Semi-Supervised Matrix Completion 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 text classification is the task of assigning labels to observed documents in a label-scarce target language domain by using a prediction model trained with labeled documents from a label-rich source language domain. Cross-lingual text classification is popularly studied in natural language processing area to reduce the expensive manual annotation effort required in the target language domain. In this work, we propose a novel semi-supervised representation learning approach to address this challenging task by inducing interlingual features via semi-supervised matrix completion. To evaluate the proposed learning technique, we conduct extensive experiments on eighteen cross language sentiment classification tasks with four different languages. The empirical results demonstrate the efficacy of the proposed approach, and show it outperforms a number of related cross-lingual learning methods.

Introduction

With the rapid development of linguistic resources in multiple languages, it is very important to develop cross-lingual natural language processing (NLP) systems. As a popular task in the NLP community, cross-lingual text classification (CLTC) aims to classify documents in a *target* language domain where there is not enough labeled training data by employing a classification model trained with a large amount of annotated documents in a *source* language domain (Bel, Koster, and Villegas 2003; Shanahan et al. 2004; Amini, Usunier, and Goutte 2009). CLTC is of great importance in the NLP area as it can exploit the existing labeled information in a source language and hence substantially reducing the expensive and time-consuming labeling effort in the target language. A variety of work in the literature has demonstrated the success of various NLP applications in the multilingual scenarios such as cross language document categorization (Wan, Pan, and Li 2011; Dai et al. 2008; Duan, Xu, and Tsang 2012; Ling et al. 2008), multilingual fine-grained genre classification (Petrenz and Webber 2012), and cross-lingual sentiment classification (Prettenhofer and Stein 2010; Xiao and Guo 2013).

One fundamental challenge for CLTC is the difference of the feature representations in the two languages due to the fact that different languages have very different word vocabularies. If we train a standard monolingual classification model with labeled documents in the source language on the original word feature representation space, it will fail to apply in the target language. Recently, cross-lingual representation learning methods have been proposed to address this challenging task by automatically learning a set of language-independent features to bridge the language barrier (Platt, Toutanova, and Yih 2010; Xiao and Guo 2013). Those methods though have demonstrated good empirical cross-lingual adaptation performance, they are conducted in a fully unsupervised fashion without using the existing valuable label information, which limits their learning capacity.

In this paper, we propose a novel semi-supervised representation learning method to address CLTC, which exploits a small set of unlabeled parallel documents and induces cross-lingual features via semi-supervised matrix completion. We first unify the word vocabularies of the two language domains and construct a dual-language document-term matrix for the documents in both languages, where parallel documents are represented as vectors filled with word features from the original documents and their translations, while non-parallel document vectors only contain word features from one language. Since we only have a small set of parallel documents, this partially filled document-term matrix contains a large portion of missing entries. We then perform matrix completion to automatically fill those missing values and recover a complete dual-language document-term matrix. To make the recovered matrix more informative for the target classification task, we incorporate the existing label information into the matrix completion process in a semi-supervised manner. Finally we perform latent semantic indexing over the completed dual-language document-term matrix to produce robust low dimensional interlingual document representations for cross-lingual classification. We empirically evaluate the proposed method on eighteen cross-lingual sentiment classification tasks constructed from Amazon product reviews with four different languages. The experimental results demonstrate the effectiveness of the proposed semi-supervised matrix completion method in learning better cross-lingual representations, comparing to a number of related cross-lingual learning methods.

Related Work

A variety of representation learning approaches have been developed in the literature to address CLTC, which induce interlingual representations by exploiting parallel documents (Littman, Dumais, and Landauer 1998; Vinokourov, Shawe-taylor, and Cristianini 2002; Platt, Toutanova, and Yih 2010; Klementiev, Titov, and Bhattarai 2012; Xiao and Guo 2013), or some other multilingual resources (Gliozzo 2006; Prettenhofer and Stein 2010; Wei and Pal 2010; Wu, Wang, and Lu 2008; Petrenz and Webber 2012). Parallel documents are usually obtained using machine translation and used to bridge two different languages. (Littman, Dumais, and Landauer 1998) proposed a cross-lingual latent semantic indexing (CL-LSI) method to extract interlingual representations. They first constructed a bilingual document-term matrix on the parallel data, where each parallel document is represented as a dual-language document vector containing word features from the original document and the corresponding translated document. Then they performed standard latent semantic indexing operation over the dual-language document-term matrix to learn low dimensional representations. Similarly, (Platt, Toutanova, and Yih 2010) employed the cross-lingual oriented principal component analysis (CL-OPCA) method over the parallel documents to induce language-independent features. (Vinokourov, Shawe-taylor, and Cristianini 2002) developed a cross-lingual kernel canonical component analysis (CL-KCCA) method to learn two language-specific projections from parallel data, and then used them to project documents from language-specific representation spaces to the language-independent feature space. Recently, (Xiao and Guo 2013) proposed a two-step learning (TSL) method to induce cross language representations by first performing matrix completion over the partially filled dual-language document-term matrix and then conducting latent semantic indexing over the completed matrix to discover interlingual features.

