Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

A Subspace Learning Framework For Cross-Lingual Sentiment Classification With Partial Parallel Data

Guangyou Zhou[1], Tingting He[1], Jun Zhao[2], and Wensheng Wu[3]
[1] School of Computer, Central China Normal University, Wuhan 430079, China
[2] National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China
[3] Computer Science Department, University of Southern California, Los Angeles, CA 90089-0781
{gyzhou,tthe}@mail.ccnu.edu.cn jzhao@nlpr.ia.ac.cn wuwens@gmail.com

Abstract

Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scarce target language by exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces of the source language data and that of the target language data. To address this challenge, previous work in the literature mainly relies on the large amount of bilingual parallel corpora to bridge the language gap. In many real applications, however, it is often the case that we have some partial parallel data but it is an expensive and time-consuming job to acquire large amount of parallel data on different languages. In this paper, we propose a novel subspace learning framework by leveraging the partial parallel data for cross-lingual sentiment classification. The proposed approach is achieved by jointly learning the document-aligned review data and un-aligned data from the source language and the target language via a non-negative matrix factorization framework. We conduct a set of experiments with cross-lingual sentiment classification tasks on multilingual Amazon product reviews. Our experimental results demonstrate the efficacy of the proposed cross-lingual approach.

1 Introduction

With the development of web 2.0, more and more user generated sentiment data have been shared on the web. They exist in the form of user reviews on shopping or opinion sites, in posts of blogs or customer feedback in different languages. These labeled user generated sentiment data are considered as the most valuable resources for the sentiment classification task. However, such resources in different languages are very imbalanced. Manually labeling each individual language is a time-consuming and labor-intensive job, which makes cross-lingual sentiment classification essential for this application.

Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scarce target language by exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces 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 engines [James G. Shanahan, 2004; Wu *et al.*, 2008; Wan, 2009; Prettenhofer and Stein, 2010; Wan, 2009; Pan *et al.*, 2011; Xiao and Guo, 2013]. Most of these studies translate documents from the source language to the target language or vice versa, and then apply the standard monolingual classification methods. However, due to the difference in language and culture, there exists a word drift problem. That is, while a word frequently appears in one language, its translated version may rarely appear in the other language [Guo, 2012].

Another group of works propose to use a large amount of bilingual parallel corpora to induce language-independent representations [Lu *et al.*, 2011; Meng *et al.*, 2012; Klementiev *et al.*, 2012]. In many real applications, however, it is often the case that we have some partial parallel data but it is a expensive and time-consuming job to acquire large amount of document-aligned (or sentence-aligned) data on different languages. For example in multilingual sentiment classification data [Prettenhofer and Stein, 2010], we have a large amount of Amazon product reviews written in four languages (English, French, German and Japanese) but only 2000 unlabeled parallel reviews between English and each of the other three languages are existed.

To handle such data, we propose a novel subspace learning approach to address the cross-lingual sentiment classification with the partial parallel data. Our assumption is that document-aligned parallel data describe the same semantics in two different languages, they should share the same latent representations under the discriminative subspace regarding the same classification task. Then the language gap between the source language and the target language can be reduced via the shared representations. Specially, our model is achieved by jointly learning the document-aligned review data and un-aligned data from the source language and the target language via a non-negative matrix factorization framework [Lee and Seung, 1999]. We derive an efficient algorithm for learning the factorization and provide proof of convergence. To evaluate the effectiveness of the proposed approach, we conduct a set of experiments with cross-lingual sentiment classification tasks on multilingual Amazon product reviews. The empirical results show that the proposed approach is effective for cross-lingual sentiment classification and outperforms other comparison methods.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes our proposed subspace learning framework on partial parallel data. Section 4 presents the experimental results. Finally, we conclude this paper in Section 5.

2 Related Work

In this Section, we present the related work on traditional monolingual sentiment classification and cross-lingual sentiment classification.

2.1 Sentiment Classification

Sentiment classification has gained widely interest in natural language processing (NLP) community. Methods for automatically classifying sentiments expressed in products and movie reviews can roughly be divided into supervised and unsupervised (or semi-supervised) sentiment analysis. Supervised techniques have been proved promising and widely used in sentiment classification [Pang *et al.*, 2002; Pang and Lee, 2008; Liu, 2012]. However, the performance of these methods relies on manually labeled training data. In some cases, the labeling work may be time-consuming and expensive. This motivates the problem of learning robust sentiment classification via an unsupervised (or semi-supervised) paradigm.

The most representative way to perform semi-supervised paradigm is to employ partial labeled data to guide the sentiment classification [Goldberg and Zhu, 2006; Sindhwani and Melville, 2008; Li *et al.*, 2011]. However, we do not have any labeled data at hand in many situations, which makes the unsupervised paradigm possible. The most representative way to perform unsupervised paradigm is to use a sentiment lexicon to guide the sentiment classification [Turney, 2002; Taboada *et al.*, 2011; Zhou *et al.*, 2014b] or learn sentiment orientation of a word from its semantically related words mined from the lexicon [Peng and Park, 2011]. Sentiment polarity of a word is obtained from off-the-shelf sentiment lexicon, the overall sentiment polarity of a document is computed as the summation of sentiment scores of the words in the document. All these works focus on the monolingual sentiment classification, we point the readers to recent books [Pang and Lee, 2008; Liu, 2012] for an in-depth survey of literature on sentiment classification.

