. Some CLSA methods rely on machine translation systems (Wan, 2009; Demirtas and Pechenizkiy, 2013; Xiao and Guo, 2012; Zhou et al., 2016a) to provide cross-lingual supervision, making themselves implicitly dependant on large-scale parallel corpus which may not be available for low-resource languages. Wan (2009) apply the co-training algorithm to translated data while other researchers have proposed multi-view learning (Xiao and Guo, 2012).

Another line of CLSA research bridges the language gap using CLWE, which saves the efforts of training a machine translation system thus requires less cross-lingual resources. Some work has proposed to map pretrained monolingual embeddings to a shared space (Barnes et al., 2018) to obtain CLWE while others proposed jointly learning CLWE and a sentiment classifier, allowing the embeddings to encode sentiment information (Zhou et al., 2016b; Xu and Wan, 2017).

Very recently, unsupervised CLSA methods that do not require either cross-lingual supervision or target language supervision have been proposed (Chen et al., 2018b,a). Chen et al. (2018a) transfer sentiment information from multiple source languages by jointly learning language invariant and language specific features. Yet, these unsupervised CLSA methods rely on unsupervised CLWE which builds on the assumption that pretrained monolingual embeddings can be properly aligned. This assumption, however, is not true in low-resource scenarios (Søgaard et al., 2018).

It is worth pointing out that the language-adversarial training model of (Chen et al., 2018b) is able to perform unsupervised CLSA without CLWE. The proposed model consists of a feature extractor, a sentiment classifier and a language discriminator. The feature extractor is trained to fool the discriminator so that the extracted features are language invariant. However, its performance is significantly lower than the variant that uses pretrained CLWE.

While traditional CLSA methods assume that data in both languages is within the same domain (e.g. English hotel reviews for training and Chinese hotel review for testing, we refer to this setting as "cross-lingual in-domain sentiment analysis"), the more challenging cross-lingual cross-domain setting has also been explored. Ziser and Reichart (2018) extend pivot-based monolingual domain adaption methods to the cross-lingual setting. However, their method is not unsupervised and requires expensive cross-lingual resources.

Cross-lingual Language Modeling Our work is also related to cross-lingual language modeling, which is a topic that has been explored by researchers very recently. Lample and Conneau (2019), pretrain a language model with a joint vocabulary on the concatenation of multiple large-scale monolingual corpora and finetune it on labeled data. However, this approach exploits cross-lingual supervision provided by shared sub-word units, which has been shown to improve performance (Lample et al., 2018), and it remains a challenge to efficiently perform cross-lingual transfer without exploiting shared identical strings. In this work, we treat identical words from different languages as different words and thus eliminate any form of cross-lingual supervision.

Wada and Iwata (2018) proposed a similar cross-lingual language modeling architecture for unsupervised word translation. They show that it outperforms mapping based approaches (Artetxe et al., 2018; Lample et al., 2017), but only when a small amount of monolingual data is used. The difference between their model and ours is that we adopt different parameter sharing strategies and consider the correlation between multiple domains.

3 Cross-lingual In-Domain Sentiment Analysis

3.1 Overview

In this section we describe our cross-lingual in-domain sentiment analysis model (CLIDSA). It assumes the training data and test data come from different languages but are within the same domain (e.g. English hotel reviews as training data and Chinese hotel reviews as test data), which is the most common setting of previous CLSA approaches.

Although we use the same set of labeled data as previous CLSA approaches, we adopt a different strategy for utilizing the unlabeled data. Suppose there is a set of languages $\mathcal{L}$ and a set of domains $\mathcal{D}$. Let $\mathcal{P} \subseteq \mathcal{L} \times \mathcal{D}$ denote a set of language-domain pairs. For each language-domain pair $(l, d) \in \mathcal{P}$, we have a set of unlabeled reviews $\mathcal{C}_{mono}^{l,d}$. We also have a annotated sentiment corpus $\mathcal{C}_{senti}^{l_s,d_s}$ in a source language-domain pair (l_s, d_s). Our goal is to predict the sentiment polarity of the examples in a target language-domain pair (l_t, d_t) (note that $d_s = d_t$ in the cross-lingual in-domain setting).