Bilingual dictionaries have also been employed to build connections between the source and target languages. (Gliozzo 2006) first used the bilingual dictionary to translate all words in a document from one language into the other language and used these translated words to augment the original document. Then they conducted latent semantic analysis over the augmented document-term matrix to discover language-independent features. (Prettenhofer and Stein 2010) proposed a cross language structural correspondence learning method to induce interlingual representations by using a large amount of unlabeled documents from the two languages in aid of some pivot bilingual-word pairs. They induced cross-lingual representation by modeling the correlations between th pivot features and non-pivot features. Other resources have also been used in a few works to address cross language text classification. For example, the universal part-of-speech tags (Petrov, Das, and McDonald 2012) are used in (Petrenz and Webber 2012), the Wikipedia is used in (Ni et al. 2011), and the multi-lingual WordNet is used in (Gliozzo 2006; A.R., Joshi, and Bhattacharyya 2012). These resources though can bridge different languages, they are not always available for any language pairs. Hence we do not evaluate them in our empirical studies.

Our proposed approach shares similarity with those representation learning methods on exploiting parallel documents. However, unlike the CL-CCA, CL-OPCA and CL-KCCA methods, which work on the pre-given dual-language document-term matrix, both our approach and the TSL method first automatically complete the partially observed bilingual document-term matrix and then learn multilingual semantic features. Moreover, the proposed approach differs from CL-LSI, CL-OPCA, CL-KCCA, and TSL by exploiting existing label information to perform semi-supervised representation learning.

Approach

In this section, we present a semi-supervised matrix completion method to induce cross-lingual representations between a source language domain and a target language domain. This method exploits a small set of unlabeled parallel documents to build connections across language vocabularies, and then automatically fills the missing features for all documents from both language domains in a unified bilingual document-term matrix. Moreover, this approach exploits the existing labels from both domains to produce label informative completion over the bilingual document-term matrix.

Formulation

Let M^0 be the $t \times d$ bilingual document-term matrix built over all the documents in the two language domains, which is partially filled with observed feature values, where t is the number of documents and d is the size of the unified bilingual vocabulary. In this matrix, each pair of parallel documents is represented as a fully observed row vector, and each non-parallel document is represented as a partially observed row vector where only entries corresponding to words in its own language vocabulary are observed. We use Ω to denote the index set of the observed features in M^0, such as $(i, j) \in \Omega$ if only if M^0_{ij} is observed.

Given the partially observed bilingual document-term matrix M^0, we seek to automatically fill the missing observations to recover a fully observed matrix M. M has three properties: (1) it is sparse since each document typically only contains a very small set of words from the big vocabulary; (2) it is low-rank since there are correlations between the word features; and (3) it is non-negative since document-term matrix usually has non-negative features, e.g., term frequency features or tf-idf features. Moreover, to complete the document-term matrix in a discriminative manner, and hence produce a label informative feature representation, we integrate the matrix completion and an implicit linear classification model training together. For simplicity of presentation, we assume the first t_ℓ documents are the labeled documents from both language domains with a label vector $\mathbf{y} \in \{-1, 1\}^{t_\ell}$. Then the joint optimization problem for semi-supervised matrix completion is formulated as

$$\min_{M \geq 0,\, \mathbf{z}, b} \quad \gamma \|[M, \mathbf{z}]\|_* + \mu \|M\|_{1,1} + \sum_{(i,j) \in \Omega} c(M_{ij}, M^0_{ij}) + \beta \sum_{i=1}^{t_\ell} c(\mathbf{z}_i + b, \mathbf{y}_i) \quad (1)$$

