Cross-Lingual Mixture Model for Sentiment Classification

Xinfan Meng[‡] [*] **Furu Wei**[†] **Xiaohua Liu**[†] **Ming Zhou**[†] **Ge Xu**[‡] **Houfeng Wang**[‡]
[‡]MOE Key Lab of Computational Linguistics, Peking University
[†]Microsoft Research Asia
[‡]{mxf, xuge, wanghf}@pku.edu.cn
[†]{fuwei,xiaoliu,mingzhou}@microsoft.com

Abstract

The amount of labeled sentiment data in English is much larger than that in other languages. Such a disproportion arouse interest in cross-lingual sentiment classification, which aims to conduct sentiment classification in the target language (e.g. Chinese) using labeled data in the source language (e.g. English). Most existing work relies on machine translation engines to directly adapt labeled data from the source language to the target language. This approach suffers from the limited coverage of vocabulary in the machine translation results. In this paper, we propose a generative cross-lingual mixture model (**CLMM**) to leverage unlabeled bilingual parallel data. By fitting parameters to maximize the likelihood of the bilingual parallel data, the proposed model learns previously unseen sentiment words from the large bilingual parallel data and improves vocabulary coverage significantly. Experiments on multiple data sets show that CLMM is consistently effective in two settings: (1) labeled data in the target language are unavailable; and (2) labeled data in the target language are also available.

1 Introduction

Sentiment Analysis (also known as opinion mining), which aims to extract the sentiment information from text, has attracted extensive attention in recent years. Sentiment classification, the task of determining the sentiment orientation (positive, negative or neutral) of text, has been the most extensively studied task in sentiment analysis. There is already a large amount of work on sentiment classification of text in various genres and in many languages. For example, Pang et al. (2002) focus on sentiment classification of movie reviews in English, and Zagibalov and Carroll (2008) study the problem of classifying product reviews in Chinese. During the past few years, NTCIR[1] organized several pilot tasks for sentiment classification of news articles written in English, Chinese and Japanese (Seki et al., 2007; Seki et al., 2008).

For English sentiment classification, there are several labeled corpora available (Hu and Liu, 2004; Pang et al., 2002; Wiebe et al., 2005). However, labeled resources in other languages are often insufficient or even unavailable. Therefore, it is desirable to use the English labeled data to improve sentiment classification of documents in other languages. One direct approach to leveraging the labeled data in English is to use machine translation engines as a *black box* to translate the labeled data from English to the target language (e.g. Chinese), and then using the translated training data directly for the development of the sentiment classifier in the target language (Wan, 2009; Pan et al., 2011).

Although the machine-translation-based methods are intuitive, they have certain limitations. First, the vocabulary covered by the translated labeled data is limited, hence many sentiment indicative words can not be learned from the translated labeled data. Duh et al. (2011) report low overlapping between vocabulary of natural English documents and the vocabulary of documents translated to English from Japanese, and the experiments of Duh

[*] Contribution during internship at Microsoft Research Asia.

[1] http://research.nii.ac.jp/ntcir/index-en.html

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 572–581,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

et al. (2011) show that vocabulary coverage has a strong correlation with sentiment classification accuracy. Second, machine translation may change the sentiment polarity of the original text. For example, the negative English sentence “It is too good to be true” is translated to a positive sentence in Chinese “这是好得是真实的” by Google Translate (http://translate.google.com/), which literally means “It is good and true”.

In this paper we propose a cross-lingual mixture model (**CLMM**) for cross-lingual sentiment classification. Instead of relying on the unreliable machine translated labeled data, CLMM leverages bilingual parallel data to bridge the language gap between the source language and the target language. CLMM is a generative model that treats the source language and target language words in parallel data as generated *simultaneously* by a set of mixture components. By “synchronizing” the generation of words in the source language and the target language in a parallel corpus, the proposed model can (1) improve vocabulary coverage by learning sentiment words from the unlabeled parallel corpus; (2) transfer polarity label information between the source language and target language using a parallel corpus. Besides, CLMM can improve the accuracy of cross-lingual sentiment classification consistently regardless of whether labeled data in the target language are present or not. We evaluate the model on sentiment classification of Chinese using English labeled data. The experiment results show that CLMM yields 71% in accuracy when no Chinese labeled data are used, which significantly improves Chinese sentiment classification and is superior to the SVM and co-training based methods. When Chinese labeled data are employed, CLMM yields 83% in accuracy, which is remarkably better than the SVM and achieve state-of-the-art performance.