In Section 5 we compare two CLIDSA variants. CLIDSA$_{full}$ exploits unlabeled data from all possible language-domain pairs, i.e., we set $\mathcal{P} = \mathcal{L} \times \mathcal{D}$. However, since most previous CLSA methods do not use multi-domain or multilingual unlabeled data, we create a variant CLIDSA$_{min}$ that requires minimal resources by setting $\mathcal{P} = \{(l_s, d_s), (l_t, d_t)\}$.

A natural way to utilize unlabeled data is to perform the language modeling task. Our CLIDSA model consists of multiple language models for mutiple language-domain pairs, with some of their parameters shared across languages or across domains. It also includes a classifier component which takes the hidden states (produced by the LSTM language model) as input features and predicts the sentiment polarity. We also adopt a language discriminator to force the features to be language invariant. The overall architecture of our model is illustrated in Figure 1. We detail each component of our CLIDSA model in the following subsections.

3.2 Multilingual Multi-Domain Language Modeling

Language modeling is the most critical part in our model since it acts as a language invariant feature extractor. Intuitively, if we share the LSTM layers of language models across languages, these layers are likely to process sentences from different languages in the same space, thus inducing language invariant features. In this subsection we detail our parameter sharing strategies for modeling sentences from multiple language-domain pairs.

Following previous work, we compute the probability of a sentence x by modeling the probability of a word w_k given the previous words:

$$p(x) = \prod_{k=1}^{|x|} p(w_k \mid w_1, \ldots, w_k - 1) \qquad (1)$$

For sentences in a certain language-domain pair (l, d), the probabilities are computed using a two-layer LSTM language model, which includes an embedding layer, two LSTM layers and a linear decoding layer. We first pass the input words through the embedding layer of language l which is parameterized by θ_{emb}^{l}. Then we forward the word embeddings to a LSTM layer parameterized by θ_{lstm1}, which is shared across all languages and all domains, generating a sequence of intermediate

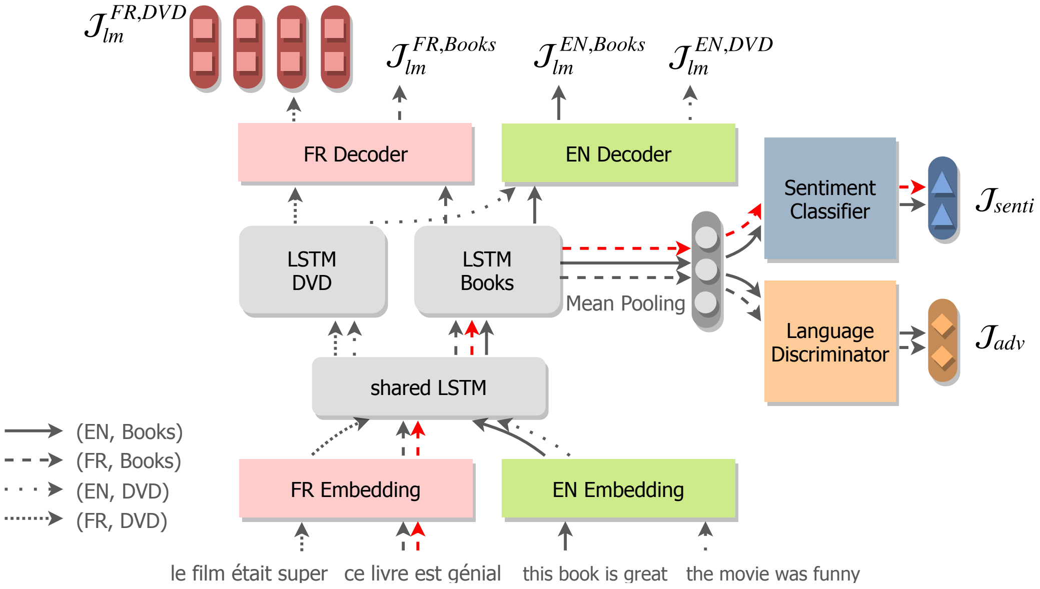

Figure 1: Illustration of the CLIDSA model. In this example, $\mathcal{L} = \{\text{EN}, \text{FR}\}$, $\mathcal{D} = \{\text{Books}, \text{DVD}\}$, $\mathcal{P} = \mathcal{L} \times \mathcal{D}$, $l_s = \text{EN}$, $l_t = \text{FR}$ and $d_s = d_t = \text{Books}$. We visualize the forward pass of input sentences from different language-domain pairs. The path shown in red only occurs at test time.