Algorithm 1 Algorithm

Input: $M^0,\ \gamma > 0,\ \beta \geq 1,\ 0 < \tau < \min(2, \frac{2}{\beta}),\ \mu$
Initialize M as the nonnegative projection of the rank-1 approximation of M^0; initialize $\mathbf{z}$ as zeros.
while not converged **do**
1. gradient descent: $[M, \mathbf{z}] = [M, \mathbf{z}] - \tau \nabla g(M, \mathbf{z})$.
2. shrinkage operation: $[M, \mathbf{z}] = \mathcal{S}_{\tau\gamma}([M, \mathbf{z}])$.
3. project M onto the feasible set: $M = \max(M, 0)$.
end while

where $[M, \mathbf{z}]$ is a combination matrix of M and $\mathbf{z} \in \mathbb{R}^t$; the cost function $c(x, y) = \frac{1}{2}(x - y)^2$; $\|\cdot\|_*$ denotes the matrix trace norm and $\|\cdot\|_{1,1}$ denotes the entrywise L1 norm. With nonnegativity constraints, the entrywise L1 norm becomes $\|M\|_{1,1} = \sum_{i,j} M_{ij}$. The $\mathbf{z}$ vector denotes a latent output column that captures the linear mapping results over the feature matrix M, e.g., $\mathbf{z} = M\mathbf{w}$; hence $\sum_i c(\mathbf{z}_i + b, \mathbf{y}_i)$ denotes the prediction loss over the labeled data, where b is a bias parameter. By enforcing the low-rank property of the combination matrix via trace norm, the implicit linear function $\mathbf{z} = M\mathbf{w}$ will be enhanced. Missing features are expected to be filled to facilitate such a linear prediction function, while being consistent with the observed data in M^0.

Let $\mathbf{z}^\ell$ be the subvector of $\mathbf{z}$ that contains the first t_ℓ entries, corresponding to the t_ℓ labeled documents. Let $A = [I_{t_\ell}, O_{t_\ell, t-t_\ell}]$, where I_{t_ℓ} denotes an identity matrix with size t_ℓ, and $O_{t_\ell, t-t_\ell}$ is a $t_\ell \times (t - t_\ell)$ matrix with all zeros. Then we have $\mathbf{z}^\ell = A\mathbf{z}$. The minimization over b in problem (1) has a closed-form solution

$$b = \frac{1}{t_\ell} \mathbf{1}^\top (\mathbf{y} - A\mathbf{z}) \tag{2}$$

where $\mathbf{1}$ denotes any column vector with all 1s. Let $H = I_{t_\ell} - \frac{1}{t_\ell}\mathbf{1}\mathbf{1}^\top$ be a centering matrix. By plugging (2) back into (1), we obtain the following equivalent problem

$$\min_{M \geq 0,\ \mathbf{z}} \ \gamma \|[M, \mathbf{z}]\|_* + \mu \|M\|_{1,1} + \sum_{(i,j) \in \Omega} c(M_{ij}, M^0_{ij}) + \beta c(HA\mathbf{z}, H\mathbf{y}) \tag{3}$$

Optimization Algorithm

The optimization problem (3) is convex but non-smooth. We treat the optimization objective function in (3) as a combination function $f(M, \mathbf{z})$ over a smooth function $g(\cdot)$ and a non-smooth trace norm, such as

$$f(M, \mathbf{z}) = \gamma \|[M, \mathbf{z}]\|_* + g(M, \mathbf{z}) \tag{4}$$

$$g(M, \mathbf{z}) = \mu \|M\|_{1,1} + \sum_{(i,j) \in \Omega} c(M_{ij}, M^0_{ij}) + \beta c(HA\mathbf{z}, H\mathbf{y}) \tag{5}$$

and then develop a projected gradient descent algorithm to solve it. The algorithm is given in Algorithm 1.