This paper makes two contributions: (1) we propose a model to effectively leverage large bilingual parallel data for improving vocabulary coverage; and (2) the proposed model is applicable in both settings of cross-lingual sentiment classification, irrespective of the availability of labeled data in the target language.

The paper is organized as follows. We review related work in Section 2, and present the cross-lingual mixture model in Section 3. Then we present the experimental studies in Section 4, and finally conclude the paper and outline the future plan in Section 5.

2 Related Work

In this section, we present a brief review of the related work on monolingual sentiment classification and cross-lingual sentiment classification.

2.1 Sentiment Classification

Early work of sentiment classification focuses on English product reviews or movie reviews (Pang et al., 2002; Turney, 2002; Hu and Liu, 2004). Since then, sentiment classification has been investigated in various domains and different languages (Zagibalov and Carroll, 2008; Seki et al., 2007; Seki et al., 2008; Davidov et al., 2010). There exist two main approaches to extracting sentiment orientation automatically. The Dictionary-based approach (Turney, 2002; Taboada et al., 2011) aims to aggregate the sentiment orientation of a sentence (or document) from the sentiment orientations of words or phrases found in the sentence (or document), while the corpus-based approach (Pang et al., 2002) treats the sentiment orientation detection as a conventional classification task and focuses on building classifier from a set of sentences (or documents) labeled with sentiment orientations.

Dictionary-based methods involve in creating or using sentiment lexicons. Turney (2002) derives sentiment scores for phrases by measuring the mutual information between the given phrase and the words “excellent” and “poor”, and then uses the average scores of the phrases in a document as the sentiment of the document. Corpus-based methods are often built upon machine learning models. Pang et al. (2002) compare the performance of three commonly used machine learning models (Naive Bayes, Maximum Entropy and SVM). Gamon (2004) shows that introducing deeper linguistic features into SVM can help to improve the performance. The interested readers are referred to (Pang and Lee, 2008) for a comprehensive review of sentiment classification.

2.2 Cross-Lingual Sentiment Classification

Cross-lingual sentiment classification, which aims to conduct sentiment classification in the target language (e.g. Chinese) with labeled data in the source language (e.g. English), has been extensively studied in the very recent years. The basic idea is to explore the abundant labeled sentiment data in source language to alleviate the shortage of labeled data in the target language.

Most existing work relies on machine translation engines to directly adapt labeled data from the source language to target language. Wan (2009) proposes to use ensemble method to train better Chinese sentiment classification model on English labeled data and their Chinese translation. English Labeled data are first translated to Chinese, and then two SVM classifiers are trained on English and Chinese labeled data respectively. After that, co-training (Blum and Mitchell, 1998) approach is adopted to leverage Chinese unlabeled data and their English translation to improve the SVM classifier for Chinese sentiment classification. The same idea is used in (Wan, 2008), but the ensemble techniques used are various voting methods and the individual classifiers used are dictionary-based classifiers.

Instead of ensemble methods, Pan et al. (2011) use matrix factorization formulation. They extend Nonnegative Matrix Tri-Factorization model (Li et al., 2009) to bilingual view setting. Their bilingual view is also constructed by using machine translation engines to translate original documents. Prettenhofer and Stein (2011) use machine translation engines in a different way. They generalize Structural Correspondence Learning (Blitzer et al., 2006) to multilingual setting. Instead of using machine translation engines to translate labeled text, the authors use it to construct the word translation oracle for pivot words translation.