hidden states:

$$h_k = \text{LSTM}(h_{k-1}, \vec{w}_k; \theta_{lstm1}) \qquad (2)$$

where $\vec{w}_k$ denotes the embedding of word w_k. These hidden states are then passed through the second LSTM layer which is domain specific but language invariant, generating a sequence of final hidden states:

$$z_k = \text{LSTM}(z_{k-1}, h_k; \theta_{lstm2}^{d}) \qquad (3)$$

where the second LSTM layer of domain d is parameterized by θ_{lstm2}^{d}. The final hidden states $(z_1, z_2, \ldots, z_{|x|})$ thus can be considered as language invariant features for cross-lingual classification.

For the purpose of language modeling, we adopt a language-specific linear decoding layer to transform the final hidden states into probability distributions for next word prediction. The decoding layer of language l is parameterized by θ_{dec}^{l} and is shared across domains.

The intuition of adopting a domain-specific LSTM layer is that the distribution of sentences varies across domains. For example, given the first three words "I love this", the next word is most likely to be "book" in a book review dataset or "movie" in a movie review dataset. While it is possible to address this issue by using domain-specific linear decoding layers, we find that sharing the decoders across domains substantially reduces the total number of parameters thus provides regularization when only limited resources are available (see Section 5.5 for the ablation study). Sharing the decoders further enables the weight tying technique (Inan et al., 2016) to tie the decoder weight with the embedding layer.

For language-domain pair (l, d), the language modeling objective is written as follows:

$$\mathcal{J}_{lm}^{l,d}(\theta_{emb}^{l}, \theta_{lstm1}, \theta_{lstm2}^{d}, \theta_{dec}^{l}) = \mathbb{E}_{x \sim \mathcal{C}_{mono}^{l,d}} \left[-\frac{1}{|x|} \sum_{k=1}^{|x|} \log p(w_k \mid w_1, \ldots, w_{k-1}) \right] \qquad (4)$$

where the sentence likelihood is normalized by the sentence length and $x \sim \mathcal{C}_{mono}^{l,d}$ indicates that x is sampled from the unlabeled text in $\mathcal{C}_{mono}^{l,d}$.

3.3 Sentiment Classifier

We adopt a simple linear classifier that takes the averaged final hidden states $\frac{1}{|x|} \sum_{k=1}^{|x|} z_k$ as input features and outputs the probabilities of different labels. The classification objective can be written as:

$$\mathcal{J}_{senti}(\theta_{emb}^{l_s}, \theta_{lstm1}, \theta_{lstm2}^{d_s}, \theta_{clf}) = \mathbb{E}_{(x,y) \sim \mathcal{C}_{senti}^{l_s,d_s}} \big[-\log p(y \mid x) \big] \qquad (5)$$

where $(x, y) \sim \mathcal{C}_{senti}^{l_s,d_s}$ indicates that the sentence x and its label y are sampled from the source sentiment corpus and θ_{clf} denotes the parameters of the linear classifier.

The classification objective is jointly minimized with the language modeling objective, allowing sentiment-specific supervision signals to backpropagate through the model so that it can learn to extract useful features for sentiment prediction.

3.4 Language Adversarial Training

To further force the features used for sentiment classification to be language invariant, we adopt the language adversarial training technique (Chen et al., 2018b). A language discriminator is trained to predict the language ID given the features by minimizing the cross entropy loss, while the LSTM network is trained to fool the discriminator by maximizing the loss:

$$\mathcal{J}_{adv}(\theta_{emb}, \theta_{lstm1}, \theta_{lstm2}^{d_s}, \theta_{dis}) = \mathbb{E}_{(x,l)}[-\log p(l \mid x)] \quad (6)$$

where $\theta_{emb} = \theta_{emb}^1 \oplus \cdots \oplus \theta_{emb}^{|\mathcal{L}|}$ denotes the parameters of all the embedding layers and θ_{dis} denotes the parameters of the language discriminator. The sentence x and the language id l are sampled from all the unlabeled data in domain $d_s = d_t$. We do not employ language adversarial training on the features of other domain since we only perform classification in a single domain.

