Semi-Supervised Representation Learning for Cross-Lingual Text Classification

Min Xiao and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
{minxiao,yuhong}@temple.edu

Abstract

Cross-lingual adaptation aims to learn a prediction model in a label-scarce target language by exploiting labeled data from a label-rich source language. An effective cross-lingual adaptation system can substantially reduce the manual annotation effort required in many natural language processing tasks. In this paper, we propose a new cross-lingual adaptation approach for document classification based on learning cross-lingual discriminative distributed representations of words. Specifically, we propose to maximize the log-likelihood of the documents from both language domains under a cross-lingual log-bilinear document model, while minimizing the prediction log-losses of labeled documents. We conduct extensive experiments on cross-lingual sentiment classification tasks of Amazon product reviews. Our experimental results demonstrate the efficacy of the proposed cross-lingual adaptation approach.

1 Introduction

With the rapid development of linguistic resources in different languages, developing cross-lingual natural language processing (NLP) systems becomes increasingly important (Bel et al., 2003; Shanahan et al., 2004). Recently, cross-lingual adaptation methods have been studied to exploit labeled information from an existing *source* language domain where labeled training data is abundant for use in a *target* language domain where annotated training data is scarce (Prettenhofer and Stein, 2010). Previous work has shown that cross-lingual adaptation can greatly reduce labeling effort for a variety of cross language NLP tasks such as document categorization (Bel et al., 2003; Amini et al., 2009), genre classification (Petrenz and Webber, 2012), and sentiment classification (Shanahan et al., 2004; Wei and Pal, 2010; Prettenhofer and Stein, 2010).

The fundamental challenge of cross-lingual adaptation stems from a lack of overlap between the feature space of the source language data and that of the target language data. To address this challenge, previous work in the literature mainly relies on automatic machine translation tools. They first translate all the text data from one language domain into the other and then apply techniques such as domain adaptation (Wan et al., 2011; Rigutini and Maggini, 2005; Ling et al., 2008) and multi-view learning (Amini et al., 2009; Guo and Xiao, 2012b; Wan, 2009) to achieve cross-lingual adaptation. However, machine translation tools may not be freely available for all languages. Moreover, translating all the text data in one language into the other language is too time-consuming in reality. As an economic alternative solution, cross-lingual representation learning has recently been used in the literature to learn language-independent representations of the data for cross language text classification (Prettenhofer and Stein, 2010; Petrenz and Webber, 2012).

In this paper, we propose to tackle cross language text classification by inducing cross-lingual predictive data representations with both labeled and unlabeled documents from the two language domains. Specifically, we propose a cross-lingual log-bilinear document model to learn distributed representations of words, which can capture both the semantic sim-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1465–1475,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

ilarities of words across languages and the predictive information with respect to the target classification task. We conduct the representation learning by maximizing the log-likelihood of all documents from both language domains under the cross-lingual log-bilinear document model and minimizing the prediction log-losses of labeled documents. We formulate the learning problem as a joint non-convex minimization problem and solve it using a local optimization algorithm. To evaluate the effectiveness of the proposed approach, we conduct experiments on the task of cross language sentiment classification of Amazon product reviews. The empirical results show the proposed approach is very effective for cross-lingual document classification, and outperforms other comparison methods.

2 Related Work

Much work in the literature proposes to construct cross-lingual representations by using aligned parallel data. Basically, they first employ machine translation tools to translate documents from one language domain to the other one and then induce low dimensional latent representations as interlingual representations (Littman et al., 1998; Vinokourov et al., 2002; Platt et al., 2010; Pan et al., 2011; Guo and Xiao, 2012a). Littman et al. (1998) proposed a cross-language latent semantic indexing method to induce interlingual representations by performing latent semantic indexing over a dual-language document-term matrix, where each dual-language document contains its original words and the corresponding translation text. Vinokourov et al. (2002) proposed a cross-lingual kernel canonical correlation analysis method, which learns two projections (one for each language) by conducting kernel canonical correlation analysis over a paired bilingual corpus and then uses the two projections to project documents from language-specific feature spaces to the shared multilingual semantic feature space. Platt et al. (2010) employed oriented principal component analysis (Diamantaras and Kung, 1996) over concatenated parallel documents, which learns a multilingual projection by simultaneously minimizing the projected distance between parallel documents and maximizing the projected covariance of documents across different languages. Pan et al. (2011) proposed a bi-view non-negative matrix tri-factorization method for cross-lingual sentiment classification on the parallel training and test data. Guo and Xiao (2012a) developed a transductive subspace representation learning method for cross-lingual text classification based on non-negative matrix factorization. Some other works exploited parallel data by using multilingual topic models to extract cross-language latent topics as interlingual representations (Mimno et al., 2009; Ni et al., 2011; Platt et al., 2010; Smet et al., 2011) and using neural probabilistic language modes to learn word embeddings as cross-lingual distributed representations (Klementiev et al., 2012). Most of them were developed by applying the latent Dirichlet allocation (LDA) model (Blei et al., 2003) in a multilingual setting, including the polylingual topic model (Mimno et al., 2009), the bilingual LDA model (Smet et al., 2011), and the multilingual LDA model (Ni et al., 2011). Platt et al. (2010) extended the probabilistic latent semantic analysis (PLSA) model (Hofmann, 1999) and presented two variants of multilingual topic models: the joint PLSA model and the coupled PLSA model. Recently, Klementiev et al. (2012) extended the neural probabilistic language model (Bengio et al., 2000) to induce cross-lingual word distributed representations on a set of word-level aligned parallel sentences. The applicability of these approaches however is limited by the availability of parallel corpus. Translating the whole set of documents to produce parallel corpus is too time-consuming, expensive and even practically impossible for some language pairs. We thus do not evaluate those approaches in our empirical study.