In this algorithm, we iteratively update the model parameters using projected gradient descent. In each iteration, we perform three steps: gradient descent, shrinkage operation, and projection onto the feasible set. The gradient should be computed over the smooth function $g(\cdot, \cdot)$ in (5). Let $Y \in \{0, 1\}^{t \times d}$ such that $Y_{ij} = 1$ if $(i, j) \in \Omega$ and $Y_{ij} = 0$ otherwise. Let E be a $t \times d$ matrix with all 1s. The gradient function is $\nabla g(M, \mathbf{z}) = [\nabla_M g(M, \mathbf{z}), \nabla_\mathbf{z} g(M, \mathbf{z})]$, where

$$\nabla_M g(M, \mathbf{z}) = \mu E + M \circ Y - M^0 \circ Y \tag{6}$$

$$\nabla_\mathbf{z} g(M, \mathbf{z}) = \beta A^\top H(A\mathbf{z} - \mathbf{y}) \tag{7}$$

and "$\circ$" denotes the Hadamard product operator. After the gradient descent, we perform a shrinkage operation to take the trace norm regularizer into account. The shrinkage operator is based on singular value decomposition. Let $Z = [M, \mathbf{z}]$ and $\nu = \tau\gamma$. Then the shrinkage operation is

$$S_\nu(Z) = U\Sigma_\nu V^\top, \tag{8}$$
$$\text{where} \quad Z = U\Sigma V^\top, \quad \Sigma_\nu = \max(\Sigma - \nu, 0).$$

Finally the updated M can be projected to the feasible nonnegative set.

Convergence Analysis Let $h(\cdot) = I(\cdot) - \tau \nabla g(\cdot)$ be the gradient descent operator used in the gradient descent step. We can verify that $h(\cdot)$ is non-expansive, i.e., $\|h(Z) - h(Z')\|_F \leq \|Z - Z'\|_F$, for $\tau \in (0, \min(2, \frac{2}{\beta}))$. Here $\|\cdot\|_F$ denotes the Frobenius norm.

According to (Ma, Goldfarb, and Chen 2011) and (Xiao and Guo 2013), the shrinkage operator and nonnegative projection operator are non-expansive as well. Hence the composite operator formed by the three steps is non-expansive, and the projected gradient updates will converge to an optimal solution $(M^*, \mathbf{z}^*)$ (Ma, Goldfarb, and Chen 2011).

Cross-lingual Representation

After obtaining the automatically filled bilingual document-term matrix M^*, we produce a low dimensional representation shared by documents from both languages by performing latent semantic indexing over M^*. That is, we first decompose $M^* = U\Sigma V$ via singular value decomposition. Then we use the top k right singular vectors V_k as a projection matrix, which produces a low dimensional representation $Z = M^* V_k$ for all the documents. We can then perform learning with this unified representation matrix.

Experiments

In this section, we report extensive empirical evaluations of the proposed semi-supervised matrix completion approach on a variety of cross language sentiment classification tasks.

Experimental Setup

Dataset We used the multilingual Amazon product review dataset in our experiments for cross-lingual sentiment classification, which contains reviews in three different categories (Books(B), DVD(D), and Music(M)), written in four different languages (English(E), French(F), German(G) and Japanese(J)). For each category of the product reviews, there are 2000 positive and 2000 negative reviews in English, and 1000 positive and 1000 negative reviews in each of the other three languages (French, German and Japanese). Moreover,

Table 1: Average classification accuracies and standard deviations for the 18 cross language sentiment classification tasks.