3.5 The Full Objective Function

Putting all the components together, the final objective function is thus:

$$\mathcal{J}_{full}(\theta_{emb}, \theta_{lstm}, \theta_{dec}, \theta_{clf}, \theta_{dis}) = \sum_{(l,d)\in\mathcal{P}} \mathcal{J}_{lm}^{l,d} + \alpha \mathcal{J}_{senti} - \beta \mathcal{J}_{adv} \quad (7)$$

where $\theta_{lstm} = \theta_{lstm1} \oplus \theta_{lstm2}^1 \oplus \cdots \oplus \theta_{lstm2}^{|\mathcal{D}|}$ denotes the parameters of all the LSTM layers, $\theta_{dec} = \theta_{dec}^1 \oplus \cdots \oplus \theta_{dec}^{|\mathcal{L}|}$ denotes the parameters of all the decoding layers, α and β are the hyperparameters controlling the importance of the classification objective and the language adversarial training objective. Parameters θ_{dis} are trained to maximize this objective function while the others are trained to minimize it:

$$\hat{\theta}_{dis} = \arg\max_{\theta_{dis}} \mathcal{J}_{full} \quad (8)$$

$$(\hat{\theta}_{emb}, \hat{\theta}_{lstm}, \hat{\theta}_{dec}, \hat{\theta}_{clf}) = \arg\min_{\theta_{emb},\theta_{lstm},\theta_{dec},\theta_{clf}} \mathcal{J}_{full} \quad (9)$$

4 Cross-lingual Cross-Domain Sentiment Analysis

In this section we focus on a more challenging CLSA setting, where the training data and test data are from different languages and different domain (e.g. English hotel reviews as training data, Chinese book reviews as test data). We show that the CLIDSA model can be applied to this setting with only slight modification.

Following previous notations, we denote the source language-domain pair as (l_s, d_s) and the target language pair as (l_t, d_t) with $l_s \neq l_t$ and $d_s \neq d_t$. Ziser and Reichart (2018) rely on expensive resources to perform cross-lingual cross-domain transfer, including unlabeled data from (l_s, d_s) and (l_t, d_t), CLWE and machine translation. In this work, we also use the unlabeled data from (l_s, d_s) and (l_t, d_t). However, instead of relying on CLWE and machine translation, we propose to leverage the unlabeled data in a "pivot" pair (l_s, d_t) to bridge the language-domain gap. This is reasonable since source languages are those with rich resources and we do not use additional annotation. Formally, we have a set of unlabeled reviews for each language-domain pair in $\mathcal{P} = \{(l_s, d_s), (l_s, d_t), (l_t, d_t)\}$ and a set of labeled reviews from (l_s, d_s).

In the CLIDSA model, inputs from (l_t, d_t) do not go through the source-domain LSTM layer (parameterized by $\theta_{lstm2}^{l_s}$), thus can not be forwarded to the sentiment classifier for sentiment prediction. Nevertheless, we now show that we can directly apply the CLIDSA model to this setting by slightly altering the forward pass of (l_t, d_t). Figure 2 illustrates our CLCDSA model for cross-lingual cross-domain sentiment analysis. The architecture of CLCDSA is identical to CLIDSA (i.e. parameterized by the same set of parameters), but the data forwarding process is slightly different. The key idea is simple: instead of viewing the sentiment classifier as a linear classifier that takes the domain-specific final hidden states $z_1, z_2, \ldots, z_{|x|}$ as input features, we consider the source-domain LSTM layer and the linear classifier together as a "LSTM+Linear" classifier (parameterized by $\theta_{lstm2}^{d_s} \oplus \theta_{clf}$) that takes the domain invariant and language invariant hidden states $h_1, h_2, \ldots, h_{|x|}$ as input features. From

