Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews

Xiaojun Wan*
Peking University

The lack of reliable Chinese sentiment resources limits research progress on Chinese sentiment classification. However, there are many freely available English sentiment resources on the Web. This article focuses on the problem of cross-lingual sentiment classification, which leverages only available English resources for Chinese sentiment classification. We first investigate several basic methods (including lexicon-based methods and corpus-based methods) for cross-lingual sentiment classification by simply leveraging machine translation services to eliminate the language gap, and then propose a bilingual co-training approach to make use of both the English view and the Chinese view based on additional unlabeled Chinese data. Experimental results on two test sets show the effectiveness of the proposed approach, which can outperform basic methods and transductive methods.

1. Introduction

Sentiment classification is the task of identifying the sentiment polarity of a given text, which is traditionally categorized as either positive or negative. In recent years, sentiment classification has drawn much attention in the natural language processing (NLP) field and it has many useful applications, such as opinion mining and summarization (Liu, Hu, and Cheng 2005; Ku, Liang, and Chen 2006; Titov and McDonald 2008).

To date, a variety of lexicon-based and corpus-based methods have been developed for sentiment classification. The lexicon-based methods rely heavily on a sentiment lexicon containing positive terms and negative terms. The corpus-based methods rely heavily on an annotated corpus for training a sentiment classifier. The sentiment lexicon and corpus are considered the most valuable resources for the sentiment classification task. However, such resources in different languages are rather unbalanced. Because most previous work focuses on English sentiment classification, many annotated sentiment lexica and corpora for English sentiment classification in various domains are freely available on the Web. However, the annotated resources for sentiment classification in many other languages are not abundant and it is time-consuming to manually label a rich and reliable sentiment lexicon or corpus in those languages. The challenge before us is leveraging rich English resources for sentiment classification in other languages. In this study, we focus on the problem of English-to-Chinese cross-lingual sentiment classification, leveraging only English sentiment resources for sentiment classification of Chinese product reviews, without using any Chinese sentiment resources.

* Institute of Computer Science and Technology, The MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China. E-mail: wanxiaojun@icst.pku.edu.cn.

Submission received: 17 March 2010; revised submission received: 21 December 2010; accepted for publication: 29 January 2011.

Note that this problem is not only defined for Chinese sentiment classification, but also for various sentiment analysis tasks in other different languages. The proposed approach in this study can also be applied for generic cross-lingual text categorization tasks.

Pilot studies have been performed to make use of English resources for subjectivity classification in Romanian (Mihalcea, Banea, and Wiebe 2007; Banea et al. 2008), and the methods are very straightforward. First, they use machine translation for translating resources (such as a lexicon or corpus) between Romanian and English, and then they employ the lexicon-based or corpus-based method for subjectivity classification in either Romanian or English. Similar experiments have been performed for subjectivity classification in Spanish (Banea et al. 2008). However, our empirical study shows that sentiment classification performance using these methods is far from satisfactory because the machine translation quality is not very good according to the recent NIST open machine translation evaluation results, and thus a language gap between the original language and the translated language still exists.

In this study, we first investigate several basic methods for cross-lingual sentiment classification, and then propose a bilingual co-training approach to improve the accuracy of corpus-based polarity classification of Chinese product reviews. Unlabeled Chinese reviews can be fully leveraged in the proposed approach. First, machine translation services are used to translate English training reviews into Chinese reviews and also translate Chinese test reviews and additional unlabeled reviews into English reviews. Then, we can view the classification problem in two different ways: the Chinese view with only Chinese features and the English view with only English features. We then use the co-training approach to make full use of the two redundant views of features. The SVM classifier (Joachims 2002) is adopted as the basic classifier in the proposed approach.

Three machine translation services (*Google Translate*, *Yahoo Babel Fish*, and *Microsoft Bing Translate*) are used for review translation in the experiments. The experimental results on two test sets show that the proposed approach based on any machine translation service can outperform a few popular baselines, including advanced transductive methods. We also find that the balanced growth of the positive and negative instances at each iteration in the co-training algorithm is very important for the success of the algorithm.

The rest of this article is organized as follows: Section 2 discusses related work. Section 3 introduces several basic methods. The proposed co-training approach is described in detail in Section 4. Sections 5 and 6 present the evaluation set-up and results, respectively. Lastly, we conclude this article and discuss future work in Section 7.

2. Related Work

2.1 Sentiment Classification

Sentiment classification can be performed on words, sentences, or documents. In this article we focus on document-level sentiment classification, and research in this area has followed a lexicon-based (i.e., rule-based) or a corpus-based (i.e., classification-based) approach.

Lexicon-based methods involve deriving a sentiment measure for text based on sentiment lexica. Turney (2002) predicts the sentiment orientation of a review as the average semantic orientation of the phrases in the review that contain adjectives or adverbs, which is known as the semantic orientation method. Kim and Hovy (2004) build

three models to assign a sentiment category to a given sentence by combining the individual sentiments of sentiment-bearing words. Kanayama, Nasukawa, and Watanabe (2004) use the technique of deep language analysis for machine translation to extract sentiment units in text documents. Kennedy and Inkpen (2006) determine the sentiment of a customer review by counting positive and negative terms and taking into account contextual valence shifters, such as negations and intensifiers. Devitt and Ahmad (2007) explore a computable metric of positive or negative polarity in financial news text.

Corpus-based methods consider the sentiment analysis task as a classification task and they use a labeled corpus to train a sentiment classifier. Since the work of Pang, Lee, and Vaithyanathan (2002), various classification models and linguistic features have been proposed to improve classification performance (Mullen and Collier 2004; Pang and Lee 2004; Read 2005; Wilson, Wiebe, and Hoffmann 2005). More recently, McDonald et al. (2007) investigate a structured model for jointly classifying the sentiment of a text at varying levels of granularity. Blitzer, Dredze, and Pereira (2007) investigate domain adaptation for sentiment classifiers, focusing on on-line reviews for different types of products. Andreevskaia and Bergler (2008) present a new system consisting of the ensemble of a corpus-based classifier and a lexicon-based classifier with precision-based vote weighting. A non-negative matrix tri-factorization approach has been proposed for sentiment classification, which learns from lexical prior knowledge in the form of domain-independent sentiment-laden terms in conjunction with domain-dependent unlabeled data and a few labeled data (Li, Zhang, and Sindhwani 2009). Dasgupta and Ng (2009) propose a semi-supervised approach to sentiment classification where they first use spectral techniques to mine the unambiguous reviews and then exploit them to classify the ambiguous reviews by a novel combination of active learning, transductive learning, and ensemble learning.

Chinese sentiment analysis has also been studied (Li and Sun 2007) and most such work uses similar lexicon-based or corpus-based methods for Chinese sentiment classification.

To date, several pilot studies have been performed to leverage rich English resources for sentiment analysis in other languages. Standard naive Bayes and SVM classifiers have been applied for subjectivity classification in Romanian and Spanish (Mihalcea, Banea, and Wiebe 2007; Banea et al. 2008), and the results show that automatic translation is a feasible alternative for the construction of resources and tools for subjectivity analysis in a new target language. Wan (2008) focuses on leveraging both Chinese and English lexica to improve Chinese sentiment analysis by using lexicon-based methods. Wei and Pal (2010) apply structural correspondence learning (SCL) to minimize the noise introduced by machine translations. In this study, we focus on developing novel approaches to improve the corpus-based method for cross-lingual sentiment classification of Chinese product reviews.

2.2 Cross-Domain Text Classification

Cross-domain text classification can be considered as a more general task than cross-lingual sentiment classification. In this task, the labeled and unlabeled data come from different domains and their underlying distributions are often different from each other, which violates the basic assumption of traditional supervised learning.

To date, many semi-supervised learning algorithms have been developed for addressing the cross-domain text classification problem by transferring knowledge across domains, and such algorithms include Transductive SVM (Joachims 1999), EM (Nigam et al. 2000), EM-based naive Bayes classifier (Dai et al. 2007a), Topic-bridged PLSA (Xue

et al. 2008), Co-Clustering–based classification (Dai et al. 2007b), and the two-stage approach (Jiang and Zhai 2007). Dai et al. (2007b) use co-clustering as a bridge to propagate the class structure and knowledge from the in-domain to the out-of-domain. Jiang and Zhai (2007) look for a set of features generalizable across domains at the first generalization stage, and then pick up useful features specific to the target domain at the second adaptation stage. Daumé III and Marcu (2006) introduce a statistical formulation of this problem in terms of a simple mixture model. In recent years, a few methods/algorithms have been proposed for cross-domain sentiment classification, including structural correspondence learning (Blitzer, Dredze, and Pereira 2007), cross-domain graph ranking (Wu et al. 2009), and spectral feature alignment (Pan et al. 2010).

Moreover, several previous studies focus on the problem of cross-lingual text classification, which can be considered a special case of cross-domain text classification. Bel, Koster, and Villegas (2003) empirically investigate three translation strategies for cross-lingual text categorization: document translation, terminology translation, and profile-based translation. A few novel models have been proposed to address the problem—for example, the EM-based algorithm (Rigutini, Maggini, and Liu 2005), the information bottleneck approach (Ling et al. 2008), multilingual domain models (Gliozzo and Strapparava 2005), and the structural correspondence learning approach (Prettenhofer and Stein 2010; Wei and Pal 2010). Shi et al. (2010) introduce a method to transfer classification knowledge across languages by translating the model features and using an EM algorithm. The most recent related work includes multilingual text categorization based on multi-view learning (Amini, Usunier, and Goutte 2009; Amini and Goutte 2010). To the best of our knowledge, co-training has not yet been investigated for cross-domain or cross-lingual text classification.

3. The Basic Methods

A straightforward method for cross-lingual sentiment classification is to use machine translation for transferring lexica or corpora of reviews between English and Chinese, and then apply the lexicon-based or corpus-based method for sentiment classification in either the English or Chinese language. Therefore, the basic methods consist of two main steps: resource translation and sentiment classification. According to different translation directions and classification methods, four basic methods are introduced as follows.