2.2 Cross-Lingual Sentiment Classification

Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scarce target language by exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces of the source language data and that of the target language data.

To bridge the language gap, previous work in the literature mainly relies on machine translation engines or bilingual lexicons to directly adapt labeled data from the source language to the target language. Banea et al. [2008] employed the machine translation engines to bridge the language gap in different languages for multilingual subjectivity analysis. Wan [2009] proposed a co-training method to train Chinese sentiment classification model on English labeled data and their Chinese translations. English labeled data are first translated into Chinese, and then the bi-view sentiment classifiers are trained on English and Chinese labeled data respectively. Pan et al. [2011] proposed a bi-view non-negative matrix tri-factorization (BNMTF) model for cross-lingual sentiment classification problem. They employed machine translation engines so that both training and test data are able to have two representations, one in a source language and the other in a target language. The proposed model is derived from the non-negative matrix factorization models in both languages in order to make more accurate prediction. Prettenhofer and Stein [2010] proposed a cross-lingual structural correspondence learning (CL-SCL) method to induce language-independent features. Instead of using machine translation engines to translate labeled text, the authors first selected a subsect of pivot features in the source language to translate them into the target language, and then use these pivot pairs to induce cross-lingual representations by modeling the correlations between pivot features and non-pivot features in an unsupervised fashion. However, due to the difference in language and culture, these translation-based methods cause a word drift problem. That is, while a word frequently appears in one language, its translated version may rarely appear in the other language [Guo, 2012].

Another group of works propose to use a large amount of unlabeled parallel corpora to induce language-independent representations [Lu *et al.*, 2011; Meng *et al.*, 2012; Klementiev *et al.*, 2012]. Meng et al. [2012] proposed a generative cross-lingual mixture model (CLMM) to learn previously unseen sentiment words from the large bilingual parallel data. Klementiev et al. [2012] proposed to induce the bilingual word embeddings by using neural language models in an unsupervised setting. Zhou et al. [2014a] proposed to learn the distributed semantics for sentiment classification. A common property of these approaches are that they need a large amount of unlabeled parallel corpora to train the word alignment model [Meng *et al.*, 2012; Klementiev *et al.*, 2012]. In many real applications, however, it is often the case that we have some partial parallel data but it is an expensive and time-consuming job to acquire large amount of parallel data on different languages. In this paper, we propose a novel subspace learning framework for cross-lingual sentiment classification. Our proposed approach only needs some partial parallel data and can jointly learn the document-aligned review data and un-aligned data from the source language and the target language.

3 Our Proposed Approach

In this section, we present the formulation and optimization of our proposed subspace learning framework on the partial parallel data for cross-lingual sentiment classification. In the cross-lingual setting, a partial parallel data set $\mathbf{X} = \{\mathbf{X}^{(s,t)}, \mathbf{X}^{(s)}, \mathbf{X}^{(t)}\}$ is given, where $\mathbf{X}^{(s,t)} = [(\mathbf{x}_1^s, \mathbf{x}_1^t); \cdots; (\mathbf{x}_c^s, \mathbf{x}_c^t)]$ denotes the documents present and only present in the parallel data, $\mathbf{X}^{(s)} = [\mathbf{x}^{(s)}_{c+1}; \cdots; \mathbf{x}^{(s)}_{c+m}]$ denotes the documents from the source language, and $\mathbf{X}^{(t)} = [\mathbf{x}^{(t)}_{c+m+1}; \cdots; \mathbf{x}^{(t)}_{c+m+n}]$ denotes the documents from the target language, c, m and n represent the number of documents present and only present in the parallel data, from the source language and the target language. The goal of cross-lingual sentiment classification with partial parallel data is to establish a latent subspace where the documents from the parallel data should share the same latent representations. The language gap can be reduced with the help of the shared latent representations.

3.1 Model Formulation

Let $\mathbf{X}^{(s,t)} = [\hat{\mathbf{X}}^{(s)}_c, \hat{\mathbf{X}}^{(t)}_c]$ be composed of documents $\hat{\mathbf{X}}^{(s)}_c \in \mathbb{R}^{c \times d_1}$, $\hat{\mathbf{X}}^{(t)}_c \in \mathbb{R}^{c \times d_2}$ coming from the the parallel sentiment review data. We now have the documents of each language denoted as $\bar{\mathbf{X}}^{(s)} = [\hat{\mathbf{X}}^{(s)}_c; \mathbf{X}^{(s)}] \in \mathbb{R}^{(c+m) \times d_1}$, $\bar{\mathbf{X}}^{(t)} = [\hat{\mathbf{X}}^{(t)}_c; \mathbf{X}^{(t)}] \in \mathbb{R}^{(c+n) \times d_2}$. As we deal with the text data, non-negative matrix factorization (NMF) [Lee and Seung, 1999] has been widely used for document representation, which actually assumes that the documents are generated by additive combination of an underlying set of hidden units. Incorporating the NMF technique into our problem, for each language, its latent subspace learning can be formulated as:

$$\min_{\bar{\mathbf{V}}^{(s)} \geq 0, \mathbf{U}^{(s)} \geq 0} \|\bar{\mathbf{X}}^{(s)} - \bar{\mathbf{V}}^{(s)}\mathbf{U}^{(s)}\|_F^2 + \lambda \Omega(\bar{\mathbf{V}}^{(s)}) \tag{1}$$

$$\min_{\bar{\mathbf{V}}^{(t)} \geq 0, \mathbf{U}^{(t)} \geq 0} \|\bar{\mathbf{X}}^{(t)} - \bar{\mathbf{V}}^{(t)}\mathbf{U}^{(t)}\|_F^2 + \lambda \Omega(\bar{\mathbf{V}}^{(t)}) \tag{2}$$

where $\mathbf{U}^{(s)} \in \mathbb{R}^{k \times d_1}$ and $\mathbf{U}^{(t)} \in \mathbb{R}^{k \times d_2}$ are the basis matrix for each language's latent space, and $\bar{\mathbf{V}}^{(s)} = [\hat{\mathbf{V}}^{(s)}_c; \mathbf{V}^{(s)}] \in \mathbb{R}^{(c+m) \times k}$, $\bar{\mathbf{V}}^{(t)} = [\hat{\mathbf{V}}^{(t)}_c; \mathbf{V}^{(t)}] \in \mathbb{R}^{(c+n) \times k}$ are the latent representation of documents in the latent space. The same latent dimension k is shared between the source language and the target language. λ is the tradeoff parameter for the regularization term $\Omega(\bar{\mathbf{V}})$. By equation (1) and equation (2), the latent space basis $\mathbf{U}$ and the corresponding document latent representation $\mathbf{V}$ are simultaneously learned to minimize the document reconstruction error, which forces all documents from each language to share the similar compact representation in the latent space.

So far, the latent spaces are learned independently for each language. For the cross-lingual setting on the partial parallel data, the document-aligned sentiment review data in $\hat{\mathbf{X}}^{(s)}_c$, $\hat{\mathbf{X}}^{(t)}_c$ describing the same data object, their latent representation $\hat{\mathbf{V}}^{(s)}_c$, $\hat{\mathbf{V}}^{(t)}_c$ should also be the same. Combining this idea and equation (1), equation (2), by enforcing $\hat{\mathbf{V}}^{(s)}_c = \hat{\mathbf{V}}^{(t)}_c = \hat{\mathbf{V}}_c$, we aim to minimize the following objective function:

$$\begin{aligned} \mathcal{O} = & \left\| \begin{bmatrix} \hat{\mathbf{X}}^{(s)}_c \\ \mathbf{X}^{(s)} \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{V}}_c \\ \mathbf{V}^{(s)} \end{bmatrix} \mathbf{U}^{(s)} \right\|_F^2 + \lambda \left\| \begin{bmatrix} \hat{\mathbf{V}}_c \\ \mathbf{V}^{(s)} \end{bmatrix} \right\|_1 \\ & + \left\| \begin{bmatrix} \hat{\mathbf{X}}^{(t)}_c \\ \mathbf{X}^{(t)} \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{V}}_c \\ \mathbf{V}^{(t)} \end{bmatrix} \mathbf{U}^{(t)} \right\|_F^2 + \lambda \left\| \begin{bmatrix} \hat{\mathbf{V}}_c \\ \mathbf{V}^{(t)} \end{bmatrix} \right\|_1 \\ & s.t. \quad \mathbf{U}^{(s)} \geq 0, \mathbf{U}^{(t)} \geq 0, \hat{\mathbf{V}}_c \geq 0, \mathbf{V}^{(s)} \geq 0, \mathbf{V}^{(t)} \geq 0 \end{aligned} \tag{3}$$

Now we can have the homogeneous feature representation for all documents as $\mathbf{V} = [\hat{\mathbf{V}}_c; \mathbf{V}^{(s)}; \mathbf{V}^{(t)}]$, whether they are originally partial or not. Any standard monolingual sentiment classification approach can be applied to such representation. The novelty of equation (3) lies in that $\bar{\mathbf{V}}^{(s)}$ and $\bar{\mathbf{V}}^{(t)}$ share the same part $\hat{\mathbf{V}}_c$ and at the same time has their own individual part $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(t)}$. Moreover, the proposed framework is learned by using all available documents of each language, the individual basis matrix $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$ are connected by the common $\hat{\mathbf{V}}_c$. Lasso is used for $\Omega(\mathbf{V})$ in this work as one of the mostly often used regularization for text analysis [Hu *et al.*, 2013].