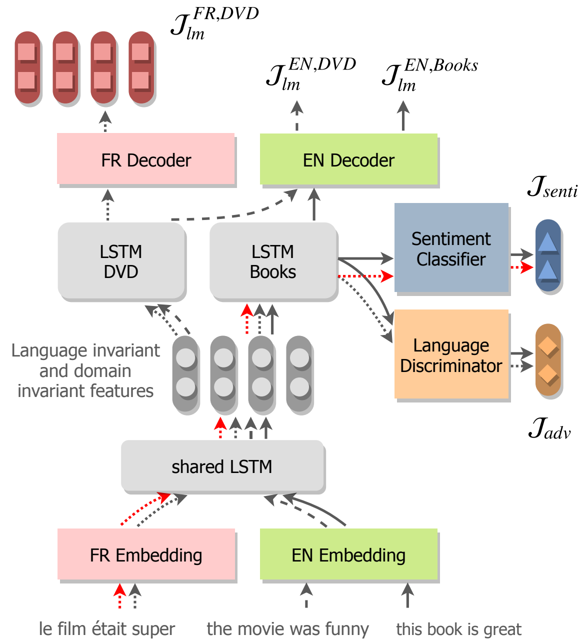

Figure 2: Illustration of the CLCDSA model. In this example, (l_s, d_s) = (EN, Books) and (l_t, d_t) = (FR, DVD). The architecture of CLCDSA is identical to CLIDSA, but the data forwarding process is different. We visualize the forward pass of input sentences from different language-domain pairs. The path shown in red only occurs at test time.

this point of view, we can pass the first-layer hidden states to the source-domain LSTM layer and the sentiment classifier to obtain the sentiment prediction at test time.

At training time, the first-layer hidden states generated from a target sentence are forwarded to the source-domain LSTM layer ($\theta_{lstm2}^{d_s}$) and the language discriminator (θ_{dis}) to compute the adversarial loss. The LSTM layers are trained to fool the language discriminator so that it cannot distinguish the examples in (l_s, d_s) from those in (l_t, d_t). We jointly optimize three language modeling objectives for each language-domain pair, the adversarial objective, and a sentiment classification objective for (l_s, d_s).

	EN	DE	FR	JA
Books	50000	165470	32870	169780
DVD	30000	91516	9358	68326
Music	25220	60392	15940	55892

Table 1: Number of unlabeled examples in the Amazon dataset.

5 Experiments

5.1 Datasets

We evaluate our model on the multilingual multi-domain Amazon review dataset (Prettenhofer and Stein, 2010) which contains product reviews in four languages (English, French, German, Japanese) and three domains (Books, DVD, Music). For each language-domain pair, there are 2000 examples for training and 2000 examples for testing. The statistics of unlabeled data is summarized in Table 1. For cross-lingual in-domain sentiment analysis, we use English as the source language and the others as target languages, resulting in nine tasks in total. For cross-lingual cross-domain sentiment analysis, we follow the setting in (Ziser and Reichart, 2018) and use English as the source language, French and German as target languages, and consider all the domain combinations, resulting in twelve tasks in total. Note that we would also want to evaluate our model on some low resource languages. However, since there isn't an public benchmark for such languages, we leave it to future work.

5.2 Implementation Details

Most of the hyperparamters are set empirically without tuning. For language modeling, we adopt the AWD-LSTM language model (Merity et al., 2017) with 1150 hidden units and a weight dropout rate of 0.5. We refer readers to (Merity et al., 2017) for a more detailed description. The sentiment classifier is a linear classifier with a dropout rate of 0.6. The language discriminator is a three-layer MLP with 400 hidden units.

As the only exceptions, hyperparameters α and β are tuned on the target development set following standard CLSA practice. We set (α, β) to $(0.01, 0.1)$ for CLIDSA$_{full}$ and CLCDSA, $(0.01, 0.03)$ for CLIDSA$_{min}$ in all tasks and do not perform any task-specific tuning.