3.1 Lexicon-Based Method in English Language: LEX(EN)

This method first translates Chinese reviews into English reviews, and then identifies the sentiment polarity of the translated English reviews based on English sentiment lexica, as illustrated in Figure 1.

For any specific language, we employ the semantic-oriented approach used in Wan (2008) to compute the semantic orientation value of a review. The unsupervised approach is quite straightforward and it makes use of the following sentiment lexica: **positive Lexicon (Positive_Dic)** containing terms expressing positive polarity, **Negative Lexicon (Negative_Dic)** containing terms expressing negative polarity, **Negation Lexicon (Negation_Dic)** containing terms that are used to reverse the semantic polarity of a particular term, and **Intensifier Lexicon (Intensifier_Dic)** containing terms that are used to change the degree to which a term is positive or negative. The semantic orientation value for a review is computed by summing the polarity values of all terms in the review, making use of both the word polarity defined in the positive and negative lexica and the contextual valence shifters defined in the negation and intensifier lexica.

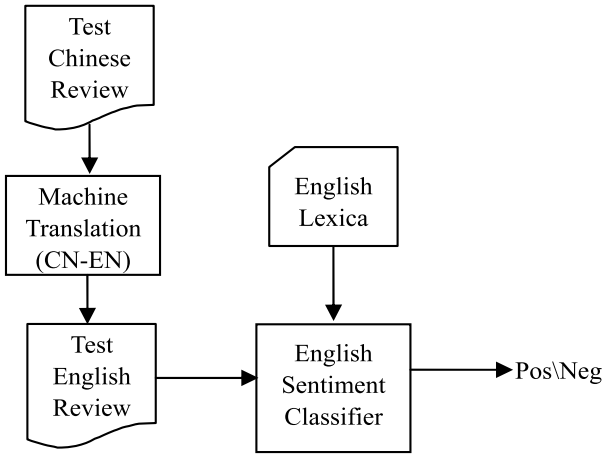

Figure 1
Framework of LEX(EN).

For example, given a review of *the image quality is not good*, although the term *good* is a positive term, the use of the negation term *not* reverses the polarity orientation value, and the overall polarity orientation value of the review is negative. Given a review of *the image quality is very good*, the use of the intensifier term *very* intensifies the polarity orientation value of *good*, and the overall polarity orientation value of the review is positive. In our study, the scope of a negation or intensifier term is simply determined by using a distance window of two words. We do not use a parser to determine the scope because the parsing results for the translated reviews are not reliable.

Finally, if the semantic orientation value of a review is less than 0, the review is labeled as negative; otherwise, the review is labeled as positive.

3.2 Lexicon-Based Method in Chinese Language: LEX(CN)

This method first translates English sentiment lexica into Chinese lexica, and then identifies the sentiment polarity of Chinese reviews based on the translated Chinese lexica, as illustrated in Figure 2.

After we retrieve the four translated Chinese lexica, we apply the algorithm for semantic orientation value computation used in Wan (2008) to predict the polarity orientation of the Chinese reviews. Each Chinese review is first segmented into Chinese terms/words by using our in-house conditional random field (CRF)–based Chinese word segmentation tool, and then the polarity orientation value for the Chinese review is computed by summing the polarity values of all terms in the review. The terms defined in the negation lexicon are used to reverse the polarity values of the nearby Chinese terms, and the terms defined in the intensifier lexicon are used to intensify the polarity values of the nearby Chinese terms. The scope of a negation or intensifier term is also simply determined by using a distance window of two words.

3.3 Corpus-Based Method in English Language: SVM(EN)

As illustrated in Figure 3, we first learn a classifier based on labeled English reviews, and then translate test Chinese reviews into English reviews. Lastly, we use the classifier to classify the translated English reviews. In this study, we use the widely used SVM

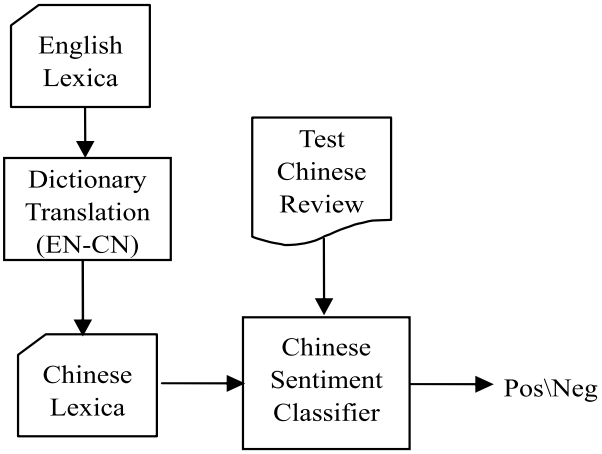

Figure 2
Framework of LEX(CN).

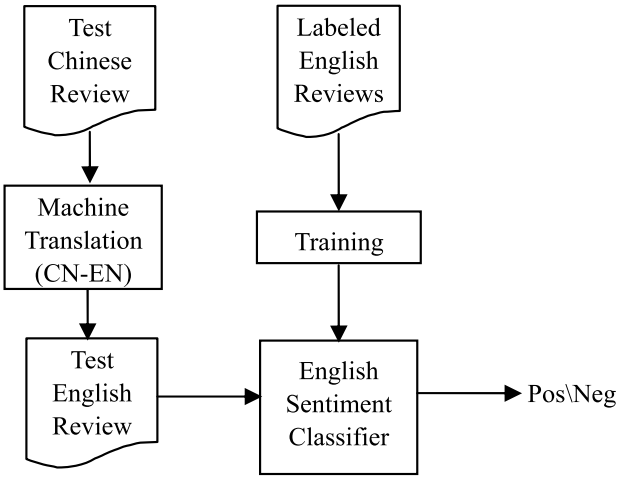

Figure 3
Framework of SVM(EN).

classifier for classification. We also use a transductive variant of the SVM classifier for making use of unlabeled Chinese reviews, which will be described in Section 5.5. All English unigrams and bigrams are used as features, and the feature weight is simply set to term frequency.[1] Finally, the sign of the prediction value of the classifier indicates the polarity orientation of the review.

3.4 Corpus-Based Method in Chinese Language: SVM(CN)

As illustrated in Figure 4, we first translate labeled English reviews into Chinese reviews, and then learn a classifier based on the translated Chinese reviews with labels.

1 Term frequency performs better than TK/IDF by our empirical analysis.

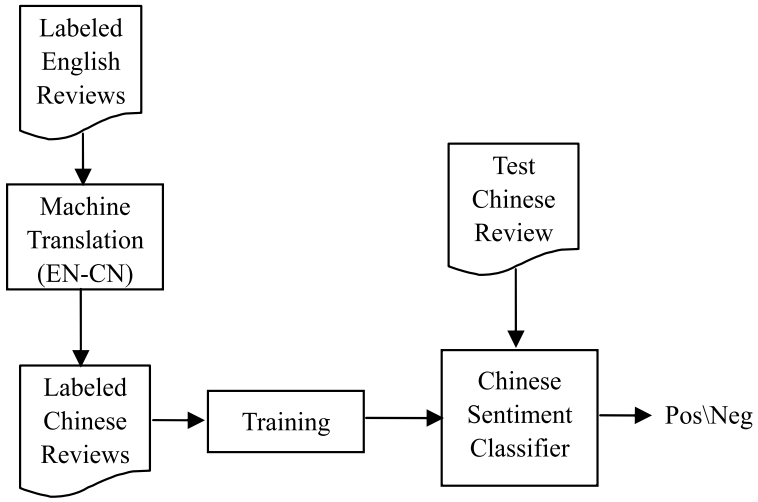

Figure 4
Framework of SVM(CN).

Lastly, we use the classifier to classify test Chinese reviews. We also use the SVM classifier (and a transductive variant) for classification, and all Chinese unigrams and bigrams are used as features.[2]

4. The Bilingual Co-Training Method

4.1 Overview

These two basic corpus-based methods have been used in Banea et al. (2008) for Romanian subjectivity analysis. As shown in our later experiments, the two methods do not perform well for Chinese sentiment classification, because the term distributions in the original reviews and the translated reviews are different. One reason is attributed to machine translation. Because current machine translation services cannot accurately translate reviews, it is inevitable that they bring errors into the translated texts. Moreover, it may happen that different terms are used to express the same meaning in the original texts and the translated texts, because each machine translation service uses particular resources and corpora for model building. The other reason is attributed to inherent domain difference. The review sets in different languages are generally in very different domains, because they are written by different users in different countries, and the writing styles, lengths, and term usages of the reviews are very different.

In order to address this problem, we propose to use the co-training approach to make use of some amounts of unlabeled Chinese reviews to improve the classification accuracy. The co-training approach can make full use of both the English features and the Chinese features in a unified framework. The framework of the proposed approach is illustrated in Figure 5.