3.2 Learning Algorithm

In this subsection, we present the solution to the optimization problem in equation (3), which is convex in latent representations $\mathbf{V}$ given the basis matrix $\mathbf{U}$ and vice versa, but not jointly convex in both, we propose an iterative update procedure and prove its convergence. Firstly, the basis matrices are initialized by the initialization step and then the following steps are repeated until convergence: (1) minimizing the objective function $\mathcal{O}$ over $\mathbf{V}$ with fixed $\mathbf{U}$; and (2) minimizing the objective function $\mathcal{O}$ over $\mathbf{U}$ with fixed $\mathbf{V}$.

Initialization Since the efficiency of the iterative optimization procedure is greatly affected by the initialization step, we learn the initial value of $\mathbf{U}$ rather than random selection:

$$\begin{aligned} \mathcal{O}_{init} = \|\hat{\mathbf{X}}^{(s)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(s)}\|_F^2 + \|\hat{\mathbf{X}}^{(t)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(t)}\|_F^2 + \lambda\|\hat{\mathbf{V}}_c\|_1 \\ s.t. \quad \mathbf{U}^{(s)} \geq 0, \mathbf{U}^{(t)} \geq 0, \hat{\mathbf{V}}_c \geq 0 \end{aligned} \tag{4}$$

From equation (4), we can see that $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$ are essentially initialized by applying traditional NMF on documents without partial parallel data. This initialization is also solved by iterative optimization. At each iteration, $\mathcal{O}_{init}$ is minimized alternatively over $\hat{\mathbf{V}}_c \geq 0$, $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$. Fixing $\hat{\mathbf{V}}_c \geq 0$, $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$ can be independently optimized by:

$$\min_{\mathbf{U}^{(s)} \geq 0} \mathcal{O}_{init}(\mathbf{U}^{(s)}) = \|\hat{\mathbf{X}}^{(s)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(s)}\|_F^2 \tag{5}$$

$$\min_{\mathbf{U}^{(t)} \geq 0} \mathcal{O}_{init}(\mathbf{U}^{(t)}) = \|\hat{\mathbf{X}}^{(t)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(t)}\|_F^2 \tag{6}$$

Fixing $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$, $\hat{\mathbf{V}}_c$ is optimized by:

$$\begin{aligned} \min_{\hat{\mathbf{V}}_c \geq 0} \mathcal{O}_{init}(\hat{\mathbf{V}}_c) = \|\hat{\mathbf{X}}^{(s)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(s)}\|_F^2 \\ + \|\hat{\mathbf{X}}^{(t)}_c - \hat{\mathbf{V}}_c\mathbf{U}^{(t)}\|_F^2 + \lambda\|\hat{\mathbf{V}}_c\|_1 \end{aligned} \tag{7}$$

(1) Minimizing $\mathcal{O}$ over $\mathbf{V}$ with fixed $\mathbf{U}$ Given the basis matrix $\mathbf{U}$ for each language, the computation of $\hat{\mathbf{V}}$ do not depend on each other. Therefore, the objective function in equation (3) reduces to:

$$\min_{\mathbf{V}^{(s)} \geq 0} \mathcal{O}(\mathbf{V}^{(s)}) = \|\mathbf{X}^{(s)} - \mathbf{V}^{(s)}\mathbf{U}^{(s)}\|_F^2 + \lambda\|\mathbf{V}^{(s)}\|_1 \tag{8}$$

$$\min_{\mathbf{V}^{(t)} \geq 0} \mathcal{O}(\mathbf{V}^{(t)}) = \|\mathbf{X}^{(t)} - \mathbf{V}^{(t)}\mathbf{U}^{(t)}\|_F^2 + \lambda\|\mathbf{V}^{(t)}\|_1 \tag{9}$$

$$\min_{\hat{\mathbf{V}}_c \geq 0} \mathcal{O}(\hat{\mathbf{V}}_c) = \|\hat{\mathbf{X}}_c^{(s)} - \hat{\mathbf{V}}_c \mathbf{U}^{(s)}\|_F^2 + \|\hat{\mathbf{X}}_c^{(t)} - \hat{\mathbf{V}}_c \mathbf{U}^{(t)}\|_F^2 + \lambda \|\hat{\mathbf{V}}_c\|_1 \tag{10}$$

Noting the same formulation of equation (10) as equation (7), so at the first iteration, $\hat{\mathbf{V}}_c$ has already been obtained from initialization.