Another group of works propose to use bilingual dictionaries to learn interlingual representations (Gliozzo, 2006; Prettenhofer and Stein, 2010). Gliozzo (2006) first translated each term from one language to the other using a bilingual dictionary and used the translated terms to augment original documents. Then they conducted latent semantic analysis (LSA) over the document-term matrix with concatenated vocabularies to obtain interlingual representations. Prettenhofer and Stein (2010) proposed a cross-language structural correspondence learning (CL-SCL) method to induce language-independent features by using word translation oracles. They first selected a subset of source language features, which have the highest mutual information with respect to the class labels in the labeled documents from the source language domain, to translate them into the target language domain, and then used these pivot pairs to induce cross-lingual representations by modeling the correlations between pivot features and non-pivot features. Our proposed approach shares a similarity with the CL-SCL method in (Prettenhofer and Stein, 2010) on only requiring a small amount of word translations. But our approach performs representation learning in a semi-supervised manner by directly incorporating discriminative information with respect to the target prediction task, while CL-SCL only exploits labels when selecting pivot features and the structural correspondence learning process is conducted in a fully unsupervised fashion.

Some other bilingual resources, such as multilingual WordNet (Fellbaum, 1998) and universal part-of-speech (POS) tags (Petrov et al., 2012), have also been exploited in the literature for interlingual learning. Gliozzo (2006) proposed to use MultiWordNet to map words from different languages to a common synset-id as language-sharing terms. A similar work was proposed in A.R. et al. (2012), which transformed words from different languages to WordNet synset identifiers as interlingual sense-based representations. However, multilingual WordNet resources are not always available for different language pairs. Recently, Petrenz and Webber (2012) used language-specific POS taggers to tag each word and then mapped those language-specific POS tags to twelve universal POS tags as interlingual features for cross language fine-grained genre classification. This approach requires a POS tagger for each language and it may be adversely affected by the POS tagging accuracy.

3 Semi-Supervised Representation Learning for Cross-Lingual Text Classification

In this section, we introduce a semi-supervised cross-lingual representation learning method and then use it for cross language text classification.

Assume we have ℓ_s labeled and u_s unlabeled documents in the source language domain $\mathcal{S}$ and ℓ_t labeled and u_t unlabeled documents in the target language domain $\mathcal{T}$. We assume all the documents are independent and identically distributed in each language domain, and each document $\mathbf{x}_i$ is represented as a bag of words, $\mathbf{x}_i = \{w_{i1}, w_{i2}, \ldots, w_{iN_i}\}$. We use $(\mathbf{x}_i^\ell, y_i)$ to denote the i-th labeled document and its label, and consider exploiting the labeled documents in the source domain $\mathcal{S}$ for learning classifiers in the target domain $\mathcal{T}$.

To build connections between the two language domains, we first construct a set of critical bilingual word pairs $M = \{(w_i^s, w_j^t)\}_{i=1}^m$, where w_i^s is a critical word in the source language domain, w_j^t is its translation in the target language domain, and m is the number of word pairs. Here being critical means the word should be discriminative for the prediction task and occur frequently in both language domains. Following the work (Prettenhofer and Stein, 2010), we select bilingual word pairs in a heuristic way. First we select a subset of words from the source language domain, which have the highest mutual information with the class labels in labeled source documents. The mutual information is computed based on the empirical distributions of words and labels in the labeled source documents. Then we translate the selected words into the target language using a translation tool to produce word pairs. Finally we produce the M set by eliminating any candidate pair (w^s, w^t), if either w^s occurs less than a predefined threshold value ϕ in all source language documents or w^t occurs less than ϕ in all target language documents. Given the constructed bilingual word pair set M, the words appearing in the source language documents but not in M can be put together to form a source specific vocabulary set $V_s = \{w_1^s, \ldots, w_{v_s}^s\}$. Similarly, the words appearing in the target language documents but not in M can be put together to form a target specific vocabulary set $V_t = \{w_1^t, \ldots, w_{v_t}^t\}$. An overall cross-lingual vocabulary set can then be constructed as $V = V_s \cup V_t \cup M$, which has a total of $v = v_s + v_t + m$ entries. This cross-lingual vocabulary set covers all words appearing in both domains, while mapping each bilingual pair in M into the same entry.