The Adam optimizer (Kingma and Ba, 2014) with a base learning rate of 0.003 and $\beta_1 = 0.7$ is used for training. In each iteration, we sample a batch from every language-domain pairs in $\mathcal{P}$ to compute the language modeling loss and discriminator loss. Then we sample a batch from the source annotated corpus to compute the sentiment classification loss. All the parameters are jointly updated using the Gradient Reversal Layer (Ganin et al., 2016) and standard backpropagation. We

	EN-DE			**EN-FR**			**EN-JA**		
	Books	DVD	Music	Books	DVD	Music	Books	DVD	Music
	Methods with cross-lingual supervsion								
CL-SCL†‡	79.50	76.92	77.79	78.49	78.80	77.92	73.09	71.07	75.11
BiDRL†‡	84.14	84.05	84.67	84.39	83.60	82.52	73.15	76.78	78.77
UMM†	81.65	81.27	81.32	80.27	80.27	79.41	71.23	72.55	75.38
CLDFA†‡	83.95	83.14	79.02	83.37	82.56	83.31	77.36	80.52	76.46
	Methods without cross-lingual supervision								
MAN-MoE	82.40	78.80	77.15	81.10	84.25	80.90	62.78	69.10	72.60
MWE	76.10	76.80	74.70	76.35	78.70	71.60	-	-	-
CLIDSA$_{min}$	86.55	80.35	83.50	86.65	85.40	84.30	75.90	71.45	71.40
CLIDSA$_{full}$	**86.65**	**84.60**	**85.05**	**87.20**	**87.95**	**87.15**	**79.35**	**81.90**	**84.05**

Table 2: Test accuracy of different CLSA methods on the Amazon review dataset in the cross-lingual in-domain setting. The highest score on each task is shown in **bold**. The second highest score is underlined. '-'indicates that MUSE fails to align the EN and JA embeddings so MWE's predictions are random. Methods that require cross-lingual resources are marked as †. Methods that require machine translation are marked as ‡.

run 50000 iterations for CLIDSA and 30000 iterations for CLCDSA without early stopping.

5.3 Baselines

We compare our model to the following CLSA baselines, including methods that require cross-lingual resources (either in the form of machine translation or parallel data), methods that rely on unsupervised CLWE, and a few variants of our proposed model. **PBLM-BE** is a cross-lingual cross-domain model, **MWE** applies to both settings, while others are cross-lingual in-domain methods.

CL-SCL Prettenhofer and Stein (2010) map the bag-of-word representations to a cross-lingual space via structural correspondence learning.

BiDRL Zhou et al. (2016b) learn bilingual document representation for CLSA. The authors translate each document into both languages and enforce a bilingual constraint between the original document and the translated version.

UMM Xu and Wan (2017) jointly learn multilingual word embeddings and a sentiment classifier using parallel corpora of multiple language pairs. Languages that do not have direct parallel corpus are bridged via a third pivot language.

CLDFA Xu and Yang (2017) propose cross-lingual distillation using translated reviews.

MAN-MoE Chen et al. (2018a) propose the state-of-the-art unsupervised CLSA model that learns language invariant features and language specific features. It relies on unsupervised CLWE for cross-lingual transfer. Unlike other CLSA approaches, it transfers the sentiment information from multiple source languages.

MWE This is a variant of our proposed model that relies on unsupervised CLWE instead of language modeling. We map all target language embeddings to the English space using the MUSE library (Conneau et al., 2017) and use them to initialize the embedding layers. We train the sentiment classifier using the labeled data in the source language-domain pair and directly apply it to the test data. The same architecture is used but we only optimize the classification objective.

PBLM-BE Ziser and Reichart (2018) extend existing pivot-based domain adaption approaches to the cross-lingual settings using CLWE and machine translation.

5.4 Results and Analysis

Cross-lingual In-Domain Results Table 2 presents the performance of different CLSA methods on various cross-lingual in-domain tasks. Our proposed model achieves new state of the art on all nine tasks. Even in the restricted setting where only minimal resources are used (no cross-lingual resources, no pretrained embeddings, no multilingual multi-domain unlabeled data), CLIDSA$_{min}$ outperforms the strongest baseline on four out of nine tasks, validating the efficacy of our proposed model. Exploiting multilingual multi-domain unlabeled data leads to an average improvement of $+4.27\%$ across all tasks. We