(2) Minimizing $\mathcal{O}$ over U with fixed V Given $\mathbf{V} = [\hat{\mathbf{V}}_c; \mathbf{V}^{(s)}; \mathbf{V}^{(t)}]$, the latent representations for documents of each language can be obtain as $[\hat{\mathbf{V}}_c; \mathbf{V}^{(s)}]$ and $[\hat{\mathbf{V}}_c; \mathbf{V}^{(t)}]$, minimizing $\mathcal{O}$ over $\mathbf{U}$ now independently reduces to:

$$\min_{\mathbf{U}^{(s)} \geq 0} \mathcal{O}(\mathbf{U}^{(s)}) = \|\bar{\mathbf{X}}^{(s)} - [\hat{\mathbf{V}}_c; \mathbf{V}^{(s)}]\mathbf{U}^{(s)}\|_F^2 \tag{11}$$

$$\min_{\mathbf{U}^{(t)} \geq 0} \mathcal{O}(\mathbf{U}^{(t)}) = \|\bar{\mathbf{X}}^{(t)} - [\hat{\mathbf{V}}_c; \mathbf{V}^{(t)}]\mathbf{U}^{(t)}\|_F^2 \tag{12}$$

To solve the optimization problem in equation (5)$\sim$ equation (12), which are lasso regularized NMF with one factor fixed, we employ the greedy coordinate descent (GCD) approach proposed by Hsieh and Dhillon [2011], which is about 10 times faster than cyclic coordinate descent scheme and proved to converge. To get a robust and stable sparsity tradeoff parameter for different data sets, λ is normalized by the data size during the optimization, i.e., for equations (8), (9) and (10), λ is timed by coefficient $(d_1 + d_2)/k$, d_1/k and d_2/k, respectively.

Theorem 1. *The objective function in equation (3) is non-increasing under the optimization procedure in Algorithm 1.*

Lemma 1. *[Hsieh and Dhillon, 2011] For least squares NMF, if a basis matrix, latent representation pair sequence $\{(\mathbf{U}_j, \mathbf{V}_j)\}$ is generated by GCD, then every limit point of this sequence is a stationary point.*

Proof. To prove Theorem 1, we only need to prove that the objective function in equation (3) is non-increasing after each step. With fixed $\mathbf{U}^{(s)}$ and $\mathbf{U}^{(t)}$, the objective function value of equation (3) with respect to $\hat{\mathbf{V}}_c, \mathbf{V}^{(s)}, \mathbf{V}^{(t)}$ equals the sum of the objective function value of equation (8)$\sim$equation (10). With fixed $\hat{\mathbf{V}}_c, \mathbf{V}^{(s)}, \mathbf{V}^{(t)}$, the objective value of equation (3) with respect to $\mathbf{U}^{(s)}$, $\mathbf{U}^{(t)}$ equals the sum of equation (11)$\sim$equation (12). By Lemma 1, the objective function value of equation (8)$\sim$equation (12) are guaranteed to converge to some local minima. So the objective function value of equation (3) is guaranteed to non-increase after each step. □

3.3 Cross-Lingual Sentiment Classification

Once the parameters $\{\hat{\mathbf{V}}_c, \mathbf{V}^{(s)}, \mathbf{V}^{(t)}\}$ and $\{\mathbf{U}^{(s)}, \mathbf{U}^{(t)}\}$ are obtained, we use the learned latent representations $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(t)}$ of the labeled data from the two languages and train a simple sentiment classification model using a linear support vector machine (e.g., Liblinear [Fan *et al.*, 2008]). Then, we predict the sentiment polarity of the test data from the target language on the latent representations $\mathbf{V}^{(t)}$ with the learned classification model.

4 Experiments

In this section, we conduct experiments for cross-lingual sentiment classification on multilingual Amazon product reviews in four languages.

4.1 Data Sets

We use the multilingual sentiment classification data set provided by [Prettenhofer and Stein, 2010], which contains Amazon product reviews in four languages (English (E), French (F), German (G) and Japanese (J)) of three categories (Books (B), DVD (D), Music (M)). The English product reviews are sampled from previous cross-domain sentiment classification data sets [Blitzer *et al.*, 2007], while the other three language product reviews are crawled from Amazon by the authors in November. For each category of the product reviews, there are 2000 positive and 2000 negative English reviews (represented with $\mathbf{X}^{(s)}$), and 1000 positive and 1000 negative reviews for each of the other three languages (represented with $\mathbf{X}^{(t)}$). Besides, we have another 2000 unlabeled parallel sentiment reviews between English and each of the other three languages (partial parallel data, represented with $\mathbf{X}^{(s,t)}$). All data described in the above is publicly available from the website.[1]

Following the literature [Xiao and Guo, 2013], each review is represented as a unigram bag-of-words vector representation and each entry is computed with tf-idf. Given the four languages from the three categories, we construct 18 cross-lingual sentiment classification tasks (EFB, EFD, EFM, EGB, EGD, EGM, EJB, EJD, EJM, FEB, FED, FEM, GEB, GED, GEM, JEB, JED, JEM) between English and the other three languages. For example, the task EFB uses English Books reivews as the source language data and uses French Books reviews as the target language data.