TASK	TBOW	CL-LSI	CL-KCCA	CL-OPCA	TSL	SSMC
EFB	67.31±0.96	79.56±0.21	77.56±0.14	76.55±0.31	81.92±0.20	**83.05±0.26**
FEB	66.82±0.43	76.66±0.34	73.45±0.13	74.43±0.53	79.51±0.21	**80.05±0.18**
EFD	67.80±0.94	77.82±0.66	78.19±0.09	70.54±0.41	81.97±0.33	**82.70±0.20**
FED	66.15±0.65	76.61±0.25	74.93±0.07	72.49±0.47	78.09±0.32	**79.40±0.28**
EFM	67.84±0.43	75.39±0.40	78.24±0.12	73.69±0.49	79.30±0.30	**80.46±0.20**
FEM	66.08±0.52	76.33±0.27	73.38±0.12	73.46±0.50	78.53±0.46	**78.82±0.17**
EGB	67.23±0.68	77.59±0.21	79.14±0.12	74.72±0.54	79.22±0.31	**81.88±0.42**
GEB	67.16±0.55	77.64±0.19	74.15±0.09	74.78±0.39	78.65±0.23	**79.06±0.23**
EGD	66.79±0.80	79.22±0.22	76.73±0.10	74.59±0.66	81.34±0.24	**82.25±0.20**
GED	66.27±0.69	77.78±0.26	74.26±0.08	74.83±0.45	79.34±0.23	**80.89±0.16**
EGM	67.65±0.45	73.81±0.49	79.18±0.05	74.45±0.59	79.39±0.39	**81.30±0.20**
GEM	66.74±0.55	77.28±0.51	72.31±0.08	74.15±0.42	79.02±0.34	**79.85±0.17**
EJB	63.15±0.69	72.68±0.35	69.46±0.11	71.41±0.48	72.57±0.52	**73.76±0.24**
JEB	66.85±0.68	74.63±0.42	67.99±0.18	73.41±0.41	77.17±0.36	**77.82±0.13**
EJD	65.47±0.50	72.55±0.28	74.79±0.11	71.84±0.41	76.60±0.49	**77.58±0.32**
JED	66.42±0.55	75.18±0.27	72.44±0.16	75.42±0.52	79.01±0.50	**79.60±0.25**
EJM	67.62±0.75	73.44±0.50	73.54±0.11	74.96±0.86	76.21±0.40	**77.53±0.25**
JEM	66.51±0.51	72.38±0.50	70.00±0.18	72.64±0.66	77.15±0.58	**77.74±0.24**

there are 2000 additional unlabeled parallel reviews between English and each of the other three languages (French, German, and Japanese) for each category. We constructed 18 cross language sentiment classification tasks (EFB, EFD, EFM, FEB, FED, FEM, EGB, EGD, EGM, GEB, GED, GEM, EJB, EJD, EJM, JEB, JED, JEM) between English and the other three languages for the three categories. For example, the task *EFB* uses the *Books(B)* reviews in *English(E)* as the source language data and the *Books(B)* reviews in *French(F)* as the target language data.

Approaches We compared the proposed semi-supervised matrix completion (*SSMC*) approach with the following methods in our experiments: (1) a target bag-of-word (*TBOW*) feature based method, which trains a sentiment classifier on the labeled data in the target language with bag-of-word features; (2) a cross-lingual latent semantic indexing (*CL-LSI*) method (Littman, Dumais, and Landauer 1998), which first learns low-dimensional interlingual representations by performing latent semantic indexing over the dual-language document-term matrix and then trains a sentiment classifier on the labeled data from the two languages; (3) a cross-lingual kernel canonical component analysis (*CL-KCCA*) method (Vinokourov, Shawe-taylor, and Cristianini 2002), which first uses the parallel data to learn two language projections and then trains a sentiment classifier on the labeled data from the two languages in the projected interlingual representation space; (4) a cross-lingual oriented principle component analysis (*CL-OPCA*) method (Platt, Toutanova, and Yih 2010), which first learns cross language representations by performing oriented principle component analysis over the dual-language document-term matrix and then trains a sentiment classifier on the labeled data from the two languages; (5) a two step learning (*TSL*) method (Xiao and Guo 2013), which learns cross-lingual representations by completing the partially filled dual-language document-term matrix with unsupervised matrix completion and performing latent semantic indexing over the completed matrix, and then trains a sentiment classifier with labeled data from the two languages.

For all the approaches, we used support vector machines (SVMs) as the base classifiers for sentiment classification. We used the LIBSVM package (Chang and Lin 2011) with linear kernels and default parameter setting.

Classification Accuracy

For each of the eighteen cross language sentiment classification tasks, in addition to the 2000 unlabeled parallel reviews which we used only for representation learning, we used all the documents in the source language as labeled data (4000 English reviews or 2000 non-English reviews) and randomly chose 100 reviews in the target language as labeled data while keeping the rest reviews in the target language as unlabeled data. We used all the data to learn cross language representations and then trained a sentiment classifier on the labeled documents from the two languages and applied it to classify the remaining unlabeled target reviews. We conducted parameter selection based on three runs over the first task *EFB* with different random selections of the 100 labeled training data in the target language. For *SSMC*, we chose γ from $\{0.01, 0.1, 1, 10, 100\}$, β from $\{1, 2, 5, 10, 100\}$, μ from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and chose the reduced dimension size k from $\{20, 50, 100, 200, 500\}$. This leads to the following setting: $\gamma = 10$, $\beta = 1$, $\mu = 10^{-4}$, $k = 50$. We used $\tau = 1$. For *TSL*, we set $\mu = 10^{-6}$, $\tau = 1$, and chose γ from $\{0.01, 0.1, 1, 10, 100\}$, ρ from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and the reduced dimension size k from $\{20, 50, 100, 200, 500\}$. This leads to the setting $\gamma = 0.1$, $\rho = 10^{-4}$, and $k = 50$. We used the same procedure to select the reduced dimension size k for the other three methods, which leads to $k = 50$ for *CL-LSI*