To tackle cross language text classification, we then propose a cross-lingual log-bilinear document model to learn a predictive cross-lingual representation of words, which maps each entry in the vocabulary set V to one row vector in a word embedding matrix $R \in \mathbb{R}^{v \times k}$. Similar to the log-bilinear language model (Mnih and Hinton, 2007) and the log-bilinear document model (Maas et al., 2011), our proposed model learns a dense feature vector for each word to capture semantic similarities between the vocabulary entries. But unlike the previous two models which only work with a monolingual language, our model also captures semantic similarities across different languages. Moreover, we explicitly incorporate the label information into our proposed approach, rendering the induced word embeddings more discriminative to the target prediction task.

3.1 Cross-Lingual Word Embeddings

As mentioned above, we assume a unified embedding matrix R which contains the distributed vector representations of words in the two language domains. However, even in a unified representation space, the distribution of words in the two domains will be different. To capture the distribution divergence of the two domains and facilitate cross-lingual learning, we split the word embedding matrix into three parts: source language specific part $R_s \in \mathbb{R}^{v \times k_s}$, common part $R_c \in \mathbb{R}^{v \times k_c}$ and target language specific part $R_t \in \mathbb{R}^{v \times k_t}$, such that $k = k_s + k_c + k_t$. Intuitively, we assume that source language words contain no target language specific representations and target language words contain no source language specific representations. Thus for words in the two language domains, we retrieve their distributed vector representations from the embedding matrix R using two mapping functions, $\Phi_{\mathcal{S}}$ and $\Phi_{\mathcal{T}}$, one for each language domain. The two mapping functions are defined as

$$\Phi_{\mathcal{S}}(w) = [R_s(w), R_c(w), \mathbf{0}_t]^T \tag{1}$$

$$\Phi_{\mathcal{T}}(w) = [\mathbf{0}_s, R_c(w), R_t(w)]^T \tag{2}$$

where $\mathbf{0}_t$ is a k_t-dimensional row vector of zeros, $\mathbf{0}_s$ is a k_s-dimensional row vector of zeros, $R_s(w)$ denotes the row vector of R_s matrix corresponding to the word w, $R_c(w)$ denotes the row vector of R_c matrix corresponding to the word w, and $R_t(w)$ denotes the row vector of R_t matrix corresponding to the word w. It is easy to see that each pair of words in M will share the same vector from R_c. To encode more information into the common part of representation for better knowledge transfer from the source language domain to the target language domain, we assume $k_c \geq k_s$ and $k_c \geq k_t$. The form of three part feature representations has been exploited in previous work of domain adaptation with heterogeneous feature spaces (Duan et al., 2012). However, their approach simply duplicates the original features as language-specific representations, while we will automatically learn those three part latent representations in our approach.

3.2 Semi-Supervised Cross-Lingual Representation Learning

Given the word representation scheme above, we conduct cross-lingual representation learning by simultaneously maximizing the log-likelihood of all documents and the conditional likelihood of labeled documents from the two language domains

$$\max_{\theta} \sum_{\mathcal{L} \in \{\mathcal{S},\mathcal{T}\}} \sum_{\mathbf{x}_i \in \mathcal{L}} \sum_{j=1}^{N_i} \log P_{\mathcal{L}}(w_{ij}|\theta) + \alpha \sum_{\mathcal{L} \in \{\mathcal{S},\mathcal{T}\}} \sum_{\mathbf{x}_i^{\ell} \in \mathcal{L}} \log P_{\mathcal{L}}(y_i|\mathbf{x}_i^{\ell}, \theta) \tag{3}$$

where θ denotes the model parameters and α is a trade-off parameter. The first part of the objective function captures the likelihood of the documents being generated with the learned representation R. $P_{\mathcal{L}}(w_{ij}|\theta)$ is the probability of word w_{ij} appearing in the document $\mathbf{x}_i$ from the language domain $\mathcal{L}$, and is defined as

$$P_{\mathcal{L}}(w_{ij}|\theta) = \frac{\exp\left(-E_{\mathcal{L}}(w_{ij}, \theta)\right)}{\sum_{w' \in V} \exp\left(-E_{\mathcal{L}}(w', \theta)\right)} \tag{4}$$

The term $E_{\mathcal{L}}(w_{ij}, \theta)$ is a log-bilinear energy function, defined as

$$E_{\mathcal{L}}(w_{ij}, \theta) = -\mathbf{d}_i^T \Phi_{\mathcal{L}}(w_{ij}) - b_{w_{ij}} \tag{5}$$

where $\mathbf{d}_i$ is a k-dimensional weight vector for document $\mathbf{x}_i$ and $b_{w_{ij}}$ is the bias for word w_{ij}. Below we will use $\mathbf{b}$ to denote a v-dimensional vector containing all words' biases.