Lu et al. (2011) focus on the task of jointly improving the performance of sentiment classification on two languages (e.g. English and Chinese) . the authors use an unlabeled parallel corpus instead of machine translation engines. They assume parallel sentences in the corpus should have the same sentiment polarity. Besides, they assume labeled data in both language are available. They propose a method of training two classifiers based on maximum entropy formulation to maximize their prediction agreement on the parallel corpus. However, this method requires labeled data in *both* the source language and the target language, which are not always readily available.

3 Cross-Lingual Mixture Model for Sentiment Classification

In this section we present the cross-lingual mixture model (**CLMM**) for sentiment classification. We first formalize the task of cross-lingual sentiment classification. Then we describe the CLMM model and present the parameter estimation algorithm for CLMM.

3.1 Cross-lingual Sentiment Classification

Formally, the task we are concerned about is to develop a sentiment classifier for the target language T (e.g. Chinese), given labeled sentiment data D_S in the source language S (e.g. English), unlabeled parallel corpus U of the source language and the target language, and *optional* labeled data D_T in target language T. Aligning with previous work (Wan, 2008; Wan, 2009), we only consider binary sentiment classification scheme (positive or negative) in this paper, but the proposed method can be used in other classification schemes with minor modifications.

3.2 The Cross-Lingual Mixture Model

The basic idea underlying CLMM is to enlarge the vocabulary by learning sentiment words from the parallel corpus. CLMM defines an intuitive generation process as follows. Suppose we are going to generate a positive or negative Chinese sentence, we have two ways of generating words. The first way is to *directly* generate a Chinese word according to the polarity of the sentence. The other way is to first generate an English word with the same polarity and meaning, and then *translate* it to a Chinese word. More formally, CLMM defines a generative mixture model for generating a parallel corpus. The *unobserved* polarities of the unlabeled parallel corpus are modeled as hidden variables, and the *observed* words in parallel corpus are modeled as generated by a set of words generation distributions conditioned on the hidden variables. Given a parallel corpus, we fit CLMM model by maximizing the likelihood of generating this parallel corpus. By maximizing the likelihood, CLMM can estimate words generation probabilities for words unseen in the labeled data but present in the parallel corpus, hence expand the vocabulary. In addition, CLMM can utilize words in both the source language and target language for determining polarity classes of the parallel sentences.

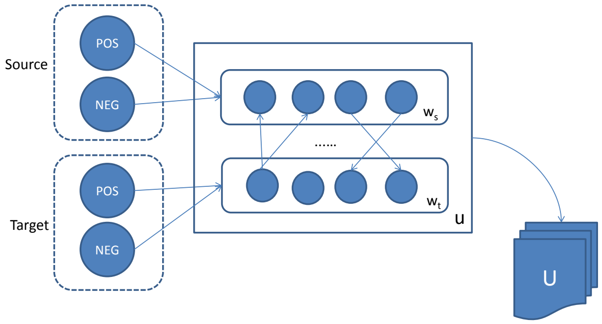

Figure 1: The generation process of the cross-lingual mixture model

Figure 1 illustrates the detailed process of generating words in the source language and target language respectively for the parallel corpus U, from the four mixture components in CLMM. Particularly, for each pair of parallel sentences $u_i \in U$, we generate the words as follows.

1. **Document class generation:** Generating the polarity class.
   (a) Generating a polarity class c_s from a Bernoulli distribution $P_s(C)$.
   (b) Generating a polarity class c_t from a Bernoulli distribution $P_t(C)$
2. **Words generation:** Generating the words
   (a) Generating source language words w_s from a Multinomial distribution $P(w_s|c_s)$
   (b) Generating target language words w_t from a Multinomial distribution $P(w_t|c_t)$
3. **Words projection:** Projecting the words onto the other language
   (a) Projecting the source language words w_s to target language words w_t by word projection probability $P(w_t|w_s)$
   (b) Projecting the target language words w_t to source language words w_s by word projection probability $P(w_s|w_t)$

CLMM finds parameters by using MLE (Maximum Likelihood Estimation). The parameters to be estimated include conditional probabilities of word to class, $P(w_s|c)$ and $P(w_t|c)$, and word projection probabilities, $P(w_s|w_t)$ and $P(w_t|w_s)$. We will describe the log-likelihood function and then show how to estimate the parameters in subsection 3.3. The obtained word-class conditional probability $P(w_t|c)$ can then be used to classify text in the target languages using Bayes Theorem and the Naive Bayes independence assumption.