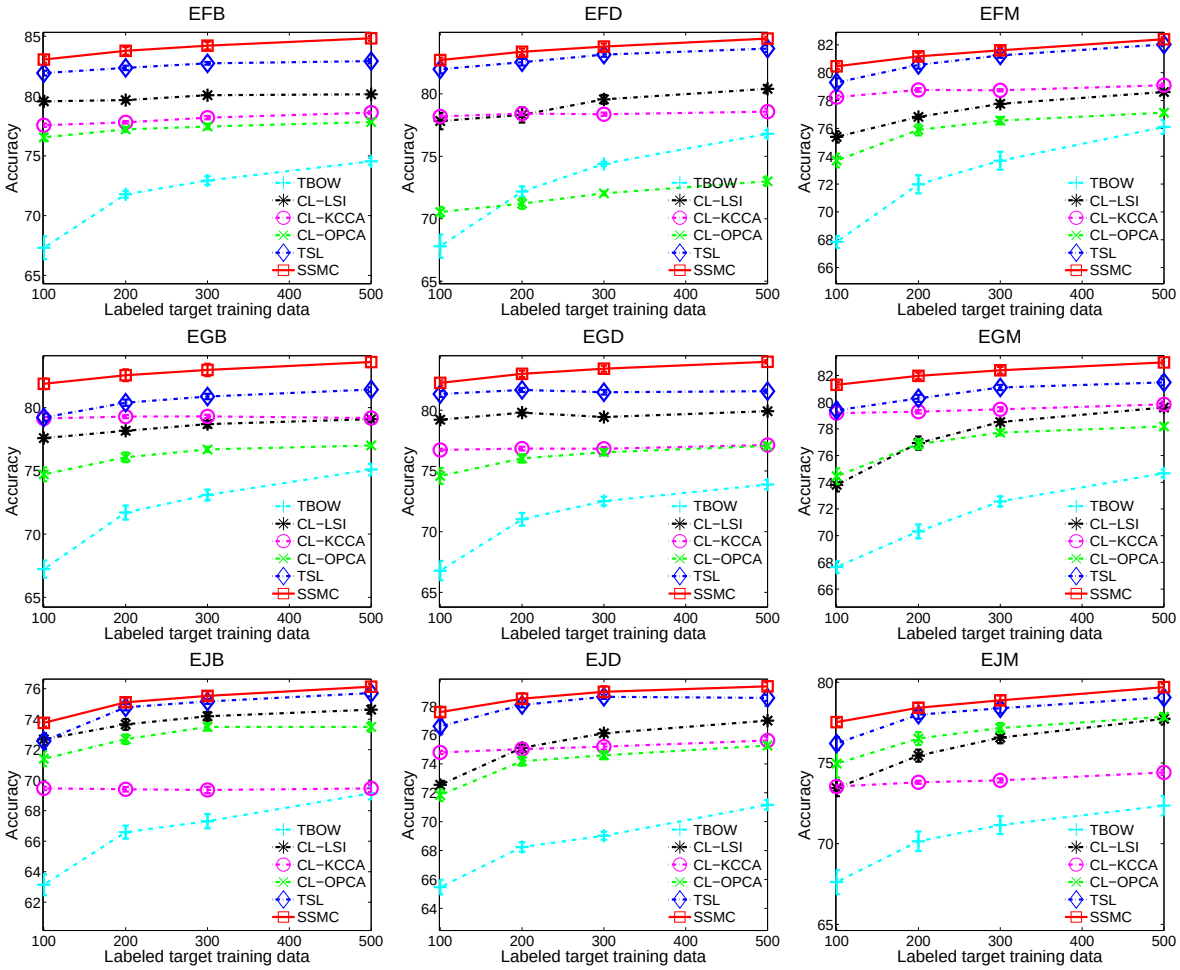

Figure 1: Average classification accuracies and standard deviations with respect to different numbers of labeled training documents in the target language for the nine CLTC tasks with English as the source language.

and *CL-OPCA* and $k = 100$ for *CL-KCCA*. After choosing the parameter settings, we run each task 10 times based on different random selections of the 100 labeled training data in the target language. The average classification accuracies and standard deviation values over 10 runs for the 18 tasks are reported in Table 1.