The second part of the objective function in (3) takes the label information into account and aims to render the latent word representations more task-predictive. We use a logistic regression model to compute the conditional probability of the class label given the document with the induced word representations, such that

$$P_{\mathcal{L}}(y_i|\mathbf{x}_i^{\ell}, \theta) = \frac{1}{1 + \exp\left(-y_i\left(\mathbf{w}^T\Psi_{\mathcal{L}}(\mathbf{x}_i^{\ell}) + q\right)\right)} \tag{6}$$

where $\mathbf{w}, q$ are model parameters of the logistic regression model, $\Psi_{\mathcal{L}}(\mathbf{x}_i)$ is the k-dimensional vector representation of the document $\mathbf{x}_i$ in the language domain $\mathcal{L}$. We compute $\Psi_{\mathcal{L}}(\mathbf{x}_i)$ by taking average over all words in the document $\mathbf{x}_i$ such as

$$\Psi_{\mathcal{L}}(\mathbf{x}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\Phi_{\mathcal{L}}(w_{ij}) \tag{7}$$

By summing over all descriptions above, we can see that the proposed semi-supervised representation learning has a set of model parameters, $\theta = \{R, \{\mathbf{d}_i\}, \mathbf{b}, \mathbf{w}, q\}$. In order to avoid overfitting, we add regularization terms for the parameters $R, \{\mathbf{d}_i\}$ and $\mathbf{w}$, which leads to the final optimization problem below

$$\begin{aligned}\max_{\theta} &\sum_{\mathcal{L}\in\{\mathcal{S},\mathcal{T}\}}\sum_{\mathbf{x}_i\in\mathcal{L}}\Big(\sum_{j=1}^{N_i}\log P_{\mathcal{L}}(w_{ij}|\theta) - \gamma\|\mathbf{d}_i\|_2^2\Big)\\ &+ \alpha\sum_{\mathcal{L}\in\{\mathcal{S},\mathcal{T}\}}\sum_{\mathbf{x}_i^{\ell}\in\mathcal{L}}\log P_{\mathcal{L}}(y_i|\mathbf{x}_i^{\ell}, \theta)\\ &- \beta\|R\|_F^2 - \eta\|\mathbf{w}\|_2^2\end{aligned} \tag{8}$$

where β, γ, η are trade-off parameters, $\|\cdot\|_F$ denote the Frobenius norm and $\|\cdot\|_2$ denote the Euclidean-norm. This objective function is not jointly convex in all model parameters. We develop a gradient-based iterative optimization procedure to seek a local optimal solution. We first randomly initialize the model parameters $\{\mathbf{d}_i\}, R, \mathbf{w}$ and set $\mathbf{b}$ and q to zeros. Then we iteratively make gradient-based updates over the model parameters until reach a local optimal solution.

3.3 Cross-Lingual Document Classification

After solving (8), we obtain a word embedding matrix R. The distributed vector representation of any given document can then be computed using Eq. (7) based on Eq. (1) or Eq. (2). Under the distributed vector representations of the documents in both language domains, we perform cross-lingual document classification by training a supervised classification model using labeled data from both language domains and then applying it to classify test documents in the target language domain .

4 Experiments

We empirically evaluate the proposed approach using the cross language sentiment classification tasks of Amazon product reviews in four languages. In this section, we report our experimental results.

4.1 Dataset

We used the multilingual sentiment classification dataset[1] provided by Prettenhofer and Stein (2010), which contains Amazon product reviews in four different languages, English (E), French (F), German (G) and Japanese (J). The English product reviews were sampled from previous cross-domain sentiment classification datasets (Blitzer et al., 2007), while the other three language product reviews were crawled from Amazon by the authors in November 2009. In the dataset, each language contains three categories of product reviews, Books (B), DVD (D) and Music (M). Each language-category pair contains a balanced training set and test set, each of which consists of 1000 positive reviews and 1000 negative reviews. Each review is represented as a unigram bag-of-word feature vector with term-frequency values. Following the work (Prettenhofer and Stein, 2010), we used the original English reviews as the source language while treating the other three languages as target languages. Thus, we construct nine cross language sentiment classification tasks (GB, GD, GM, FB, FD, FM, JB, JD, JM), one for each target language-category pair. For example, the task *GB* means that the target language is *German* and the training and test data are samples from *Books* reviews.