4.2 Baseline Algorithms

In our experiments, we re-implement the following baseline algorithms in order to compare with the proposed approach:

- **TB:** This is a target bag-of-words baseline method, which trains a supervised monolingual classifier on the labeled training data from the target language without using the unlabeled parallel data.
- **CL-LSA:** This is the cross-lingual learning method described in [Gliozzo and Strapparava, 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-lingual structural correspondence learning (SCL) method described in [Prettenhofer and Stein, 2010], which first chooses some pivot features and then automatically induce the cross-lingual correspondences with a bilingual dictionary. We implement this approach by using the source codes provided by the authors.[2]
- **CL-MT:** This is a machine translation based method, which first uses the Google Translate (http://translate.google.com) tool to translate the target language documents into the source language and

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

[2] https://github.com/pprett/nut

Table 1: Average classification accuracies (%) and standard derivations (%) over 10 test runs for the 18 cross-lingual sentiment classification tasks. The bold format indicates that the difference between the results of our proposed approach and state-of-the-art CL-TS is significant with $p < 0.05$ under a McNemar paired test for labeling disagreements.

Task	TB	CL-LSA	CL-SCL	CL-MT	CL-OPCA	CL-TS	`this work`
EFB	66.89±0.87	79.38±0.25	79.86±0.22	78.01±0.45	76.47±0.35	81.83±0.25	**82.61±0.25**
EFD	67.42±0.91	77.69±0.54	78.80±0.25	77.75±0.68	70.36±0.43	81.92±0.35	**82.70±0.45**
EFM	67.55±0.50	75.26±0.43	75.95±0.31	75.86±0.51	73.43±0.37	79.06± 0.28	**80.19±0.40**
EGB	67.31±0.72	77.60±0.37	77.77±0.28	77.02±0.60	74.65±0.48	79.30±0.34	**79.91±0.47**
EGD	66.54±0.73	79.16±0.26	79.93±0.23	79.75±0.58	74.47±0.56	81.27±0.26	81.86± 0.31
EGM	67.61±0.46	73.67±0.52	73.95±0.30	73.69±0.55	74.38±0.61	79.26±0.33	79.59± 0.42
EJB	62.91±0.62	72.55±0.41	72.91±0.25	72.20±0.80	71.17±0.50	72.40±0.48	**73.45±0.27**
EJD	65.33±0.58	72.51±0.32	72.82±0.28	72.68±0.56	71.70±0.45	76.51±0.37	**77.06±0.32**
EJM	67.28±0.67	73.40±0.47	73.75±0.35	73.33±0.65	74.82±0.64	76.17±0.43	**76.83±0.52**
FEB	66.77±0.51	76.61±0.40	77.26±0.22	77.43±0.55	74.29±0.52	79.29±0.30	**80.48±0.33**
FED	65.98±0.57	76.39±0.34	76.57±0.20	76.80±0.52	72.30±0.57	77.88±0.34	**78.76±0.38**
FEM	65.92±0.55	76.24±0.35	76.76±0.25	76.19±0.48	73.41±0.55	78.31±0.41	**79.18±0.33**
GEB	67.05±0.68	77.43±0.26	77.85±0.27	77.50±0.66	74.66±0.42	78.45±0.29	78.61± 0.34
GED	66.30±0.60	77.55±0.30	77.83±0.33	77.52±0.54	74.68±0.54	79.22±0.28	**80.27±0.35**
GEM	66.55±0.46	77.03±0.42	77.37±0.34	77.63±0.51	74.07±0.46	78.90±0.37	**79.80±0.26**
JEB	66.72±0.63	74.49±0.37	75.25±0.30	74.41±0.47	73.38±0.45	77.11±0.30	**77.97±0.35**
JED	66.32±0.49	75.17±0.24	75.34±0.27	75.15±0.49	75.37±0.48	78.95±0.46	**80.63±0.38**
JEM	66.48±0.55	72.29±0.45	73.21±0.33	73.16±0.50	72.55±0.61	77.13±0.51	**77.78±0.37**

then trains a monolingual classifier with labeled training data from both languages.

- **CL-OPCA:** This is the cross-lingual oriented principal component analysis (OPCA) method described in [Platt *et al.*, 2010], which first learn cross-lingual representations with all data from both languages by performing OPCA and then train a monolingual classifier with labeled data from both languages in the induced feature space.
- **CL-TS:** This is the state-of-the-art method for cross-lingual sentiment classification described in [Xiao and Guo, 2013], which formulates the cross-lingual representation as a matrix completion problem to infer unobserved feature values of the concatenated document-term matrix in the space of unified vocabulary set from the source and target languages by using unlabeled parallel bilingual documents.[3]

In all experiments, we train the sentiment classification model on the latent representations using the Liblinear [Fan *et al.*, 2008]. For **CL-LSA**, **CL-OPCA** and **CL-TS**, we use the same parameter value $k = 50$ as suggested in the paper [Xiao and Guo, 2013]. For **CL-SCL**, we use 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. We choose the above parameter values empirically because these parameter settings have shown superior performance on the same benchmark [Prettenhofer and Stein, 2010].