Formally, we have the following log-likelihood function for a parallel corpus U[2].

$$L(\theta|U) = \sum_{i=1}^{|U_s|}\sum_{j=1}^{|C|}\sum_{s=1}^{|V_s|}\left[N_{si}\log\left(P(w_s|c_j) + P(w_s|w_t)P(w_t|c_j)\right)\right] + \sum_{i=1}^{|U_t|}\sum_{j=1}^{|C|}\sum_{t=1}^{|V_t|}\left[N_{ti}\log\left(P(w_t|c_j) + P(w_t|w_s)P(w_s|c_j)\right)\right] \quad (1)$$

where θ is the model parameters; N_{si} (N_{ti}) is the occurrences of the word w_s (w_t) in document d_i; $|D_s|$ is the number of documents; $|C|$ is the number of class labels; V_s and V_t are the vocabulary in the source language and the vocabulary in the target language.$|U_s|$ and $|U_t|$ are the number of unlabeled sentences in the source language and target language.

Meanwhile, we have the following log-likelihood function for labeled data in the source language D_s.

$$L(\theta|D_s) = \sum_{i=1}^{|D_s|}\sum_{j=1}^{|C|}\sum_{s=1}^{|V_s|} N_{si}\log P(w_s|c_j)\delta_{ij} \quad (2)$$

where $\delta_{ij} = 1$ if the label of d_i is c_j, and 0 otherwise.

In addition, when labeled data in the target language is available, we have the following log-likelihood function.

$$L(\theta|D_t) = \sum_{i=1}^{|D_t|}\sum_{j=1}^{|C|}\sum_{t=1}^{|V_t|} N_{ti}\log P(w_t|c_j)\delta_{ij} \quad (3)$$

Combining the above three likelihood functions together, we have the following likelihood function.

$$L(\theta|D_t, D_s, U) = L(\theta|U) + L(\theta|D_s) + L(\theta|D_t) \quad (4)$$

Note that the third term on the right hand side ($L(\theta|D_t)$) is optional.

[2] For simplicity, we assume the prior distribution $P(C)$ is uniform and drop it from the formulas.

3.3 Parameter Estimation

Instead of estimating word projection probability ($P(w_s|w_t)$ and $P(w_t|w_s)$) and conditional probability of word to class ($P(w_t|c)$ and $P(w_s|c)$) simultaneously in the training procedure, we estimate them separately since the word projection probability stays invariant when estimating other parameters. We estimate word projection probability using word alignment probability generated by the Berkeley aligner (Liang et al., 2006). The word alignment probabilities serves two purposes. First, they connect the corresponding words between the source language and the target language. Second, they adjust the strength of influences between the corresponding words. Figure 2 gives an example of word alignment probability. As is shown, the three words "tour de force" altogether express a positive meaning, while in Chinese the same meaning is expressed with only one word "杰作" (masterpiece). CLMM use word alignment probability to decrease the influences from "杰作" (masterpiece) to "tour", "de" and "force" individually, using the word projection probability (i.e. word alignment probability), which is 0.3 in this case.

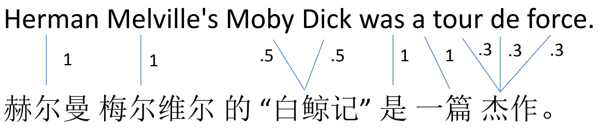
Herman Melville's Moby Dick was a tour de force.
1
1
.5
.5
1
1
.3
.3
.3
赫尔曼 梅尔维尔 的"白鲸记" 是 一篇 杰作。

Figure 2: Word Alignment Probability

We use Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to estimate the conditional probability of word w_s and w_t given class c, $P(w_s|c)$ and $P(w_t|c)$ respectively. We derive the equations for EM algorithm, using notations similar to (Nigam et al., 2000).