The framework consists of a training phase and a classification phase. In the training phase, the input is the labeled English reviews and some amount of unlabeled Chinese

2 For Chinese text, a unigram refers to a Chinese word and a bigram refers to two adjacent Chinese words.

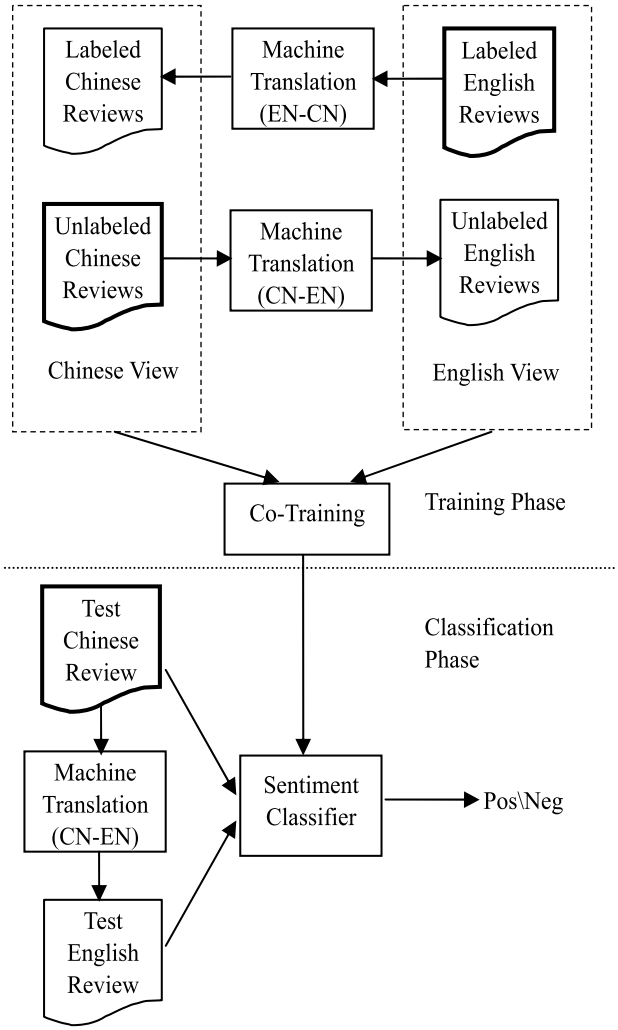

Figure 5
Framework of the bilingual co-training approach.

reviews.[3] The labeled English reviews are translated into labeled Chinese reviews and the unlabeled Chinese reviews are translated into unlabeled English reviews by using machine translation services. Therefore, each review is associated with an English version and a Chinese version. The English features (i.e., all English unigrams and bigrams) and the Chinese features (i.e., all Chinese unigrams and bigrams) for each review are considered two different and redundant views of the review. The co-training algorithm is then applied to learn two classifiers, and finally the two classifiers are combined into a single sentiment classifier. In the classification phase, each unlabeled Chinese review

3 The unlabeled Chinese reviews used for co-training do not include the unlabeled Chinese reviews for testing, that is, the Chinese reviews for testing are blind to the training phase.

for testing is first translated into an English review, and then the learned classifier is applied to predict the polarity orientation of the review as either positive or negative.

4.2 The Co-Training Algorithm

The co-training algorithm (Blum and Mitchell 1998) is a bootstrapping method; it starts with a set of labeled data, and increases the amount of annotated data using some amount of unlabeled data in an incremental way. One important aspect of co-training is that two conditionally independent views are required for co-training to work, but the independence assumption can be relaxed. In the past, co-training has been successfully applied to statistical parsing (Sarkar 2001), reference resolution (Ng and Cardie 2003), part-of-speech tagging (Clark, Curran, and Osborne 2003), word sense disambiguation (Mihalcea 2004), and e-mail classification (Kiritchenko and Matwin 2001). Co-training has not yet been used for cross-domain or cross-lingual text categorization, however.

The intuition behind the co-training algorithm is that if one classifier can confidently predict the class of an example, it can provide one more training example for the other classifier. Of course, if this example happens to be easily classified by the first classifier, it does not mean that this example will be easily classified by the second classifier, so the second classifier will get useful information to improve itself, and vice versa (Kiritchenko and Matwin 2001).

In the context of cross-lingual sentiment classification, each labeled English review or unlabeled Chinese review has two sets of features: English features and Chinese features. Here, a review is used to indicate both its Chinese version and its English version, unless stated otherwise. Now we describe the details of the co-training algorithm.

The notations in the co-training algorithm are as follows:

- F_{en} and F_{cn} are redundantly sufficient sets of features, where F_{en} represents the English features, and F_{cn} represents the Chinese features;
- L is a set of labeled training reviews;
- U is a set of unlabeled reviews;
- C is a basic classification algorithm, and C_{en} and C_{cn} represent two component classifiers based on C; and
- p and n are positive integer numbers.

The following steps loop for I iterations in the co-training algorithm:

(1) Learn the first classifier C_{en} from L based on F_{en}.

(2) Use C_{en} to label reviews in U based on F_{en}.

(3) Choose p positive and n negative most confidently predicted reviews (E_{en}) from U.

(4) Learn the second classifier C_{cn} from L based on F_{cn}.

(5) Use C_{cn} to label reviews in U based on F_{cn}.

(6) Choose p positive and n negative most confidently predicted reviews (E_{cn}) from U.

(7) Remove the reviews in $E_{en} \bigcup E_{cn}$ from U. Note that the examples with conflicting labels are not included in $E_{en} \bigcup E_{cn}$. In other words, if an example is in both E_{en} and E_{cn}, but the labels for the example are conflicting, the example will be excluded from $E_{en} \bigcup E_{cn}$.

(8) Add the reviews in $E_{en} \bigcup E_{cn}$ with the corresponding labels to L.

From the point of view of the algorithm, the unlabeled reviews are added to the model during the bootstrapping phase, and only for these reviews are the labels obtained using an average of the normalized prediction values determined by the component classifiers. In the algorithm, p and n are two parameters controlling the growth size in the labeled data. At each iteration, at most $2(p + n)$ reviews are added into L. The two parameters also maintain the class distribution in the labeled data by balancing the parameter values of p and n at each iteration. If p is similar to n, the growth is called **balanced growth**, otherwise the growth is called **unbalanced growth**. Note that the co-training algorithm used in this study differs slightly from the original co-training algorithm in that the original co-training algorithm is dependent on the sequence of the two component classifiers, whereas our co-training algorithm is independent of the classifier sequence. Moreover, each classifier in our co-training algorithm not only makes use of a few examples confidently predicted by the other classifier, but also makes use of a few examples confidently predicted by itself.

In the co-training algorithm, a basic classification algorithm is required to construct C_{en} and C_{cn}. Typical text classifiers include Support Vector Machine (SVM), naive Bayes (NB), Maximum Entropy (ME), K-Nearest Neighbor (KNN), and so forth. In this study, we adopt the widely used SVM classifier (Joachims 2002), as in the basic corpus-based methods. Viewing input data as two sets of vectors in a feature space, SVM constructs a separating hyperplane in this space by maximizing the margin between the two data sets. The output value of the SVM classifier for a review indicates the confidence level of the review's classification. The sentiment polarity of a review is indicated by the sign of the prediction value. Note that we use all unigrams and bigrams in each language as features and the feature weight is simply set to term frequency.

In the training phase, the co-training algorithm learns two separate classifiers: C_{en} and C_{cn}. Therefore, in the classification phase, we can obtain two prediction values for a test review. We normalize the prediction values into $[-1, 1]$ by dividing the maximum absolute value. Finally, the average of the normalized values is used as the overall prediction value of the review.[4]

Several theoretical studies have been performed on co-training in the machine learning field. Blum and Mitchell (1998) prove that co-training can be successful if the two sufficient and redundant views are conditionally independent of each other. Abney (2002) shows that weak dependence between the two views can also guarantee successful co-training. Balcan, Blum, and Yang (2005) prove that a weaker assumption called ε-expansion is sufficient for iterative co-training to succeed. Wang and Zhou (2010) view the co-training process as a combinative label propagation over two views, and they

4 Though this method of combining scores is unprincipled due to the fact that the scores themselves are not calibrated, we found it worked well in practice.

provide the sufficient and necessary condition for co-training to succeed. As can be seen, the assumption about the dependence between the two views is much relaxed, which can guarantee that although the English features and the Chinese features are not conditionally independent of each other, the use of the two views for co-training is acceptable. In the extreme case, if both classifiers agree on all the unlabeled data, labeling the data does not create new information, and thus the co-training algorithm will not work at all. We will show in the experiments that the English classifier and the Chinese classifier disagree on many unlabeled examples, which can also guarantee the success of the co-training approach.

5. Evaluation Set-up

5.1 English Sentiment Resources

The basic LEX(EN) and LEX(CN) methods require English sentiment lexica. In this study, we collected and used the following popular and publicly available English sentiment lexica,[5] without any further filtering and labeling:

Positive_Dic^{en}: 2,718 English positive terms (e.g., *amazing*, *gorgeous*) were collected from a feature file[6] containing the subjectivity clues used in the work (Wilson, Wiebe, and Hoffmann 2005; Wilson et al. 2005). The clues in this file were collected from a number of sources. Some were culled from manually developed resources (e.g., *General Inquirer*[7] [Stone et al. 1966]). Others were identified automatically using both annotated and unannotated data. A majority of the clues were collected as part of the work reported in Riloff and Wiebe (2003).

Negative_Dic^{en}: 4,910 English negative terms (e.g., *boring*, *idiot*) were collected from the same file.

Negation_Dic^{en}: 88 negation terms (e.g., *never*, *lack*) were collected from a feature file[8] used in Wilson, Wiebe, and Hoffmann 2005; Wilson et al. 2005.

Intensifier_Dic^{en}: 244 intensifier terms (e.g., *very*, *absolutely*) were collected from a feature file[9] used in Wilson, Wiebe, and Hoffmann 2005; Wilson et al. 2005.

We then used a large English-to-Chinese dictionary (LDC_EC_DIC2.0[10]) with 110,834 entries for projecting English lexica into Chinese lexica via term-to-term translation. If an English term corresponds to multiple Chinese terms, we simply use the first Chinese term for translation because the first one is the dominant translation.

The basic SVM(EN), SVM(CN) methods and the co-training method require a labeled English sentiment corpus. In this study, we used the following popular English sentiment corpus:

Training Set (Labeled English Reviews): There are many labeled English corpora available on the Web; we used the corpus constructed for multi-domain sentiment classification (Blitzer, Drezde, and Perreira 2007),[11] because the corpus was large-scale

5 In this study, we focus on using a few popular English resources for comparative study, instead of trying to collect and use all available resources.
6 The file *subjclueslen1-HLTEMNLP05.tff* can be downloaded from `http://www.cs.pitt.edu/mpqa/`.
7 `http://www.wjh.harvard.edu/~inquirer/homecat.htm`.
8 The file *valenceshifters.tff* can be downloaded from `http://www.cs.pitt.edu/mpqa/`.
9 The file *intensifiers2.tff* can be downloaded from `http://www.cs.pitt.edu/mpqa/`.
10 `http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm`.
11 `http://www.cis.upenn.edu/~mdredze/datasets/sentiment/`.

and it was within similar domains to the test set. The data set consists of 8,000 Amazon product reviews (4,000 positive reviews + 4,000 negative reviews) for four different product types: books, DVDs, electronics, and kitchen appliances.