4.3 Classification Accuracy

For each of the 18 cross-lingual sentiment classification task, we use all documents from the two languages (including 2000 parallel sentiment data). Then we use all documents from the source language and randomly choose 100 documents from the target language as labeled data to build the classification model, and use the rest data from the target language as test data. For our proposed algorithm, we choose the latent dimension k values from $\{10, 20, 50, 100, 150, 250, 300\}$, choose the parameter λ value from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We perform the model parameter selection by running the algorithm 3 times based on random selection of 100 labeled target training data (in short, `Tuning set`). Finally, we set parameters $k = 100$ and $\lambda = 10^{-2}$. We then use the selected model parameters for all the 18 tasks and run each experiment for 10 times based on random selections of 100 labeled target documents (in short, `Test set`).[4] The average sentiment classification accuracies and standard deviations are presented in Table 1.

From Table 1, we can see that our proposed approach outperforms all other six comparison methods in general. The target baseline TB performs poorly on all the 18 tasks, which indicates that 100 labeled documents from the target language is far from enough to obtain an accurate and robust sentiment classifier in the target language domain. All the other five cross-lingual sentiment classification methods, CL-LSA, CL-SCL, CL-MT, CL-OPCA and CL-TS, consistently outperform the baseline method TB across all the 18 tasks, which

[3] There is a slight exception: using the same data set and parameter setting, our re-implemented systems have slight differences with the results reported in [Xiao and Guo, 2013].

[4] During the experiments, we try our best to eliminate the overlap between the `Tuning set` and `Test set`. In this paper, we would do several experiments by random sampling rather than perform cross-validation in order to make a fair comparison with the previous work [Prettenhofer and Stein, 2010; Xiao and Guo, 2013].

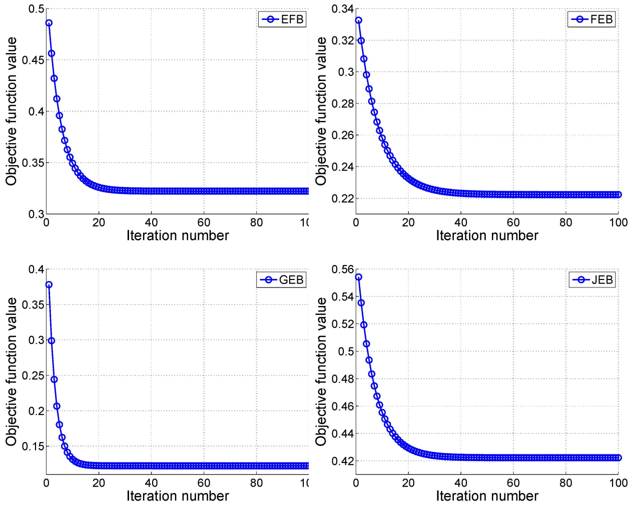

Figure 1: Convergence curve of of the proposed algorithm.

demonstrates that the labeled training data from the source language domain is useful for classifying the target language data. Nevertheless, the improvements achieved by these five methods over the baseline are much smaller than the proposed approach. Among all the 18 tasks, our proposed approach increases the average test accuracy over the baseline TB by at least 9.21% on the EJM task and up to 15.30% on the EFB task. Moreover, our propose approach also outperforms CL-LSA, CL-SCL, CL-MT and CL-OPCA across all the 18 tasks, outperforms CL-TS on 17 out of the 18 tasks and achieves the slight lower performance than CL-TS on the GEB task.

We also conduct significance tests for our proposed approach and the state-of-the-art CL-TS using a McNemar paired test for labeling disagreements [Gillick and Cox, 1989]. The results in bold formate indicate that they are significant with $p < 0.05$. All these results demonstrate the efficacy and robustness of the proposed subspace learning framework on partial parallel data for cross-lingual sentiment classification.

4.4 Convergence Analysis

In section 3.2, we have shown that the proposed objective function is convergent. Here, we empirically show the convergence analysis for the selected four tasks due to space limitation. Figure 1 shows the convergence curves of our proposed algorithm on the four tasks. From the figures, y-axis is the value of the objective function and x-axis denotes the iteration number. It can be seen that the objective function value monotonically decreases as the iteration number increases. Though it takes a lot of rounds to converge, the GCD in each iteration runs very fast, usually within 60 iterations.

4.5 Parameter Study

In this section, we explore the effect on the parameters k and λ to the sentiment classification performance. We choose k value from $\{10, 20, 50, 100, 150, 250, 300\}$ and λ value from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Due to space limitation, we only present the experimental study on the selected

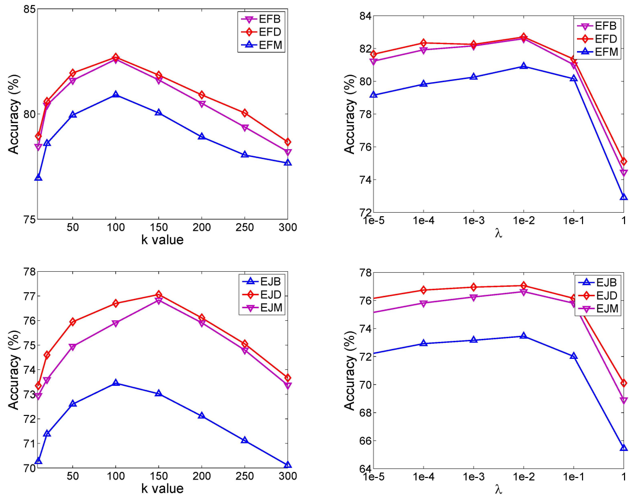

Figure 2: Influence of k and λ on the selected six tasks.

six tasks (EFB, EFD, EFM, EJB, EJD and EJM) in Figure 2. From Figure 2, we can see that the proposed algorithm achieves stable good performance when k is around 100 and λ is around 10^{-2}.