In the E-step, the distribution of hidden variables (i.e. class label for unlabeled parallel sentences) is computed according to the following equations.

$$P(c_j|u_{si}) = Z(c_{u_{si}} = c_j) = \frac{\prod_{w_s \in u_{si}}[P(w_s|c_j) + \sum_{P(w_s|w_t)>0} P(w_s|w_t)P(w_t|c_j)]}{\sum_{c_j}\prod_{w_s \in u_{si}}[P(w_s|c_j) + \sum_{P(w_s|w_t)>0} P(w_s|w_t)P(w_t|c_j)]} \tag{5}$$

$$P(c_j|u_{ti}) = Z(c_{u_{ti}} = c_j) = \frac{\prod_{w_t \in u_{ti}}[P(w_t|c_j) + \sum_{P(w_t|w_s)>0} P(w_t|w_s)P(w_s|c_j)]}{\sum_{c_j}\prod_{w_t \in u_{ti}}[P(w_t|c_j) + \sum_{P(w_t|w_s)>0} P(w_t|w_s)P(w_s|c_j)]} \tag{6}$$

where $Z(c_{u_s i} = c_j)$ $\big(Z(c_{u_t i}) = c_j\big)$ is the probability of the source (target) language sentence u_{si} (u_{ti}) in the i-th pair of sentences u_i having class label c_j.

In the M-step, the parameters are computed by the following equations.

$$P(w_s|c_j) = \frac{1 + \sum_{i=1}^{|D_s|} \Lambda_s(i) N_{si} P(c_j|d_i)}{|V| + \sum_{s=1}^{|V_s|} \Lambda(i) N_{si} P(c_j|d_i)} \tag{7}$$

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D_t|} \Lambda_t(i) N_{ti} P(c_j|d_i)}{|V| + \sum_{t=1}^{|V_t|} \Lambda(i) N_{ti} P(c_j|d_i)} \tag{8}$$

where $\Lambda_s(i)$ and $\Lambda_t(i)$ are weighting factor to control the influence of the unlabeled data. We set $\lambda_s(i)$ $\big(\lambda_t(i)\big)$ to λ_s $\big(\lambda_t\big)$ when d_i belongs to unlabeled data, 1 otherwise. When d_i belongs to labeled data, $P(c_j|d_i)$ is 1 when its label is c_j and 0 otherwise. When d_i belongs to unlabeled data, $P(c_j|d_i)$ is computed according to Equation 5 or 6.

4 Experiment

4.1 Experiment Setup and Data Sets

Experiment setup: We conduct experiments on two common cross-lingual sentiment classification settings. In the first setting, no labeled data in the target language are available. This setting has realistic significance, since in some situations we need to quickly develop a sentiment classifier for languages that we do not have labeled data in hand. In this case, we classify text in the target language using only labeled data in the source language. In the second setting, labeled data in the target language are also available. In this case, a more reasonable strategy is to make full use of both labeled data in the source language and target language to develop the sentiment classifier for the target language. In our experiments, we consider English as the source language and Chinese as the target language.

Data sets: For Chinese sentiment classification, we use the same data set described in (Lu et al., 2011). The labeled data sets consist of two English data sets and one Chinese data set. The English data set is from the Multi-Perspective Question Answering (MPQA) corpus (Wiebe et al., 2005) and the NTCIR Opinion Analysis Pilot Task data set (Seki et al., 2008; Seki et al., 2007). The Chinese data set also comes from the NTCIR Opinion Analysis Pilot Task data set. The unlabeled parallel sentences are selected from ISI Chinese-English parallel corpus (Munteanu and Marcu, 2005). Following the description in (Lu et al., 2011), we remove neutral sentences and keep only high confident positive and negative sentences as predicted by a maximum entropy classifier trained on the labeled data. Table 1 shows the statistics for the data sets used in the experiments. We conduct experiments on two data settings: (1) MPQA + NTCIR-CH and (2) NTCIR-EN + NTCIR-CH.