4.2 Approaches

We compare our proposed semi-supervised cross-lingual representation learning (**CL-RL**) approach to the following approaches for cross-lingual document classification.

[1] http://www.webis.de/research/corpora/

Table 1: Average classification accuracies and standard deviations for the 9 cross-lingual sentiment classification tasks. The bold format indicates that the difference between the results of *CL-RL* and *MT* is significant with $p < 0.05$ under a McNemar paired test for labeling disagreements.

Task	TB	CL-Dict	CLD-LSA	CL-SCL	MT	CL-RL
GB	66.25±0.64	69.40±0.61	70.30±0.44	73.78±0.32	78.05±0.64	**79.89**±0.30
GD	63.16±0.66	66.37±0.63	66.85±0.46	71.99±0.25	75.75±0.58	**77.14**±0.16
GM	65.42±0.77	68.81±0.51	68.93±0.58	71.58±0.35	74.85±0.62	**77.27**±0.16
FB	65.98±0.51	69.35±0.48	69.98±0.51	73.89±0.16	78.00±0.49	78.25±0.32
FD	63.76±0.37	67.96±0.60	68.88±0.43	73.79±0.28	75.75±0.71	74.83±0.30
FM	65.94±0.56	67.98±0.69	68.42±0.60	71.20±0.28	74.85±0.49	**78.71**±0.32
JB	63.86±0.80	59.40±0.29	62.62±0.62	62.49±0.23	67.20±0.80	**71.11**±0.21
JD	63.59±0.74	62.13±0.26	63.87±0.72	65.54±0.29	67.70±0.57	**73.12**±0.23
JM	65.84±0.90	63.01±0.46	65.67±0.72	65.49±0.36	68.30±0.61	**74.38**±0.40

- **TB:** This is a target baseline method, which trains a supervised monolingual classifier on the labeled training data from the target language domain without representation learning.

- **CL-Dict:** This is a simple baseline comparison method, which uses the bilingual word pairs directly to align features from different language domains into a unified feature dictionary and then trains a supervised classifier on this aligned feature space with labeled training data from both language domains.

- **CLD-LSA:** This is the cross-lingual representation learning method developed in (Gliozzo, 2006), which first translates each document from one language into the other language via a bilingual dictionary to produce augmenting features, and then performs latent semantic analysis (LSA) over the augmented bilingual document-term matrix.

- **CL-SCL:** This is the cross language structural correspondence learning method developed in (Prettenhofer and Stein, 2010).

- **MT:** This is a machine translation based comparison method, which first uses an existing machine translation tool (google translation) to translate the target language documents into the source language and then trains a monolingual classifier with labeled training data from both domains in the source language.

In all experiments, we used a linear support vector machine (SVM) for sentiment classification. For implementation, we used the liblinear package (Fan et al., 2008) with all of its default parameters. For the *CL-SCL* method, we used the same parameter setting as suggested in the paper (Prettenhofer and Stein, 2010): the number of pivot features is set as 450, the threshold value for selecting pivot features is 30, and the reduced dimensionality after singular value decomposition is 100. For the *CLD-LSA* method, we set the dimensionality of latent representation as 1000. Similarly, for our proposed approach, we built the cross-lingual vocabulary M by setting $m = 450$ and $\phi = 30$. For our representation learning, we set $\alpha = 1$, $\beta = \gamma = \eta = 1e^{-4}$, and set k_s, k_c, k_t to be 25, 50, 25, respectively. The values of α, β, γ and η are selected using the first cross language classification task GB. We selected the α value from the set $\{0.01, 0.1, 1, 10, 100\}$ and selected β, γ, η values from the set $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-5}\}$ by repeating the experiment three times with random data partitions and choosing the parameter values that led to the best average classification accuracy.

4.3 Classification Accuracy

For each of the nine cross language sentiment classification tasks with different target language-category pairs, we used the training set in the source language domain (English) as labeled data while treating the test set in the source language domain as unlabeled.