From Table 1, we can see that the proposed semi-supervised matrix completion method outperforms all the other comparison methods across the eighteen tasks. The baseline *TBOW* method has poor performance on all the eighteen tasks, which shows that 100 labeled reviews in the target language are too few to develop a robust sentiment classifier. By exploiting the existing labeled data in the source language, all the five cross-lingual adaptation learning methods (*CL-LSI, CL-KCCA, CL-OPCA, TSL, SSMC*) consistently outperform the *TBOW* method across all the eighteen tasks, showing that the labeled data in the source language is useful for developing a better sentiment classification model in the target language and hence can reduce the expensive manual annotation cost in the target language. Among the unsupervised representation learning methods, *TSL* outperforms *CL-KCCA* and *CL-OPCA* on all of the eighteen tasks, and outperforms *CL-LSI* on seventeen out of the eighteen tasks (one exception is the task *EJB*). But by further incorporating the label information into the matrix completion process, the proposed *SSMC* method consistently outperforms all the unsupervised representation learning methods across all the eighteen tasks. All these results demonstrate that the proposed semi-supervised matrix completion method can produce more effective cross-lingual representations.

Impact of the Labeled Data in Target Language

Since the *SSMC* method learns interlingual representations in a semi-supervised manner by incorporating the existing label information, we then studied how the number of labeled training documents in the target language domain affects the classification performance. We considered a set of different values for the number of labeled training documents, ℓ_t, in the target language, such as $\ell_t \in$