	MPQA	NTCIR-EN	NTCIR-CH
Positive	1,471(30%)	528 (30%)	2,378 (55%)
Negative	3,487(70%)	1,209(70%)	1,916(44%)
Total	4,958	1,737	4,294

Table 1: Statistics about the Data

CLMM includes two hyper-parameters (λ_s and λ_t) controlling the contribution of unlabeled parallel data. Larger weights indicate larger influence from the unlabeled data. We set the hyper-parameters by conducting cross validations on the labeled data. When Chinese labeled data are unavailable, we set λ_t to 1 and λ_s to 0.1, since no Chinese labeled data are used and the contribution of target language to the source language is limited. When Chinese labeled data are available, we set λ_s and λ_t to 0.2.

To prevent long sentences from dominating the parameter estimation, we preprocess the data set by normalizing the length of all sentences to the same constant (Nigam et al., 2000), the average length of the sentences.

4.2 Baseline Methods

For the purpose of comparison, we implement the following baseline methods.

MT-SVM: We translate the English labeled data to Chinese using Google Translate and use the translation results to train the SVM classifier for Chinese.

SVM: We train a SVM classifier on the Chinese labeled data.

MT-Cotrain: This is the co-training based approach described in (Wan, 2009). We summarize the main steps as follows. First, two monolingual SVM classifiers are trained on English labeled data and Chinese data *translated* from English labeled data. Second, the two classifiers make prediction on Chinese unlabeled data and their English translation, respectively. Third, the 100 most confidently predicted English and Chinese sentences are added to the training set and the two monolingual SVM classifiers are re-trained on the expanded training set. The second and the third steps are repeated for 100 times to obtain the final classifiers.

Para-Cotrain: The training process is the same as MT-Cotrain. However, we use a different set of English unlabeled sentences. Instead of using the corresponding machine translation of Chinese unlabeled sentences, we use the parallel English sentences of the Chinese unlabeled sentences.

Joint-Train: This is the state-of-the-art method described in (Lu et al., 2011). This model use English labeled data and Chinese labeled data to obtain initial parameters for two maximum entropy classifiers (for English documents and Chinese documents), and then conduct EM-iterations to update the parameters to gradually improve the agreement of the two monolingual classifiers on the unlabeled parallel sentences.

4.3 Classification Using Only English Labeled Data

The first set of experiments are conducted on using only English labeled data to create the sentiment classifier for Chinese. This is a challenging task, since we do not use any Chinese labeled data. And MPQA and NTCIR data sets are compiled by different groups using different annotation guidelines.

Method	NTCIR-EN NTCIR-CH	MPQA-EN NTCIR-CH
MT-SVM	62.34	54.33
SVM	N/A	N/A
MT-Cotrain	65.13	59.11
Para-Cotrain	67.21	60.71
Joint-Train	N/A	N/A
CLMM	**70.96**	**71.52**

Table 2: Classification Accuracy Using Only English Labeled Data

Table 2 shows the accuracy of the baseline systems as well as the proposed model (CLMM). As is shown, sentiment classification does not benefit much from the direct machine translation. For NTCIR-EN+NTCIR-CH, the accuracy of MT-SVM is only 62.34%. For MPQA-EN+NTCIR-CH, the accuracy is 54.33%, even lower than a trivial method, which achieves 55.4% by predicting all sentences to be positive. The underlying reason is that the vocabulary coverage in machine translated data is low, therefore the classifier learned from the labeled data is unable to generalize well on the test data. Meanwhile, the accuracy of MT-SVM on NTCIR-EN+NTCIR-CH data set is much better than that on MPQA+NTCIR-CH data set. That is because NTCIR-EN and NTCIR-CH cover similar topics. The other two methods using machine translated data, MT-Cotrain and Para-Cotrain also do not perform very well. This result is reasonable, because the initial Chinese classifier trained on machine translated data (MT-SVM) is relatively weak. We also observe that using a parallel corpus instead of machine translations can improve classification accuracy. It should be noted that we do not have the result for Joint-Train model in this setting, since it requires both English labeled data and Chinese labeled data.