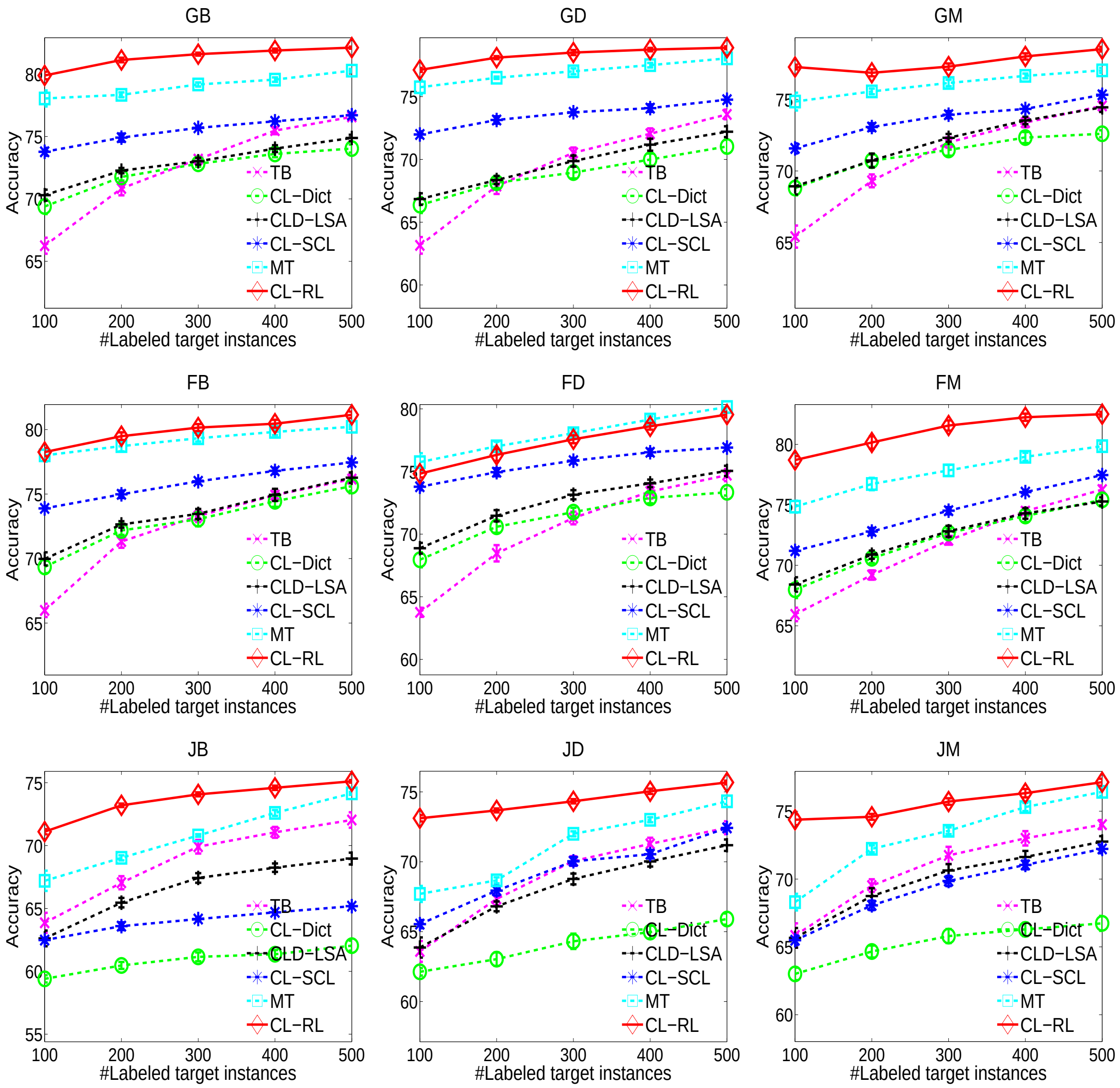

Figure 1: Average classification accuracies and standard deviations for 10 runs with respect to different numbers of labeled training documents in the target language domain.

For target language domain, we used the test set as test data while randomly selecting 100 documents from the training set as labeled data and treating the rest as unlabeled data. Thus, for each task, we have 2000 labeled documents and 2000 unlabeled documents from the source language domain, and 100 labeled and 1900 unlabeled documents from the target language domain for training. We have 2000 test documents from the target language domain as testing data. In each experiment, a classifier is produced by each approach with the training data and tested on the testing data. We repeated each experiment 10 times with different random selections of 100 labeled training documents from the target language domain. The average classification accuracies and standard deviations are reported in Table 1.

From Table 1, we can see that the proposed semi-supervised cross-lingual representation learning approach, CL-RL, clearly outperforms all other comparison methods on eight out of the nine tasks. The target baseline *TB* performs poorly on all the nine tasks, which suggests that 100 labeled instances from the target language is far from enough to obtain an accurate sentiment classifier in the target lan-