5.2 Chinese Review Sets

The following two data sets were collected and used as test sets in the experiments:

IT168 Test Set (Labeled Chinese Reviews from IT168): In order to assess the performance of the proposed approach, we collected and labeled 886 product reviews (451 positive reviews + 435 negative reviews) from the popular Chinese IT product Web site IT168.[12] The reviews focus on such products as mp3 players, mobile phones, digital cameras, and laptop computers.

360BUY Test Set (Labeled Chinese Reviews from 360BUY): In addition, we collected and labeled 930 product reviews (560 positive reviews + 370 negative reviews) from another popular Chinese online shopping Web site (360BUY).[13] The reviews focus on such products as electronics and furniture.

For these two test sets, two subjects participated in the annotation procedure. The polarity tags of the reviews were first annotated by one subject and then checked by the other subject. The conflicts were resolved by discussion.

We also collected the following unlabeled Chinese review set for transductive methods and the co-training method:

Unlabeled Set (Unlabeled Chinese Reviews): We downloaded 2,000 additional Chinese product reviews from IT168 and used the reviews as the unlabeled set.[14] The unlabeled set and the IT168 test set were in the same domain and had similar underlying feature distributions, but the unlabeled set and the 360BUY test set may be in different domains.

Note that the training set and the unlabeled set were used in the training phase, and the test set was blind to the training phase. All these data sets are available upon request.

5.3 Review Translation

For all the data sets described in Sections 5.1 and 5.2, each Chinese review was translated into an English review, and each English review was translated into a Chinese review.[15] Therefore, each review has two views: the English view and the Chinese view. A review is represented by both its English view and its Chinese view.

Fortunately, machine translation techniques have been well developed in the NLP field (Lopez 2008), though the translation performance is far from satisfactory. A few commercial machine translation services can be publicly accessed, for example, *Google Translate* (GoogleTranslate),[16] *Yahoo Babel Fish* (YahooTranslate),[17] and

12 http://www.it168.com.
13 http://www.360buy.com.
14 Only 1,000 unlabeled Chinese reviews were used in Wan (2009).
15 We used the recently updated MT services for machine translation; we believe that the translation results are better than those in Wan (2009).
16 http://translate.google.com/translate_t.
17 http://babelfish.yahoo.com/translate_txt.

Microsoft Bing Translate (MicrosoftTranslate).[18] The three MT systems are considered to be state-of-the-art commercial machine translation systems, and all three MT systems provide Chinese-to-English and English-to-Chinese translation services. However, it is not easy to accurately compare the translation performance of the three MT systems because the three systems are updated frequently. In the experiments, we adopt all of them for both English-to-Chinese translation and Chinese-to-English translation.

Here are two examples of Chinese reviews and the corresponding translated English reviews, including the manual translation results (HumanTranslate):

Positive Example: 功能还是比较多的，用的也不错。
HumanTranslate: *More functional and well used.*
GoogleTranslate: *Or more functional, well used.*
YahooTranslate: *The function are quite many, with also good.*
MicrosoftTranslate: *or more, with gamers.*

Negative example: 价格过高，最好*3000*以下。
HumanTranslate: *The price is too high and it should be below 3000.*
GoogleTranslate: *Prices were too high, preferably below 3000.*
YahooTranslate: *The price is excessively high, best below 3000.*
MicrosoftTranslate: *The best price too high. 3000 following.*

Here are two examples of English reviews and the corresponding translated Chinese reviews:

Positive Example: *The book arrived as expected and was in great shape. Thanks*
HumanTranslate: 这本书按预期到了，外形美观。谢谢。
GoogleTranslate: 这本书抵达预期，并在伟大的形状。谢谢
YahooTranslate: 书到达了正如所料并且在了不起的形状。 谢谢
MicrosoftTranslate: 本书到达按预期方式，并在很大的形状。 谢谢你

Negative example: *We had to return this item for a refund. It arrived and never worked.*
HumanTranslate: 我们不得不退回该项目，要求退款。它到了之后从未工作。
GoogleTranslate: 我们不得不返回本项目的退款。到达和从未工作。
YahooTranslate: 我们必须退回退款的这个项目。 它到达和未曾运作。
MicrosoftTranslate: 我们不得不返回此项退款。 它到达，从不工作。

5.4 Evaluation Metric

We used the standard precision (*P*), recall (*R*), and F-measure (*F*) to measure the performance of positive and negative classes, and employed the accuracy metric (*Acc*) to measure the overall performance of each system. The metrics are defined in the same way as in generic text categorization tasks.

5.5 Baseline Methods

In the experiments, the proposed co-training approach (CoTrain) is compared with two groups of baseline methods.

18 http://www.microsofttranslator.com/.

The first group includes the following three **monolingual baselines**, which perform sentiment classification of Chinese reviews based on Chinese resources.

BaseCN1: This method is a monolingual baseline for Chinese sentiment classification, and it is lexicon-based. It uses the most popular and publicly available Chinese sentiment lexica[19] for Chinese sentiment classification by applying the same algorithm as LEX(CN) for semantic orientation value computation. The four Chinese lexica were collected as follows:

Positive_Dic^{cn}: 3,730 Chinese positive terms (e.g., 好看/*good-looking*, 吉祥/*lucky*) were collected from the Chinese Vocabulary for Sentiment Analysis (VSA)[20] released by HOWNET.

Negative_Dic^{cn}: 3,116 Chinese negative terms (e.g., 昂贵/*expensive*, 笨头笨脑/*clumsy*) were collected from the Chinese VSA released by HOWNET.

Negation_Dic^{cn}: 13 negation terms (e.g., 不是/*be not*, 欠缺/*be lack in*) were collected from related papers.

Intensifier_Dic^{cn}: 148 intensifier terms (e.g., 完完全全/*totally*, 万分/*extremely*) were collected from the Chinese VSA released by HOWNET.

BaseCN2: This method is a monolingual baseline based on supervised classification in the Chinese language. We downloaded a very large number of product reviews and their associated tags from the popular Chinese online shopping Web site Amazon China.[21] The data set consists of 45,898 positive reviews and 24,146 negative reviews. The reviews are about various products such as consumer electronics, mobile phones, digital products, books, and so on. The polarity tag of each review was automatically judged by the number of the user-assigned stars attached to the review. If the star number is equal to or less than two, the review is labeled as negative, and otherwise the review is labeled as positive. We adopt the inductive SVM classifier and use the large corpus for training. Finally, the classifier is applied for sentiment classification of Chinese reviews.

BaseCN3: This method is a monolingual baseline based on transductive classification in the Chinese language. We adopt the transductive SVM classifier, and use the automatically crawled corpus used in BaseCN2 and the unlabeled Chinese reviews for training.

In addition, we perform five-fold cross-validation on each test set. The method first randomly partitions the original test set into five subsets. During each cross-validation process, a single subset is retained as the validation set, and the remaining four subsets are used as the training set. The inductive SVM classifier is trained on the training set and tested on the validation set. The cross-validation process is then repeated five times, and the five results are then averaged. Note that the results are produced by five different classification models that are different from other methods. The performance of the cross-validation method can be seen as an upper bound for the monolingual methods, because the method uses human-labeled Chinese reviews for training, and moreover, the training reviews and the test reviews come from the same Web site and thus they are in the same domain. The method is denoted **UpperBound(CrossValidation)**.

The second group includes the following eleven **cross-lingual baselines**, which perform sentiment classification of Chinese reviews based only on English resources.

19 Very few Chinese sentiment lexica are freely available on the Web.

20 http://www.keenage.com/html/e_index.html.

21 http://www.amazon.cn.

LEX(CN): This method uses the lexicon-based method in the Chinese language, as described in Section 3.2.

LEX(EN): This method uses the lexicon-based method in the English language, as described in Section 3.1.

SVM(CN): This method applies the inductive SVM with only Chinese features for sentiment classification in the Chinese view, as described in Section 3.4. Only English-to-Chinese translation is needed. The inductive SVM learner aims to build a decision function based on the training set, and the unlabeled set is not used by this method.

SVM(EN): This method applies the inductive SVM with only English features for sentiment classification in the English view, as described in Section 3.3. Only Chinese-to-English translation is needed. The unlabeled set is not used by this method.

SVM(ENCN): This method combines the results of SVM(EN) and SVM(CN) by averaging the prediction values of the two SVM classifiers in the same way as in the co-training approach.

TSVM(CN): This method applies the transductive SVM with only Chinese features for sentiment classification in the Chinese view. Only English-to-Chinese translation is needed. The unlabeled set is used by this method. Transductive SVM has been widely used to treat partially labeled data in semi-supervised learning. Different from inductive SVM, it can leverage unlabeled data and try to separate both labeled and unlabeled data with a maximum margin. For more details, refer to Joachims (1999).

TSVM(EN): This method applies the transductive SVM with only English features for sentiment classification in the English view. Only Chinese-to-English translation is needed. The unlabeled set is used by this method.

TSVM(ENCN): This method combines the results of TSVM(EN) and TSVM(CN) by averaging the prediction values of the two TSVM classifiers.

SelfTrain(CN): This method uses the self-training algorithm in Mihalcea (2004) and the unlabeled set for sentiment classification in the Chinese view. The algorithm is a single-view weakly supervised algorithm. It starts with a set of labeled reviews, and builds a SVM classifier. The classifier is then applied to the unlabeled reviews, and the p positive and n negative most confidently predicted reviews are added to the labeled set. The classifier is then retrained on the new labeled set. The process continues for I iterations. The parameters p, n, and I are defined in the same way as for the co-training algorithm.

SelfTrain(EN): This method uses the self-training algorithm and the unlabeled set for sentiment classification in the English view.

SelfTrain(ENCN): This method combines the results of SelfTrain(EN) and SelfTrain(CN) by averaging the prediction values of the two self-training classifiers. It is noteworthy that SelfTrain(ENCN) differs from CoTrain in that there is no mutual interaction between the English component classifier and the Chinese component classifier in SelfTrain(ENCN).

Note that the three transductive methods and the three self-training methods are strong baselines because they have been widely used for improving classification accuracy by leveraging additional unlabeled examples. We use the SVMLight toolkit[22] with the linear kernel and default parameter values for both inductive SVM classification and transductive SVM classification.