4.4 Classification Using English and Chinese Labeled Data

The second set of experiments are conducted on using both English labeled data and Chinese labeled data to develop the Chinese sentiment classifier. We conduct 5-fold cross validations on Chinese labeled data. We use the same baseline methods as described in Section 4.2, but we use *natural* Chinese sentences instead of *translated* Chinese sentences as labeled data in MT-Cotrain and Para-Cotrain. Table 3 shows the accuracy of baseline systems as well as CLMM.

Method	NTCIR-EN	MPQA-EN
	NTCIR-CH	NTCIR-CH
MT-SVM	62.34	54.33
SVM	80.58	80.58
MT-Cotrain	82.28	80.93
Para-Cotrain	82.35	82.18
Joint-Train	83.11	83.42
CLMM	82.73	83.02

Table 3: Classification Accuracy Using English and Chinese Labeled Data

As is seen, SVM performs significantly better than MT-SVM. One reason is that we use natural Chinese labeled data instead of translated Chinese labeled data. Another reason is that we use 5-fold cross validations in this setting, while the previous setting is an *open test* setting. In this setting, SVM is a strong baseline with 80.6% accuracy. Nevertheless, all three methods which leverage an unlabeled parallel corpus, namely Para-Cotrain, Joint-Train and CLMM, still show big improvements over the SVM baseline. Their results are comparable and all achieve state-of-the-art accuracy of about 83%, but in terms of training speed, CLMM is the fastest method (Table 4). Similar to the previous setting, We also have the same observation that using a parallel corpus is better than using translations.

Method	Iterations	Total Time
Para-Cotrain	100	6 hours
Joint-Train	10	55 seconds
CLMM	10	30 seconds

Table 4: Training Speed Comparison

4.5 The Influence of Unlabeled Parallel Data

We investigate how the size of the unlabeled parallel data affects the sentiment classification in this subsection. We vary the number of sentences in the unlabeled parallel from 2,000 to 20,000. We use only English labeled data in this experiment, since this more directly reflects the effectiveness of each model in utilizing unlabeled parallel data. From Figure 3 and Figure 4, we can see that when more unlabeled parallel data are added, the accuracy of CLMM consistently improves. The performance of CLMM is remarkably superior than Para-Cotrain and MT-Cotrain. When we have 10,000 parallel sentences, the accuracy of CLMM on the two data sets quickly increases to 68.77% and 68.91%, respectively. By contrast, we observe that the performance of Para-Cotrain and MT-Cotrain is able to obtain accuracy improvement only after about 10,000 sentences are added. The reason is that the two methods use machine translated labeled data to create initial Chinese classifiers. As is depicted in Table 2, these classifiers are relatively weak. As a result, in the initial iterations of co-training based methods, the predictions made by the Chinese classifiers are inaccurate, and co-training based methods need to see more parallel