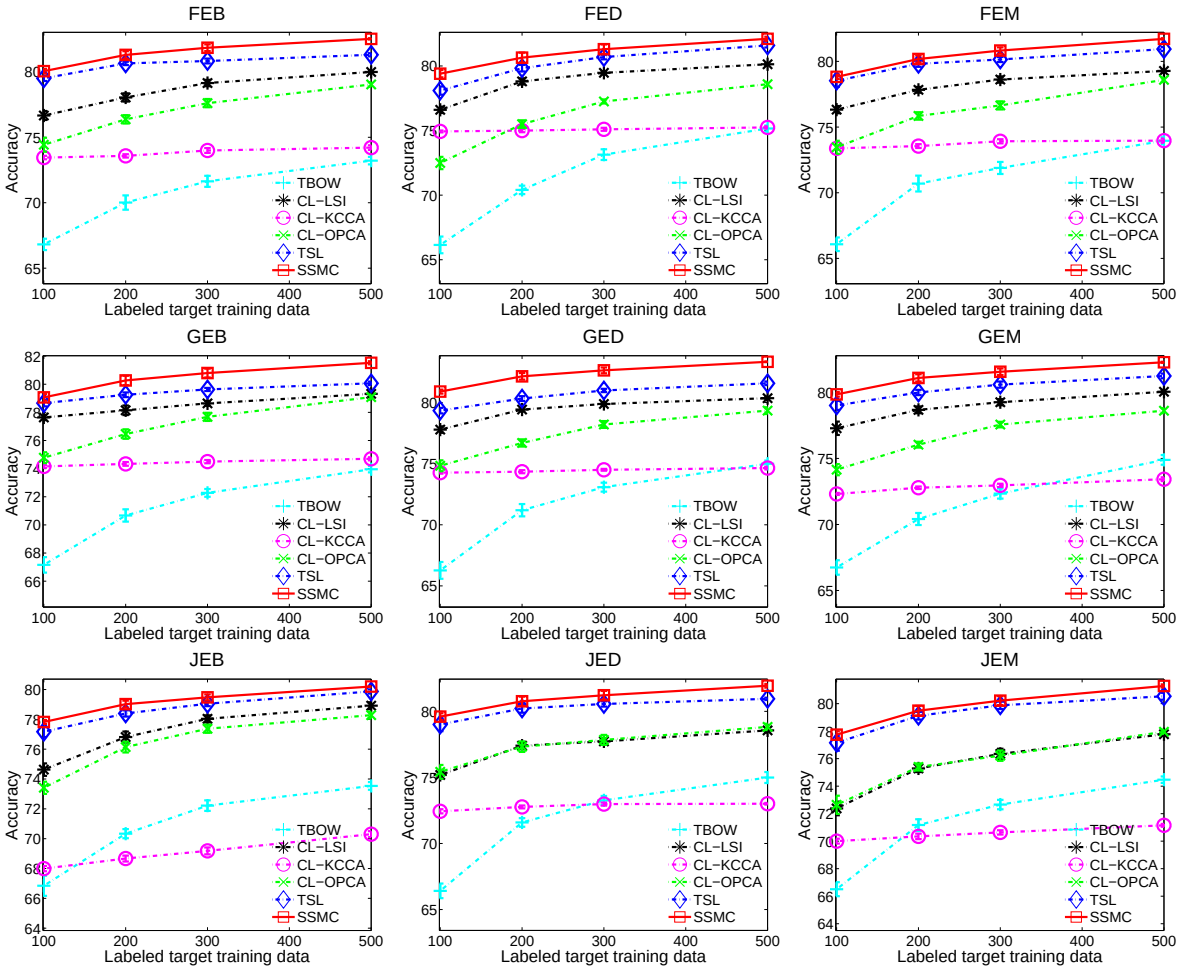

Figure 2: Average classification accuracies and standard deviations with respect to different numbers of labeled training documents in the target language for the nine CLTC tasks with English as the target language.

$\{100, 200, 300, 500\}$. We used the same experimental setting as above for each ℓ_t value. The average classification accuracies and standard deviations over 10 runs for the set of ℓ_t values are reported in Figure 1 and in Figure 2. Figure 1 presents the results for the nine cross-lingual sentiment classification tasks that use English as the source language and Figure 2 presents the results on the remaining nine cross language sentiment classification tasks that use English as the target language.

From Figure 1 and Figure 2, we can see that with the increasing of the number of labeled documents in the target language domain, the performance of all methods in general increases. But the *CL-OPCA* and the *CL-KCCA* methods fail to produce consistent improvements over the baseline *TBOW* method. For example, *CL-OPCA* performs worse than *TBOW* on the task *EFD* when the number of labeled training documents in the target language gets larger. Similar trends can be observed for *CL-KCCA* on the tasks *JEB* and *JEM*. Both *CL-LSI* and *TSL* perform better than *CL-KCCA* and *CL-OPCA* on most of the eighteen tasks. The proposed *SSMC* method however achieves the best classification accuracy across the range of different settings, and consistently outperforms all the other comparison methods. These results again justify the effectiveness of the proposed semi-supervised representation learning method.

Conclusion

In this paper, we proposed a semi-supervised representation learning approach to address cross language text classification, which exploits a small set of unlabeled parallel documents to bridge the disjoint feature spaces of the two languages. We formulated the representation learning problem as a convex semi-supervised matrix completion problem over the concatenation of the dual-language document-term matrix and the label vector, and solved it with a projected gradient descent algorithm. We evaluated the proposed learning technique on eighteen cross language sentiment classification tasks constructed from the Amazon product reviews in four different languages. The empirical results demonstrated the efficacy of the proposed method.

Acknowledgments

This research was supported by NSF grant IIS-1065397.

References

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

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

Bel, N.; Koster, C.; and Villegas, M. 2003. Cross-lingual text categorization. In *Proceedings of European Conference on Digital Libraries (ECDL)*.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Dai, W.; Chen, Y.; Xue, G.; Yang, Q.; and Yu, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *Advances in Neural Information Processing Systems (NIPS)*.

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

Gliozzo, A. 2006. 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 (COLING-ACL)*.

Klementiev, A.; Titov, I.; and Bhattarai, B. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

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

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

Ma, S.; Goldfarb, D.; and Chen, L. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming: Series A and B archive* 128, Issue 1-2.

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

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

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

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

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

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

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

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

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

Wu, K.; Wang, X.; and Lu, B. 2008. Cross language text categorization using a bilingual lexicon. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Xiao, M., and Guo, Y. 2013. A novel two-step method for cross language representation learning. In *Advances in Neural Information Processing Systems (NIPS)*.