Though feature selection methods (e.g., Document Frequency [DF], Information Gain [IG], and Mutual Information [MI]) can be used for dimension reduction, we

22 `http://svmlight.joachims.org`.

Table 1
Results for monolingual methods on the IT168 test set.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
BaseCN1	0.681	0.929	0.786	0.882	0.549	0.677	0.743
BaseCN2	0.716	**0.945**	0.815	**0.914**	0.611	0.733	0.781
BaseCN3	**0.724**	0.942	**0.819**	0.913	**0.628**	**0.744**	**0.788**
UpperBound (**CrossValidation**)	0.909	0.867	0.888	0.868	0.910	0.889	0.888

use all the features in the experiments for comparative analysis because there is no significant performance improvement after applying the feature selection techniques in our empirical study.

6. Evaluation Results

6.1 Method Comparison

In the experiments, we first compare the proposed co-training approach with the baseline methods. The parameter values for CoTrain and SelfTrain are set as $I = 80$ and $p = n = 5$. The three parameters are empirically set by considering the total number (i.e., 2,000) of the unlabeled Chinese reviews. In our empirical study, the proposed approach can perform well with a wide range of parameter values, which will be shown later.

Tables 1 and 5 show the results for three monolingual baselines and the upper bound on the two test sets, respectively. Tables 2 through 4 show the comparison results for the cross-lingual methods based on the three machine translation services on the IT168 test set, respectively. Tables 6 through 8 show the comparison results for the cross-lingual methods based on the three machine translation services on the 360BUY test set, respectively. Note that we also present the classification results for the two component classifiers (Chinese component classifier C_{cn} and English component classifier C_{en}) of our proposed co-training approach. Tables 9 and 10 show the results of significance tests between CoTrain and the baseline methods on the two test sets, respectively. We adopt the sign test as a significance test because it is widely used in the field of text categorization (Yang and Liu 1999). In particular, we use an on-line service[23] for performing sign tests in the experiments. The p-values for sign tests are presented; the performance difference between CoTrain and a baseline method is statistically significant at a 95% level if the p-value is smaller than 0.05.

As can be seen in Tables 1 through 8, no matter which machine translation service is used, the proposed co-training approach (CoTrain) outperforms all baseline methods on the overall accuracy metric and most other metrics on the two test sets. In particular, on the IT168 test set, the best accuracy is achieved by CoTrain with GoogleTranslate, and on the 360BUY test set, the best result is achieved by CoTrain with YahooTranslate. Even the two component classifiers in CoTrain can perform as well as or better than the baseline methods. As can be seen from Tables 9 and 10, the performance difference

23 http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html.

Table 2
Comparison results for cross-lingual methods on the IT168 test set with GoogleTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.615	0.772	0.684	0.678	0.499	0.575	0.638
LEX(EN)	0.770	**0.914**	0.836	**0.889**	0.717	0.794	0.817
SVM(CN)	0.735	0.843	0.785	0.808	0.685	0.741	0.765
SVM(EN)	0.737	0.800	0.767	0.773	0.703	0.736	0.753
SVM(ENCN)	0.758	0.856	0.804	0.828	0.717	0.768	0.788
TSVM(CN)	0.735	0.732	0.733	0.723	0.726	0.725	0.729
TSVM(EN)	0.816	0.847	0.831	0.835	0.802	0.818	0.825
TSVM(ENCN)	0.817	0.840	0.828	0.829	0.805	0.817	0.823
SelfTrain(CN)	0.742	0.747	0.745	0.736	0.731	0.734	0.739
SelfTrain(EN)	0.801	0.847	0.823	0.831	0.782	0.806	0.815
SelfTrain(ENCN)	0.804	0.836	0.820	0.823	0.789	0.805	0.813
C_{cn} **in CoTrain**	0.828	0.834	0.831	0.826	0.821	0.824	0.827
C_{en} **in CoTrain**	0.833	0.863	0.847	0.852	0.821	0.836	0.842
CoTrain	**0.858**	0.882	**0.870**	0.874	**0.848**	**0.861**	**0.866**

between CoTrain and any baseline method is always statistically significant when GoogleTranslate or YahooTranslate is used for machine translation. We can also see that the performance difference between CoTrain and any baseline method is almost always statistically significant when MicrosoftTranslate is used for machine translation, except for the TSVM(CN) baseline on the IT168 test set and the TSVM(ENCN) and TSVM(CN) baselines on the 360BUY test set.

Table 3
Comparison results for cross-lingual methods on the IT168 test set with YahooTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.615	0.772	0.684	0.678	0.499	0.575	0.638
LEX(EN)	0.759	**0.874**	0.812	**0.845**	0.713	0.773	0.795
SVM(CN)	0.728	0.814	0.769	0.780	0.685	0.729	0.751
SVM(EN)	0.699	0.736	0.717	0.710	0.671	0.690	0.704
SVM(ENCN)	0.735	0.792	0.762	0.765	0.703	0.733	0.748
TSVM(CN)	0.762	0.809	0.785	0.789	0.738	0.762	0.774
TSVM(EN)	0.820	0.776	0.797	0.780	0.823	0.801	0.799
TSVM(ENCN)	0.816	0.818	0.817	0.811	0.809	0.810	0.814
SelfTrain(CN)	0.733	0.767	0.750	0.746	0.710	0.728	0.739
SelfTrain(EN)	0.799	0.827	0.813	0.814	0.784	0.799	0.806
SelfTrain(ENCN)	0.788	0.823	0.805	0.807	0.770	0.788	0.797
C_{cn} **in CoTrain**	0.807	0.836	0.821	0.823	0.793	0.808	0.815
C_{en} **in CoTrain**	**0.831**	0.831	0.831	0.825	**0.825**	0.825	0.828
CoTrain	0.828	0.845	**0.836**	0.836	0.818	**0.827**	**0.832**

Table 4
Comparison results for cross-lingual methods on the IT168 test set with MicrosoftTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.615	0.772	0.684	0.678	0.499	0.575	0.638
LEX(EN)	0.744	0.909	0.818	**0.878**	0.676	0.764	0.795
SVM(CN)	0.669	**0.925**	0.777	0.871	0.526	0.656	0.729
SVM(EN)	0.702	0.855	0.771	0.806	0.623	0.703	0.742
SVM(ENCN)	0.694	0.905	0.785	0.856	0.586	0.696	0.748
TSVM(CN)	0.801	0.865	0.832	0.847	0.777	0.811	0.822
TSVM(EN)	0.789	0.745	0.766	0.750	0.793	0.771	0.769
TSVM(ENCN)	0.812	0.854	0.832	0.840	0.795	0.817	0.825
SelfTrain(CN)	0.759	0.851	0.803	0.824	0.720	0.768	0.787
SelfTrain(EN)	0.785	0.776	0.780	0.770	0.779	0.775	0.778
SelfTrain(ENCN)	0.802	0.860	0.830	0.843	0.779	0.810	0.821
C_{cn} **in CoTrain**	0.818	0.858	0.838	0.845	0.802	0.823	0.831
C_{en} **in CoTrain**	0.803	0.820	0.811	0.809	0.791	0.800	0.806
CoTrain	**0.829**	0.874	**0.851**	0.861	**0.814**	**0.837**	**0.844**

Among the baselines, the best baseline is TSVM(ENCN). Actually, TSVM(ENCN) is very similar to CoTrain, and it combines the results of two classifiers in the same way. However, the co-training approach can train two more effective component classifiers than those used in TSVM(ENCN). As suggested from the tables, the accuracy values of the component classifiers (C_{cn} and C_{en}) in CoTrain are almost always higher than those of the corresponding TSVM(CN) and TSVM(EN), based on any machine translation service. The reason is that TSVM(CN) and TSVM(EN) leverage the unlabeled data independently, while the two component classifiers in the co-training approach leverage the unlabeled data in a mutual way, and more useful knowledge in the unlabeled data can be incorporated into the co-training approach. We can also see that the co-training approach outperforms the baseline self-training approach, which further demonstrates the great importance of the mutual influence of the two views during the bootstrapping phase.

As mentioned in Section 4.2, the English classifier and the Chinese classifier in the co-training approach are required to disagree on some unlabeled examples, and we

Table 5
Results for monolingual methods on the 360BUY test set.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
BaseCN1	0.747	0.845	0.793	0.707	**0.568**	0.630	0.734
BaseCN2	0.752	**0.927**	0.830	0.829	0.538	0.652	0.772
BaseCN3	**0.761**	**0.927**	**0.836**	**0.835**	0.559	**0.670**	**0.781**
UpperBound (**CrossValidation**)	0.880	0.946	0.912	0.909	0.805	0.854	0.890

Table 6
Comparison results for cross-lingual methods on the 360BUY test set with GoogleTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.714	**0.895**	0.794	0.741	0.457	0.565	0.720
LEX(EN)	0.761	0.864	0.809	0.741	0.589	0.657	0.755
SVM(CN)	0.791	0.825	0.808	0.717	0.670	0.693	0.763
SVM(EN)	0.765	0.795	0.779	0.670	0.630	0.649	0.729
SVM(ENCN)	0.801	0.834	0.817	0.732	0.686	0.709	0.775
TSVM(CN)	0.784	0.877	0.828	0.773	0.635	0.697	0.781
TSVM(EN)	0.824	0.818	0.821	0.727	0.735	0.731	0.785
TSVM(ENCN)	0.826	0.857	0.841	0.771	0.727	0.748	0.805
SelfTrain(CN)	0.791	0.861	0.825	0.757	0.657	0.703	0.780
SelfTrain(EN)	0.797	0.813	0.805	0.708	0.686	0.697	0.762
SelfTrain(ENCN)	0.814	0.850	0.831	0.757	0.705	0.730	0.792
C_{cn} **in CoTrain**	**0.849**	0.877	0.863	0.804	**0.765**	**0.784**	0.832
C_{en} **in CoTrain**	0.816	0.834	0.825	0.740	0.716	0.728	0.787
CoTrain	0.846	0.884	**0.865**	**0.812**	0.757	0.783	**0.833**

show the disagreement ratio between the two classifiers at each iteration in Figure 6. At each iteration in the co-training algorithm, we use the two classifiers to predict the polarity tags of the unlabeled examples, respectively. The disagreement ratio is computed by dividing the number of the consistently predicted examples by the size of the unlabeled set. We can see from the figure that the disagreement ratio is always higher than 20%, which guarantees the success of the co-training approach.