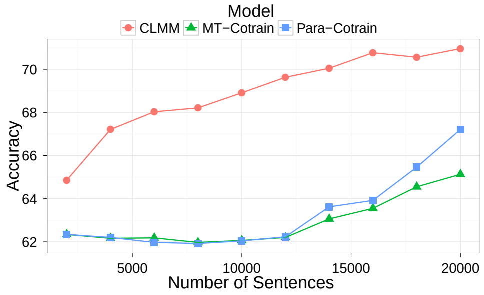

Figure 3: Accuracy with different size of unlabeled data for NTICR-EN+NTCIR-CH

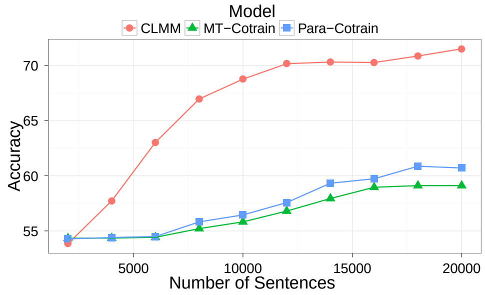

Figure 4: Accuracy with different size of unlabeled data for MPQA+NTCIR-CH

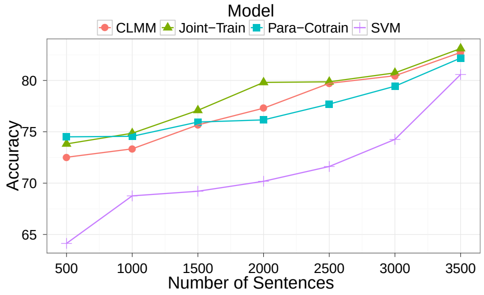

Figure 5: Accuracy with different size of labeled data for NTCIR-EN+NTCIR-CH

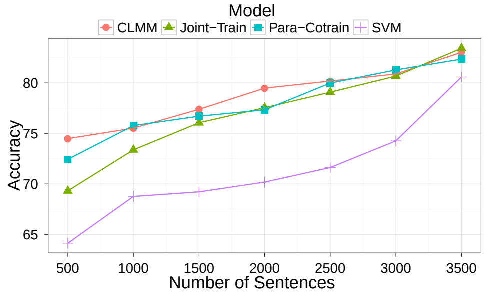

Figure 6: Accuracy with different size of labeled data for MPQA+NTCIR-CH

sentences to refine the initial classifiers.

4.6 The Influence of Chinese Labeled Data

In this subsection, we investigate how the size of the Chinese labeled data affects the sentiment classification. As is shown in Figure 5 and Figure 6, when only 500 labeled sentences are used, CLMM is capable of achieving 72.52% and 74.48% in accuracy on the two data sets, obtaining 10% and 8% improvements over the SVM baseline, respectively. This indicates that our method leverages the unlabeled data effectively. When more sentences are used, CLMM consistently shows further improvement in accuracy. Para-Cotrain and Joint-Train show similar trends. When 3500 labeled sentences are used, SVM achieves 80.58%, a relatively high accuracy for sentiment classification. However, CLMM and the other two models can still gain improvements. This further demonstrates the advantages of expanding vocabulary using bilingual parallel data.

5 Conclusion and Future Work

In this paper, we propose a cross-lingual mixture model (CLMM) to tackle the problem of cross-lingual sentiment classification. This method has two advantages over the existing methods. First, the proposed model can learn previously unseen sentiment words from large unlabeled data, which are not covered by the limited vocabulary in machine translation of the labeled data. Second, CLMM can effectively utilize unlabeled parallel data regardless of whether labeled data in the target language are used or not. Extensive experiments suggest that CLMM consistently improve classification accuracy in both settings. In the future, we will work on leveraging parallel sentences and word alignments for other tasks in sentiment analysis, such as building multilingual sentiment lexicons.

Acknowledgment We thank Bin Lu and Lei Wang for their help. This research was partly supported by National High Technology Research and Development Program of China (863 Program) (No. 2012AA011101) and National Natural Science Foundation of China (No.91024009, No.60973053)

References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, page 120–128.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, page 92–100.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, page 241–249.

Arthur Dempster, Nan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 1–38.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for Cross-Lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 429–433, Portland, Oregon, USA, June. Association for Computational Linguistics.

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841.

Mingqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 168–177.

Tao Li, Yi Zhang, and Vikas Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, page 244–252, Suntec, Singapore, August. Association for Computational Linguistics.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, page 104–111.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, page 320–330.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2):103–134.

Junfeng Pan, Gui-Rong Xue, Yong Yu, and Yang Wang. 2011. Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. *Advances in Knowledge Discovery and Data Mining*, page 289–300.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, page 79–86.

Peter Prettenhofer and Benno Stein. 2011. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):13.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of NTCIR-6 Workshop Meeting*, page 265–278.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. 2008. Overview of multilingual opinion analysis task at NTCIR-7. In *Proc. of the Seventh NTCIR Workshop*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based methods for sentiment analysis. *Comput. Linguist.*, page to appear.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 417–424.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

Taras Zagibalov and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, page 1073–1080.