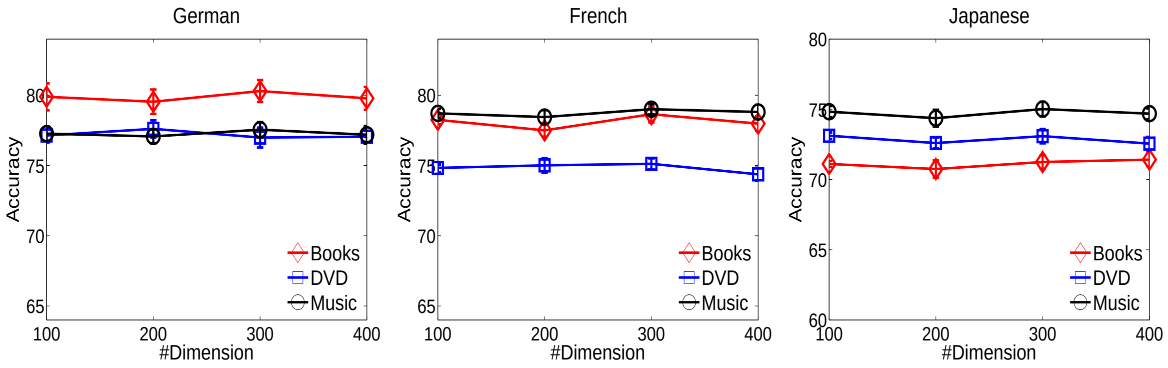

Figure 2: Average classification accuracy and standard deviation results for the proposed approach over 10 runs with respect to different dimensionality for the induced cross-lingual representations.

guage domain. By exploiting the large amount of labeled training data from the source language domain, even the simple cross-lingual adaptation approach, *CL-Dict*, produces effective improvements over *TB*. However, its performance is not consistent across the nine tasks. It has inferior performance than *TB* on the three tasks of adapting English to the Japanese language domain. This suggests the simple bilingual word-pair based feature space unification method is far from ideal for providing effective cross-lingual representations, especially when two languages (English, Japanese) are very different. With a better designed representation learning, *CLD-LSA* outperforms *CL-Dict* on all the nine tasks, but the improvements are very small on some tasks (e.g., GM). *CL-SCL* not only outperforms *CL-Dict* on all tasks, but also performs much better than *CLD-LSA* on most tasks. Its performance nevertheless is inferior to the method of *MT*. Though *MT* can greatly increase the test accuracies comparing to the other four methods, *TB*, *CL-Dict*, *CLD-LSA*, and *CL-SCL*, the benefit is obtained at the cost of whole document translations. In contrast, our proposed approach does not require whole document translations, but relies on the same simple word-pair translations used in *CL-Dict*. It however consistently and significantly outperforms *TB*, *CL-Dict*, *CLD-LSA*, and *CL-SCL* on all tasks, and outperforms *MT* on eight out of the nine tasks.

We also conduct significance tests for our proposed approach and *MT* using a McNemar paired test for labeling disagreements (Gillick and Cox, 1989). The results in bold format indicate that they are significant with $p < 0.05$. All these results demonstrate the efficacy of our cross-lingual representation learning method.

4.4 Classification Accuracy vs the Number of Labeled Target Documents

Next, we investigated the performance of the six approaches by varying the number of labeled training documents from the target language domain. We maintained the same experimental setting as before, but investigated a range of different values, $\ell_t = \{100, 200, 300, 400, 500\}$, as the number of labeled training documents from the target language domain. In each experiment, for a given value ℓ_t, we randomly selected ℓ_t documents from the training set of the target language domain as labeled data and used the rest as unlabeled data. We still performed prediction on the same 2000 test documents in the target language domain. We repeated each experiment 10 times based on different random selections of the labeled training data from the target language domain. The average classification accuracies and standard deviations across different ℓ_t values for all comparison methods on all the nine tasks are plotted in Figure 1.

We can see when the number of labeled target documents is small, *TB* performs poorly, especially for the first six tasks (GB, GD, GM, FB, FD, FM). By increasing the size of labeled target training data, *TB* can greatly increase its prediction accuracies and even outperform the *CL-Dict* method. The simple *CL-Dict* method has inconsistent performance across the nine tasks. Its performance is better than

TB when the labeled training data in the target language domain is very limited and is poor than *TB* when the labeled target data reaches 300 for the six tasks using German and French as target languages. Moreover, when adapting a system from English to a much more different target language (Japanese), *CL-Dict* produces much lower accuracies for all the three tasks comparing with *TB*. These results show that *CL-Dict* has very limited capacity on transferring labeled information from a related source language domain. Similar performance is observed for *CLD-LSA*. With a more sophisticated representation learning, the *CL-SCL* method consistently outperforms *CL-Dict*. However, it produces inferior performance than *CLD-LSA* on the tasks of *JB* and *JM*. By using more translation resources, the *MT* method outperforms *TB*, *CL-Dict*, *CLD-LSA*, *CL-SCL* in all the nine tasks across almost all scenarios. Our proposed method *CL-RL* significantly outperforms all the other five comparison methods across all experiments except on the task of FD, where *MT* produces similar performance. Moreover, it is especially important to notice that *CL-RL* achieves high test accuracies even when the number of labeled target instances is small. This is important for transferring knowledge from a source language to reduce the labeling effort in the target language.