Table 7
Comparison results for cross-lingual methods on the 360BUY test set with YahooTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.714	**0.895**	0.794	0.741	0.457	0.565	0.720
LEX(EN)	0.783	0.846	0.814	0.735	0.646	0.688	0.767
SVM(CN)	0.798	0.816	0.807	0.711	0.686	0.699	0.765
SVM(EN)	0.757	0.786	0.771	0.656	0.619	0.637	0.719
SVM(ENCN)	0.793	0.834	0.813	0.727	0.670	0.698	0.769
TSVM(CN)	0.806	0.854	0.829	0.757	0.689	0.721	0.788
TSVM(EN)	0.822	0.839	0.830	0.749	0.724	0.736	0.794
TSVM(ENCN)	0.832	0.877	0.854	0.797	0.732	0.763	0.819
SelfTrain(CN)	0.804	0.848	0.825	0.749	0.686	0.717	0.784
SelfTrain(EN)	0.804	0.830	0.817	0.730	0.695	0.712	0.776
SelfTrain(ENCN)	0.820	0.861	0.840	0.772	0.714	0.742	0.802
C_{cn} **in CoTrain**	0.857	0.857	0.857	0.784	0.784	0.784	0.828
C_{en} **in CoTrain**	0.830	0.845	0.837	0.758	0.738	0.748	0.802
CoTrain	**0.869**	0.866	**0.868**	**0.798**	**0.803**	**0.801**	**0.841**

Table 8
Comparison results for cross-lingual methods on the 360BUY test set with MicrosoftTranslate.

	Positive			Negative			Total
Method	*P*	*R*	*F*	*P*	*R*	*F*	*Acc*
LEX(CN)	0.714	0.895	0.794	0.741	0.457	0.565	0.720
LEX(EN)	0.749	0.864	0.803	0.732	0.562	0.636	0.744
SVM(CN)	0.730	**0.904**	0.808	0.772	0.495	0.603	0.741
SVM(EN)	0.727	0.795	0.759	0.638	0.549	0.590	0.697
SVM(ENCN)	0.737	0.896	0.809	0.767	0.516	0.617	0.745
TSVM(CN)	0.845	0.875	0.860	0.800	0.757	0.778	0.828
TSVM(EN)	0.810	0.700	0.751	0.623	0.751	0.681	0.720
TSVM(ENCN)	**0.856**	0.839	0.848	0.764	**0.786**	0.775	0.818
SelfTrain(CN)	0.787	0.879	0.830	0.777	0.641	0.702	0.784
SelfTrain(EN)	0.792	0.739	0.765	0.641	0.705	0.672	0.726
SelfTrain(ENCN)	0.823	0.870	0.845	0.784	0.716	0.749	0.809
C_{cn} **in CoTrain**	0.834	0.886	0.859	0.809	0.732	0.769	0.825
C_{en} **in CoTrain**	0.800	0.779	0.789	0.678	0.705	0.691	0.749
CoTrain	0.843	0.884	**0.863**	**0.810**	0.751	**0.780**	**0.831**

For the three lexicon-based baseline methods (BaseCN1, LEX(CN), and LEX(EN)), the LEX(EN) method performs better than the BaseCN1 method, but the LEX(CN) method performs worse than the BaseCN1 method. The reason is that the sentiment lexica used in LEX(CN) are automatically translated from the original English lexica, and the translation is very inaccurate because there are no contexts or clues for sense disambiguation during the translation process.

For the three monolingual baseline methods (BaseCN1, BaseCN2, and BaseCN3), the BaseCN2 and BaseCN3 methods outperform the BaseCN1 method. However, the BaseCN2 and BaseCN3 methods cannot outperform the strong cross-lingual baseline

Table 9
p-values for sign tests between the results of CoTrain and baseline methods on the IT168 test set.

	GoogleTranslate	YahooTranslate	MicrosoftTranslate
CoTrain vs. BaseCN1	2.85E-12	1.04E-06	1.82E-08
CoTrain vs. BaseCN2	1.8E-07	0.00257	0.000182
CoTrain vs. BaseCN3	1.27E-06	0.00922	0.000765
CoTrain vs. LEX(CN)	6.09E-29	3.72E-21	1.61E-24
CoTrain vs. LEX(EN)	0.0018	0.0276	0.00329
CoTrain vs. SVM(CN)	1.26E-13	6.45E-10	2.7E-14
CoTrain vs. SVM(EN)	2.15E-18	1.1E-17	3.46E-13
CoTrain vs. SVM(ENCN)	2.08E-13	8.13E-12	1.05E-12
CoTrain vs. TSVM(CN)	1.2E-19	1.07E-06	0.0624
CoTrain vs. TSVM(EN)	1.41E-05	0.00311	2.17E-08
CoTrain vs. TSVM(ENCN)	1.37E-08	0.0113	0.0396
CoTrain vs. SelfTrain(CN)	7.07E-18	2.79E-11	6.53E-07
CoTrain vs. SelfTrain(EN)	1.01E-07	0.0192	1.35E-07
CoTrain vs. SelfTrain(ENCN)	6.4E-11	0.000194	0.000508

Table 10
p-values for sign tests between the results of CoTrain and baseline methods on the 360BUY test set.

	GoogleTranslate	YahooTranslate	MicrosoftTranslate
CoTrain vs. BaseCN1	6.01E-08	4.45E-09	2.1E-07
CoTrain vs. BaseCN2	0.000144	4.77E-05	0.000247
CoTrain vs. BaseCN3	0.0009	0.000287	0.00139
CoTrain vs. LEX(CN)	9.53E-10	7.15E-11	1.17E-09
CoTrain vs. LEX(EN)	1.87E-05	1.64E-05	8.92E-07
CoTrain vs. SVM(CN)	5.7E-08	2.91E-09	2.27E-11
CoTrain vs. SVM(EN)	3.74E-15	5.77E-17	1.18E-20
CoTrain vs. SVM(ENCN)	8.07E-09	3.76E-10	1.31E-12
CoTrain vs. TSVM(CN)	2.38E-05	1.28E-05	0.838
CoTrain vs. TSVM(EN)	4.27E-06	2.48E-05	1.78E-14
CoTrain vs. TSVM(ENCN)	0.000306	0.00779	0.0884
CoTrain vs. SelfTrain(CN)	1.08E-05	3.31E-06	5.64E-05
CoTrain vs. SelfTrain(EN)	1.85E-10	2.48E-08	5.17E-14
CoTrain vs. SelfTrain(ENCN)	4.52E-07	2.57E-05	0.00107

methods (e.g., TSVM(ENCN), SelfTrain(ENCN)), because the Chinese training corpus is automatically collected without human checking and thus about 10% of the reviews are mistakenly labeled. Moreover, the corpus is collected from a different Web site, and thus the training set and the test set may be in different domains. We also note that no methods can outperform the monolingual upper bound (the cross-validation method), because it leverages in-domain human-labeled training set for model learning.

Given any machine translation service, the transductive SVM classifiers can almost always outperform the corresponding inductive SVM classifiers on the two test sets. More specifically, the BaseCN3 method outperforms the BaseCN2 method; the TSVM(CN), TSVM(EN), and TSVM(ENCN) methods almost always outperform the SVM(CN), SVM(EN), and SVM(ENCN) methods, respectively, except that TSVM(CN) cannot outperform SVM(CN) on the IT168 test set with GoogleTranslate. In most cases, SelfTrain(CN), SelfTrain(EN), and SelfTrain(ENCN) can outperform the SVM(CN),

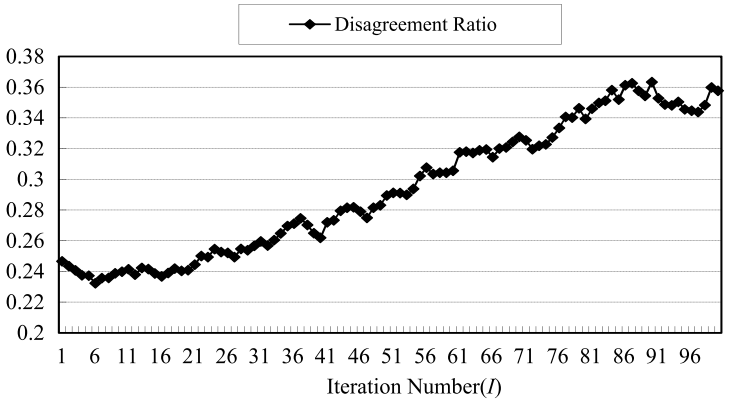

Figure 6
The disagreement ratio between the two component classifiers on the remaining unlabeled examples at each iteration for co-training ($p = n = 5$) with GoogleTranslate.

SVM(EN), and SVM(ENCN) methods, respectively. The results demonstrate that the use of unlabeled reviews is beneficial to the classification task.

Overall, the use of unlabeled data and the combination of English and Chinese views are beneficial to the final classification accuracy, and the co-training approach is more suitable for making use of the unlabeled Chinese reviews than the transductive SVM and the self-training approach.

Moreover, we find that the three machine translation services perform differently on the two test sets, and no particular service can always outperform the other two services on the two test sets. Although machine translation is very important in the proposed methods, the quality of the three machine translation services offers no significant differences.

6.2 Influences of Iteration Number (*I*)

Figures 7 and 8 show the accuracy curves of the co-training approach and two strong baselines (SVM(ENCN) and SelfTrain(ENCN)) with respect to different numbers of iterations on the two test sets with GoogleTranslate, respectively. The parameter values for CoTrain and SelfTrain are set as $p = n = 5$. The iteration number I varies from 1 to 100. When I is set to 1, both the co-training approach and the self-training approach degenerate into SVM(ENCN). The accuracy curves of the component English and Chinese classifiers learned in the co-training approach are also shown in the figures. We omit the very similar figures obtained with YahooTranslate and MicrosoftTranslate.