	EN-DE						EN-FR					
	D-B	M-B	B-D	M-D	B-M	D-M	D-B	M-B	B-D	M-D	B-M	D-M
PMLM-BE†	78.7	78.6	80.6	79.2	81.7	78.5	81.1	74.7	76.3	75.0	75.1	76.8
MWE	76.3	72.8	74.7	72.5	74.2	76.0	74.8	72.4	76.0	74.2	72.5	74.3
CLCDSA	**85.4**	**81.7**	**79.3**	**81.0**	**83.4**	**81.7**	**86.2**	**81.8**	**84.3**	**82.8**	**83.7**	**85.0**

Table 3: Test accuracy of different CLSA methods on the Amazon review dataset in the cross-lingual cross-domain setting. The highest score on each task is shown in **bold**. Methods that require cross-lingual resources are marked as †. The abbreviations {B, D, M} stand for {Books, DVD, Music}.

	EN-DE	EN-FR	EN-JA
CLIDSA$_{full}$	**84.6**	**88.0**	**81.9**
- decoder sharing	83.0	87.0	78.9
- LSTM-1 sharing	82.4	87.1	81.4
- discriminator	82.6	87.4	81.6
- joint training	81.2	86.7	79.9

Table 4: Ablation results in the cross-lingual in-domain setting. English is used as the source language and DVD is used as the source/target domain. The highest score for each language pair is shown in **bold**.

also find that it is most beneficial to sentiment analysis on distant language pairs, with an average improvement of $+8.85\%$ on EN-JA.

Among methods that do not require cross-lingual resources, CLWE based methods are lower than the proposed cross-lingual language modeling based methods. This is interesting because it has been shown in previous work (Wada and Iwata, 2018) that cross-lingual language modeling does not perform well on the word translation task when sufficient monolingual data is available. Nevertheless, we demonstrate that this is not the case for cross-lingual sentiment analysis.

Cross-lingual Cross-Domain Results Table 3 shows the results of various cross-lingual cross-domain tasks. MWE suffers greatly from domain discrepancy compared to the in-domain results. Nevertheless, our model outperforms all baselines on all tasks, with an average improvement of $+5\%$ across all tasks.

5.5 Ablation Study

We perform an ablation study to investigate the contribution of individual components. The results are summarized in Table 4. We first create a variant that does not share the decoding layers across domain, and another one that does not share the first LSTM layer across domain. Disabling parameter sharing hurts the performance most on EN-JA (-1.75%). We also observed that the performance gap is much more significant when less training data is used (not shown here).

Surprisingly, removing the language discriminator does not lead to significant performance drop, which indicates that the language modeling alone is able to produce language invariant features. Intuitively, parameter sharing would force the LSTM layers to process sentences from different languages in the same space, thus inducing cross-lingual feature representation. Note that we also try removing the language modeling objective and rely on language adversarial training to provide cross-lingual features, but find that the resulting performance is rather poor.

Finally, we explore a different training strategy where the sentiment classifier is not jointly trained with the other components. Instead, we use the labeled data to train the classifier only after we have trained the other components on the unlabeled data. We observe that the resulting performance drop is due to underfitting, i.e., the extracted features do not encode enough information for sentiment prediction. This highlights the importance of end-to-end training.

6 Conclusion and Future Work

In this work we present an end-to-end approach for cross-lingual sentiment analysis. Our method is fully unsupervised thus does not rely on any cross-lingual supervision and target language supervision. We rely on language modeling to provide language invariant feature representations. We propose two model variants, one for cross-lingual in-domain transfer and the other for cross-lingual cross-domain transfer. Both models achieve state of the art on the Amazon review dataset. Experimental results also show that exploiting multilingual multi-domain unlabeled data greatly benefits CLSA on distant language pairs.

There are several straight-forward extensions of our model: cross-lingual in-domain sentiment analysis with multiple source languages, cross-lingual cross-domain sentiment analysis with multiple target languages, etc. We leave the exploration of these extensions to future work.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61772036) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. *arXiv preprint arXiv:1805.09016*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2018a. Zero-resource multilingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018b. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 9. ACM.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*.

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

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*.

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.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. *arXiv preprint arXiv:1805.03620*.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2012. Multi-view adaboost for multilingual subjectivity analysis. *Proceedings of COLING 2012*, pages 2851–2866.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520.

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

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1403–1412.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249.