4.5 Sensitivity Analysis

We also investigated the sensitivity of the proposed approach over the dimensionality of the induced cross-lingual representations. We used the same experimental setting as before, and conducted experiments with a set of different dimensionality values, $k = \{100, 200, 300, 400\}$. For each value k, we set $k_s = 0.25k, k_c = 0.5k, k_t = 0.25k$. We repeated each experiment for 10 times based on different random selections of labeled target training data and plotted the average prediction accuracies and standard deviations in Figure 2 for all the nine cross-lingual sentiment classification tasks. We can see the proposed approach produces stable accuracy results across the range of different k values. This suggests the proposed approach is not very sensitive to the dimensionality of the cross-lingual embedding features within the considered range of values, and with a small dimensionality of 100, the induced representation can already perform very well.

4.6 Cross-Lingual Word Representations

Finally, we used the first task *GB*, which adapts the *Books* reviews from English to German, to gain intuitive understandings over the learned cross-lingual word representations. Given an English word as seed word, we find its five closest neighboring English words and German words according to the Euclidean distances calculated in the induced cross-lingual representation space. We present a few results in Table 2. From Table 2, we can see that the retrieved words in both language domains are semantically close to the seed words, which indicates that our proposed method can capture semantic similarities of words not only in a monolingual setting but also in a multilingual setting.

5 Conclusion

In this paper, we proposed a semi-supervised cross-lingual representation learning approach to address cross-lingual text classification. The distributed word representation induced by the proposed approach can capture semantic similarities of words across languages while maintaining predictive information with respect to the target classification tasks. To evaluate the proposed approach, we conducted experiments on nine cross language sentiment classification tasks constructed from the Amazon product reviews in four languages, comparing to a number of comparison methods. The empirical results showed that the proposed approach can produce effective cross-lingual adaptation performance and significantly outperform other comparison methods.

References

M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

B. A.R., A. Joshi, and P. Bhattacharyya. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2012.

N. Bel, C. Koster, and M. Villegas. Cross-lingual

Table 2: Examples of source seed words together with five closest English words and five closest German words estimated using the Euclidean distance in the cross-lingual representation space on the task *GB*.

books		**absolutely**		**love**	
English	German	English	German	English	German
books	buch	absolutely	absolut	love	liebe
book	bücher	definitely	absolute	loved	lieben
text	text	completely	definitiv	like	wie
page	blatt	certainly	komplett	fond	wieder
words	wörter	totally	sicher	feel	fühlen

expensive		**good**		**not**	
English	German	English	German	English	German
expensive	teuer	good	gut	not	nicht
expense	höher	better	besser	no	nie
overpriced	höchsten	well	nett	cannot	nein
costly	hoch	nice	großartig	non	keine
price	preis	great	größten	never	keines

text categorization. In *Proceedings of European Conference on Digital Libraries (ECDL)*, 2003.

Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Asso. for Computational Linguistics (ACL)*, 2007.

K. Diamantaras and S. Kung. *Principal component neural networks: theory and applications*. Wiley-Interscience, 1996.

L. Duan, D. Xu, and I. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research (JMLR)*, 9:1871–1874, 2008.

C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989.

A. Gliozzo. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics (ICCL-ACL)*, 2006.

Y. Guo and M. Xiao. Transductive representation learning for cross-lingual text classification. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2012a.

Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012b.

T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 1999.

A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words.

In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2012.

X. Ling, G. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? In *Proceedings of the International Conference on World Wide Web (WWW)*, 2008.

M. Littman, S. Dumais, and T. Landauer. *Automatic Cross-Language Information Retrieval using Latent Semantic Indexing*, chapter 5, pages 51–62. Kluwer Academic Publishers, 1998.

A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, 2011.

D. Mimno, H. Wallach, J. Naradowsky, D. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, 2009.

A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

X. Ni, J. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.

J. Pan, G. Xue, Y. Yu, and Y. Wang. Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *Proceedings of the Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD)*, 2011.

P. Petrenz and B. Webber. Label propagation for fine-grained cross-lingual genre classification. In *Proceedings of the NIPS xLiTe workshop*, 2012.

S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012.

J. Platt, K. Toutanova, and W. Yih. Translingual document representations from discriminative projections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

L. Rigutini and M. Maggini. An em based training algorithm for cross-language text categorization. In *Proceedings of the Web Intelligence Conference*, 2005.

J. Shanahan, G. Grefenstette, Y. Qu, and D. Evans. Mining multilingual opinions through classification and translation. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

W. Smet, J. Tang, and M. Moens. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD)*, 2011.

A. Vinokourov, J. Shawe-taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

C. Wan, R. Pan, and J. Li. Bi-weighting domain adaptation for cross-language text classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009.

B. Wei and C. Pal. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the Annual Meeting of the Asso. for Computational Linguistics (ACL)*, 2010.