We can see that the proposed co-training approach (CoTrain) can outperform the two strong baselines after a few iterations. After a large number of iterations, the performance of the co-training approach does not rise any more, because the algorithm runs out of all useful examples in the unlabeled set. The performance finally has a slight decline because some noisy training examples may be selected from the remaining unlabeled set. Fortunately, the proposed approach performs well with a wide range of iteration values.

We can also see that the two component classifiers show a similar trend to the co-training approach. It is encouraging that either the component English classifier or the component Chinese classifier alone can perform better than the strong baselines after a

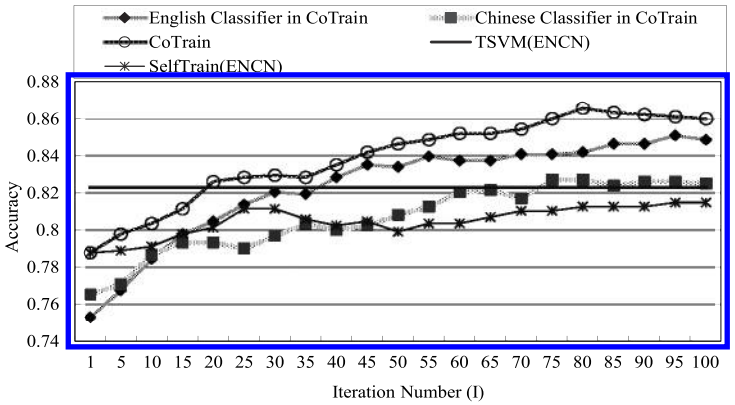

Figure 7
Accuracy vs. number of iterations for co-training and baselines ($p = n = 5$) on the IT168 test set with GoogleTranslate.

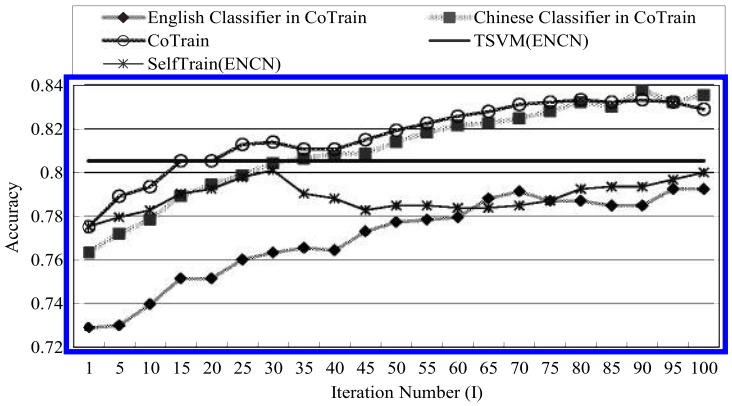

Figure 8
Accuracy vs. number of iterations for co-training and baselines ($p = n = 5$) on the 360BUY test set with GoogleTranslate.

few iterations. The results show that the effectiveness of the co-training approach can be attributed to the effectiveness of its two component classifiers.

6.3 Influences of Growth Size (p, n)

Figures 9 and 10 show how the growth size at each iteration (p positive and n negative confident examples) influences the accuracy of the proposed co-training approach on the two test sets with GoogleTranslate, respectively. In these experiments, we set $p = n$, which is considered a balanced growth. When p differs very much from n, the growth is considered unbalanced. Balanced growth of (2, 2), (5, 5), (10, 10), and (15, 15) examples and unbalanced growth of (1, 5), (5, 1), (1, 10), and (10, 1) examples are compared in the figures. We omit the very similar figures obtained with YahooTranslate and MicrosoftTranslate.

We can see that the performance of the co-training approach with balanced growth can be improved after a few iterations. The performance of the co-training approach with larger $p = n$ will rise more sharply, because the approach can make use of more

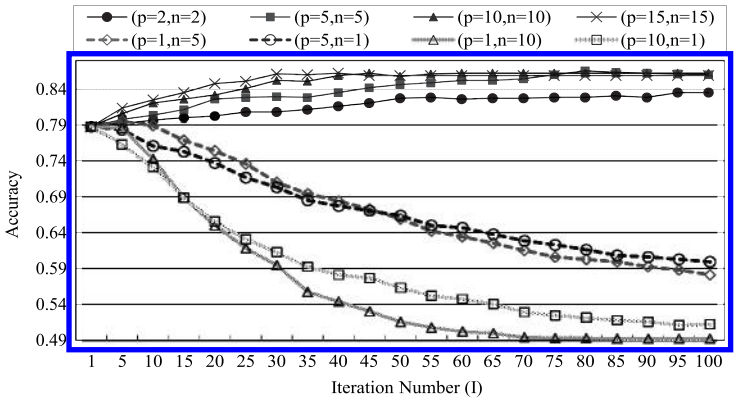

Figure 9
Accuracy vs. different (p, n) for co-training on the IT168 test set with GoogleTranslate.

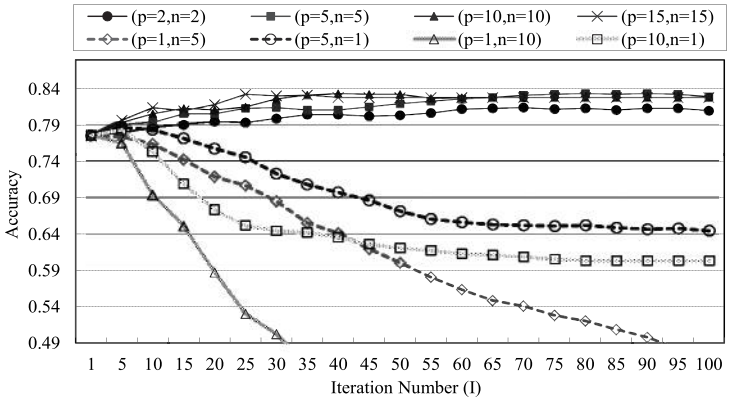

Figure 10
Accuracy vs. different (p, n) for co-training on the 360BUY test set with GoogleTranslate.

selected examples to improve the classifiers at each iteration. Also, the performance of the co-training approach with larger $p = n$ will become stable more quickly, because the approach runs out of the limited examples in the unlabeled set more quickly.

The performance of the co-training approaches with the unbalanced growth always declines quite rapidly, however, because the selected unbalanced examples hurt the performance at each iteration. We also find that the more p is different from n, the faster the performance declines. Actually, in the generic text categorization task, unbalanced training data will lead to poor classification results (Japkowicz and Stephen 2002).

Overall, the growth size has a great impact on the final performance. A balanced growth can lead to performance improvement, but an unbalanced growth can hurt the final performance.

6.4 Influences of Feature Selection

In these experiments, all features (unigrams + bigrams) are used. As mentioned earlier, feature selection techniques are widely used for dimensionality reduction. In this section, we conduct further experiments to investigate the influences of feature selection techniques on the classification results. We use the simple but effective DF for feature selection. Figures 11 and 12 show the comparison results of different feature sizes for the co-training approach and two baselines on the two test sets with GoogleTranslate, respectively. The feature size is measured as the proportion of the selected features against the total features (i.e., 100%), and we select 10%, 25%, and 50% features in the experiments. We omit the very similar figures obtained with YahooTranslate and MicrosoftTranslate.

We can see from the figures that the feature selection technique has a very slight influence on the classification accuracy of each individual method. This can be explained by the fact that sentiment classification is different from topic-based text classification, and the useful feature sets for the two classification tasks are very different. The popular feature selection techniques are helpful for topic-based text classification, but they cannot select good features for sentiment classification. Though the feature selection techniques cannot improve the sentiment classification accuracy significantly, they can reduce the feature size to 10% while not significantly lowering the classification accuracy. The large reduction of feature size can improve system efficiency.

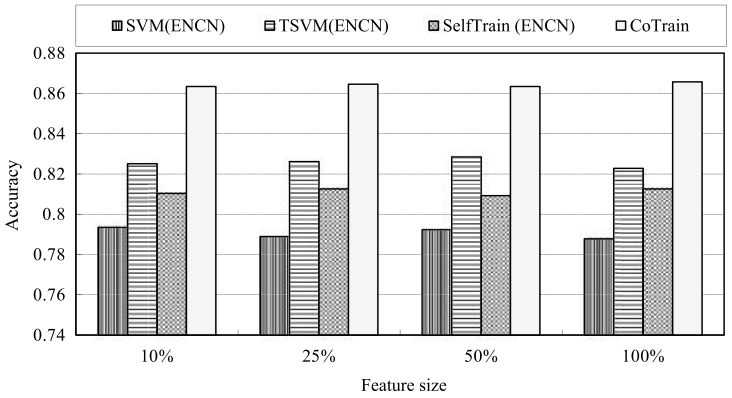

Figure 11
Influences of feature selection on the IT168 test set with GoogleTranslate.

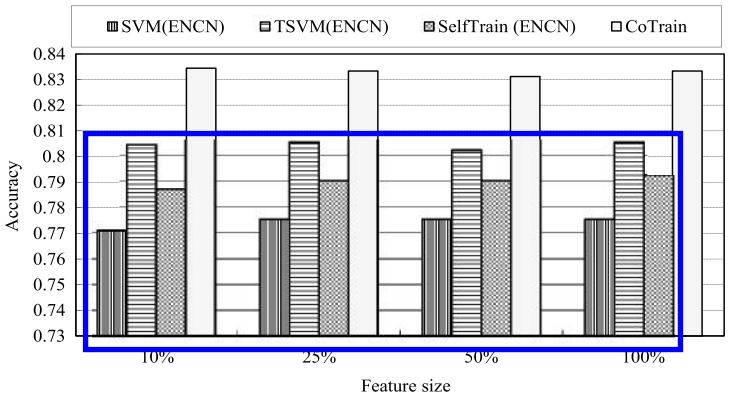

Figure 12
Influences of feature selection on the 360Buy test set with GoogleTranslate.