5 Conclusion

In this paper, we propose a novel subspace learning framework by leveraging the partial parallel data for cross-lingual sentiment classification. The proposed approach is achieved by jointly learning the document-aligned review data and unaligned data from the source language and the target language via a non-negative matrix factorization framework. We also derive an efficient algorithm for learning the factorization and provide proof of convergence. To evaluate the effectiveness of the proposed approach, we conduct a set of experiments with cross-lingual sentiment classification tasks on multilingual Amazon product reviews. The empirical results show that the proposed approach is effective for cross-lingual sentiment classification and outperforms other comparison methods. In the future, we will study how to extend the proposed subspace learning framework to multilingual settings and to nonlinear latent subspace cases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61303180 and No. 61272332), the Beijing Natural Science Foundation (No. 4144087), the Major Project of National Social Science Found (No. 12&2D223), the Fundamental Research Funds for the Central Universities (No. CCNU15ZD003), and also Sponsored by CCF-Tencent Open Research Fund. We thank the anonymous reviewers for their insightful comments.

References

[Banea *et al.*, 2008] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *EMNLP*, pages 127–135, 2008.

[Blitzer *et al.*, 2007] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: domain adaptation for sentiment classification. In *ACL*, 2007.

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

[Gillick and Cox, 1989] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recoginition algorithms. In *ICASSP*, 1989.

[Gliozzo and Strapparava, 2006] Alfio Gliozzo and Carlo Strapparava. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *ACL*, pages 553–560, 2006.

[Goldberg and Zhu, 2006] Andrew B. Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, 2006.

[Guo, 2012] Yuhong Guo. Cross language text classification via subspace co-regularized multi-view learning. In *ICML*, 2012.

[Hsieh and Dhillon, 2011] C. Hsieh and I. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *KDD*, 2011.

[Hu *et al.*, 2013] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, 2013.

[James G. Shanahan, 2004] Yan Qu David A. Evans James G. Shanahan, Gregory Grefenstette. Mining multilingual opinions through classification and translation. *AAAI*, 2004.

[Klementiev *et al.*, 2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, Bombay, India, December 2012.

[Lee and Seung, 1999] Lee and Seung. Learning the parts of ojbect by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Li *et al.*, 2011] Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-supervised learning for imbalanced sentiment classification. In *IJCAI*, pages 1826–1831, 2011.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 2012.

[Lu *et al.*, 2011] Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. Joint bilingual sentiment classification with unlabeled parallel corpora. In *ACL-HLT*, pages 320–330, 2011.

[Meng *et al.*, 2012] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. Cross-lingual mixture model for sentiment classification. In *ACL*, pages 572–581, 2012.

[Pan *et al.*, 2011] Junfeng Pan, Gui-Rong Xue, Yong Yu, and Yang Wang. Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'11, pages 289–300, Berlin, Heidelberg, 2011. Springer-Verlag.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.

[Pang *et al.*, 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.

[Peng and Park, 2011] Wei Peng and Dae Hoon Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *ICWSM*, 2011.

[Platt *et al.*, 2010] J. Platt, K. Toutanova, and W. Yih. Translingual document representations from dicriminative projections. In *EMNLP*, 2010.

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

[Sindhwani and Melville, 2008] Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, 2008.

[Taboada *et al.*, 2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[Turney, 2002] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424, 2002.

[Wan, 2009] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *ACL*, pages 235–243, 2009.

[Wu *et al.*, 2008] Ke Wu, Xiaolin Wang, and Bao liang Lu. Cross language text categorization using a bilingual lexicon. In *IJCAI*, 2008.

[Xiao and Guo, 2013] Min Xiao and Yuhong Guo. A novel two-step method for cross language representation learning. In *NIPS*, pages 1259–1267, 2013.

[Zhou *et al.*, 2014a] Guangyou Zhou, Tingting He, and Jun Zhao. Bridging the language gap: Learning distributed semantics for cross-lingual sentiment classification. In *Proceedings of Natural Language Processing and Chinese Computing (NLPCC 2014)*, pages 138–149, 2014.

[Zhou *et al.*, 2014b] Guangyou Zhou, Jun Zhao, and Daojian Zeng. Sentiment classification with graph co-regularization. In *Proceedings of COLING 2014*, pages 1331–1340, 2014.