More importantly we can see that the TSVM and SelfTrain baselines always outperform the inductive SVM baseline, and the co-training approach can always outperform all the three baselines with different feature sizes. The results further demonstrate the effectiveness and robustness of the proposed co-training approach.

6.5 Influences of Different Training Sets

In the experiments, the training set provided by Blitzer, Dredze, and Pereira (2007) is a very balanced set (4,000 positive reviews + 4,000 negative reviews). In this section, we sample the following two training sets from the original set: One training set consists of 4,000 positive reviews and randomly selected 2,000 negative reviews (#pos:#neg=2:1), and the other training set consists of 2,000 randomly selected positive reviews and 4,000 negative reviews (#pos:#neg=1:2). The two sampled training sets are not balanced. The proposed co-training approach is compared with the three strong baselines (SVM(ENCN), TSVM(ENCN), and SelfTrain(ENCN)) on the two training sets. Figures 13 and 14 show the comparison results on the two training sets, respectively. We can see that based on the two training sets, our proposed co-training approach can

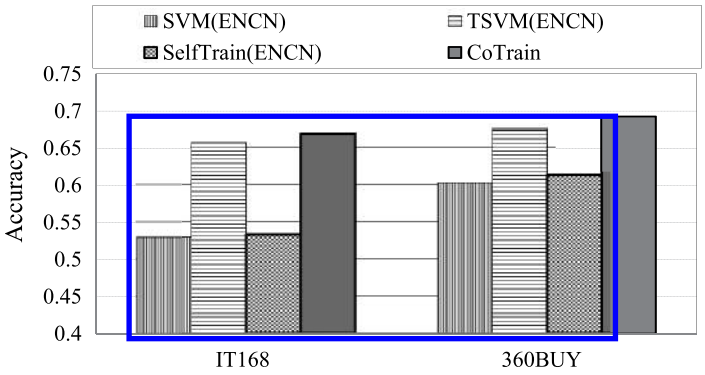

Figure 13
Comparison results based on one sampled training set (#pos:#neg=2:1) with GoogleTranslate.

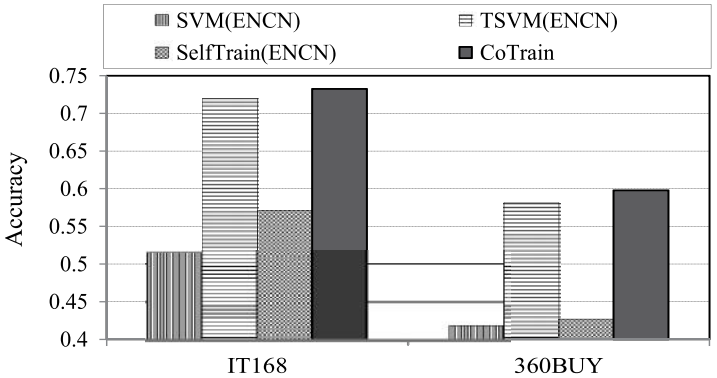

Figure 14
Comparison results based on the second sampled training set (#pos:#neg=1:2) with GoogleTranslate.

consistently outperform all three baselines, which further demonstrates the robustness of our proposed approach.

7. Conclusion and Discussion

In this article, we proposed to use the co-training approach to address the problem of cross-lingual sentiment classification. The approach leverages only labeled English reviews and unlabeled Chinese reviews for Chinese sentiment classification. First, the labeled English reviews are translated into labeled Chinese reviews by using English-to-Chinese machine translation services, and the unlabeled Chinese reviews are translated into unlabeled English reviews by using Chinese-to-English machine translation services. The English view and the Chinese view are considered two redundant views. Then, the co-training algorithm is employed to learn two component classifiers in the two views by mutually helping each other. Finally, given a test Chinese review and its translated English review, the two classifiers are used to obtain two prediction values, and the final polarity tag of the review is decided by the average of the two prediction values.

In the experiments, three machine translation services and two test sets are used for evaluation, and the evaluation results show the overall effectiveness and robustness of the proposed co-training approach. The approach can significantly outperform the lexicon-based baselines, the inductive classification baselines, the transductive classification baselines, and the self-training baselines. We also find that the growth size (i.e., the numbers of positive and negative examples selected in the labeled data) is a very important factor in the proposed approach, which has great influence over the final performance. In particular, a balanced growth leads to performance improvement, but an unbalanced growth hurts the final performance.

Though we focus on English-to-Chinese cross-language sentiment classification in this study, the proposed approach can be easily applied to cross-language sentiment classification in other languages, because the three machine translation services cover many of the most frequently used language pairs. For most western languages, feature extraction is very easy because word segmentation is not required. However, for some Asian languages (e.g., Japanese, Korean), the step of word segmentation is required in order to split a text into words, and thus a word segmentation tool for the specific language is necessary. Fortunately, with the progress of NLP research, word segmentation tools with good performance can be easily obtained for each specific language, and unigram/bigram features can be easily extracted after word segmentation.

The feature distributions of the translated text and the natural text in the same language are still different due to the inaccuracy of the machine translation service and the domain difference between the training set and the test set. In future work, we will try to develop advanced methods to minimize the feature gap in the two review sets. Moreover, we will translate both English and Chinese reviews into a few other languages, and then exploit the multi-view learning techniques for making use of the multiple views in different languages.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 60873155, the Beijing Nova Program under Grant No. 2008B03, and the MOE Program for New Century Excellent Talents in University under Grant No. NCET-08-0006. We are very grateful to the anonymous reviewers for their insightful and constructive comments and suggestions.

References

Abney, Steven P. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 360–367, Philadelphia, PA.

Andreevskaia, Alina and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 290–298, Columbus, OH.

Amini, Massih-Reza and Cyril Goutte. 2010. A co-classification approach to learning from multilingual corpora. *Machine Learning Journal*, 79(1-2):105–121.

Amini, Massih R., Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views—an application to multilingual text categorization. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 28–36, Vancouver.

Balcan, Maria-Florina, Avrim Blum, and Ke Yang. 2005. Co-training and expansion: Towards bridging theory and practice. In *Proceedings of the Nineteenth Annual Conference on Neural Information Processing Systems (NIPS)*, pages 89–96, Vancouver.

Banea, Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 127–135, Waikiki, HI.

Bel, Nuria, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *Proceedings of the Seventh*

European Conference on Research and Advanced Technology for Digital Libraries (ECDL), pages 126–139, Trondheim.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 440–447, Prague.

Blum, Avrim and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, Madison, WI.

Clark, Stephen, James R. Curran, and Miles Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the 2003 Conference on Computational Natural Language Learning (CoNLL)*, pages 49–55, Edmonton.

Dai, Wenyuan, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007a. Transferring naïve Bayes classifiers for text classification. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 540–545, Vancouver.

Dai, Wenyuan, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007b. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 210–219, San Jose, CA.

Dasgupta, Sajib and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP)*, pages 701–709, Suntec.

Daumé III, Hal and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126.

Devitt, Ann and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 984–991, Prague.

Gliozzo, Alfio and Carlo Strapparava. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 9–16, Ann Arbor, MI.

Japkowicz, Nathalie and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5): 429–449.

Jiang, Jing and ChengXiang Zhai. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM)*, pages 401–410, Lisbon.

Joachims, Thorsten. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 200–209, Bled.

Joachims, Thorsten. 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA.

Kanayama, Hiroshi, Tetsuya Nasukawa, and Hideo Watanabe. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 494–500, Geneva.

Kennedy, Alistair and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 1367–1373, Geneva.

Kiritchenko, Svetlana and Stan Matwin. 2001. Email classification with co-training. In *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON)*, pages 192–201, Toronto.

Ku, Lun-Wei, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW)*, pages 100–107, Stanford, CA.

Li, Jun and Maosong Sun. 2007. Experimental study on sentiment classification of Chinese review using machine learning techniques. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 393–400, Beijing.

Li, Tao, Yi Zhang, and Vikas Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP)*, pages 244–252, Suntec.

Ling, Xiao, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can Chinese web pages be classified with English data sources? In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 969–978, Beijing.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web (WWW)*, pages 342–351, Chiba.

Lopez, Adam. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3), pages 1–49.

McDonald, Ryan, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 432–439, Prague.

Mihalcea, Rada. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CONLL)*, pages 33–40, Boston, MA.

Mihalcea, Rada, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 976–983, Prague.

Mullen, Tony and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418, Barcelona.

Ng, Vincent and Claire Cardie. 2003. Weakly supervised natural language learning without redundant views. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 94–101, Edmonton.

Nigam, Kamal, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134.

Pan, Sinno J., Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 751–760, Raleigh, NC.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, PA.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 271–278, Barcelona.

Prettenhofer, Peter and Benno Stein. 2010. Cross-language text classfication using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1118–1127, Uppsala.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48, Ann Arbor, MI.

Rigutini, Leonardo, Marco Maggini, and Bing Liu. 2005. An EM based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 529–535, Compiegne.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 105–112, Stroudsburg, PA.

Sarkar, Anoop. 2001. Applying co-training methods to statistical parsing. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 175–182, Pittsburgh, PA.

Shi, Lei, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1057–1067, Cambridge, MA.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, MA.

Titov, Ivan and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 308–316, Columbus, OH.

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, Philadelphia, PA.

Wan, Xiaojun. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 553–561, Honolulu, HI.

Wan, Xiaojun. 2009. Co-Training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP)*, pages 235–243, Suntec.

Wang, Wei and Zhi-Hua Zhou. 2010. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1135–1142, Haifa.

Wei, Bin and Christopher Pal. 2010. Cross-lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 258–262, Uppsala.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 347–354, Vancouver.

Wilson, Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLP/EMNLP on Interactive Demonstrations*, pages 34–35, Vancouver.

Wu, Qiong, Songbo Tan, Haijun Zhai, Gang Zhang, Miyi Duan, and Xueqi Cheng. 2009. SentiRank: Cross-domain graph ranking for sentiment classification. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 309–314, Milan.

Xue, Gui-Rong, Wenyuan Dai, Qiang Yang, and Yong Yu. 2008. Topic-bridged PLSA for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 627–634, Singapore.

Yang, Yiming and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 42–49, Berkeley, CA.