Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis

Xiaojun Wan
Institute of Compute Science and Technology
Peking University
Beijing 100871, China
wanxiaojun@icst.pku.edu.cn

Abstract

It is a challenging task to identify sentiment polarity of Chinese reviews because the resources for Chinese sentiment analysis are limited. Instead of leveraging only monolingual Chinese knowledge, this study proposes a novel approach to leverage reliable English resources to improve Chinese sentiment analysis. Rather than simply projecting English resources onto Chinese resources, our approach first translates Chinese reviews into English reviews by machine translation services, and then identifies the sentiment polarity of English reviews by directly leveraging English resources. Furthermore, our approach performs sentiment analysis for both Chinese reviews and English reviews, and then uses ensemble methods to combine the individual analysis results. Experimental results on a dataset of 886 Chinese product reviews demonstrate the effectiveness of the proposed approach. The individual analysis of the translated English reviews outperforms the individual analysis of the original Chinese reviews, and the combination of the individual analysis results further improves the performance.

1 Introduction

In recent years, sentiment analysis (including subjective/objective analysis, polarity identification, opinion extraction, etc.) has drawn much attention in the NLP field. In this study, the objective of sentiment analysis is to annotate a given text for polarity orientation (positive/negative). Polarity orientation identification has many useful applications, including opinion summarization (Ku et al., 2006) and sentiment retrieval (Eguchi and Lavrenko, 2006).

To date, most of the research focuses on English and a variety of reliable English resources for sentiment analysis are available, including polarity lexicon, contextual valence shifters, etc. However, the resources for other languages are limited. In particular, few reliable resources are available for Chinese sentiment analysis[1] and it is not a trivial task to manually label reliable Chinese sentiment resources.

Instead of using only the limited Chinese knowledge, this study aims to improve Chinese sentiment analysis by making full use of bilingual knowledge in an unsupervised way, including both Chinese resources and English resources. Generally speaking, there are two unsupervised scenarios for "borrowing" English resources for sentiment analysis in other languages: one is to generate resources in a new language by leveraging on the resources available in English via cross-lingual projections, and then perform sentiment analysis in the English language based on the generated resources, which has been investigated by Mihalcea et al. (2007); the other is to translate the texts in a new language into English texts, and then perform sentiment analysis in the English language, which has not yet been investigated.

In this study, we first translate Chinese reviews into English reviews by using machine translation services, and then identify the sentiment polarity of English reviews by directly leveraging English resources. Furthermore, ensemble methods are employed to combine the individual analysis results in each language (i.e. Chinese and English) in order to obtain improved results. Given machine translation services between the selected target language and English, the proposed approach can be applied to any other languages as well.

Experiments have been performed on a dataset of 886 Chinese product reviews. Two commercial

[1] This study focuses on Simplified Chinese.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 553–561,
Honolulu, October 2008.

machine translation services (i.e. Google Translate and Yahoo Babel Fish) and a baseline dictionary-based system are used for translating Chinese reviews into English reviews. Experimental results show that the analysis of English reviews translated by the commercial translation services outperforms the analysis of original Chinese reviews. Moreover, the analysis performance can be further improved by combining the individual analysis results in different languages. The results also demonstrate that our proposed approach is more effective than the approach that leverages generated Chinese resources.

The rest of this paper is organized as follows: Section 2 introduces related work. The proposed approach is described in detail in Section 3. Section 4 shows the experimental results. Lastly we conclude this paper in Section 5.

2 Related Work

Polarity identification can be performed on word level, sentence level or document level. Related work for word-level polarity identification includes (Hatzivassiloglou and McKeown, 1997; Kim and Hovy. 2004; Takamura et al., 2005; Yao et al. 2006; Kaji and Kitsuregawa, 2007), and related work for sentence-level polarity identification includes (Yu and Hatzivassiloglou, 2003; Kim and Hovy. 2004) Word-level or sentence-level sentiment analysis is not the focus of this paper.

Generally speaking, document-level polarity identification methods can be categorized into unsupervised and supervised.

Unsupervised methods involve deriving a sentiment metric for text without training corpus. Turney (2002) predicates the sentiment orientation of a review by the average semantic orientation of the phrases in the review that contain adjectives or adverbs, which is denoted as the semantic oriented 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. Hiroshi et al. (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.

Supervised methods consider the sentiment analysis task as a classification task and use labeled corpus to train the classifier. Since the work of Pang et al. (2002), various classification models and linguistic features have been proposed to improve the classification performance (Pang and Lee, 2004; Mullen and Collier, 2004; Wilson et al., 2005a; Read, 2005). Most recently, McDonald et al. (2007) investigate a structured model for jointly classifying the sentiment of text at varying levels of granularity. Blitzer et al. (2007) investigate domain adaptation for sentiment classifiers, focusing on online 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.

Research work focusing on Chinese sentiment analysis includes (Tsou et al., 2005; Ye et al., 2006; Li and Sun, 2007; Wang et al., 2007). Such work represents heuristic extensions of the unsupervised or supervised methods for English sentiment analysis.

To date, the most closely related work is Mihalcea et al. (2007), which explores cross-lingual projections to generate subjectivity analysis resources in Romanian by leveraging on the tools and resources available in English. They have investigated two approaches: a lexicon-based approach based on Romanian subjectivity lexicon translated from English lexicon, and a corpus-based approach based on Romanian subjectivity-annotated corpora obtained via cross-lingual projections. In this study, we focus on unsupervised sentiment polarity identification and we only investigate the lexicon-based approach in the experiments.

Other related work includes subjective/objective analysis (Hatzivassiloglon and Wiebe, 2000; Riloff and Wiebe, 2003) and opinion mining and summarization (Liu et al., 2005; Popescu and Etzioni. 2005; Choi et al., 2006; Ku et al., 2006; Titov and McDonald, 2008).

3 The Proposed Approach

3.1 Overview

The motivation of our approach is to make full use of bilingual knowledge to improve sentiment analysis in a target language, where the resources

for sentiment analysis are limited or unreliable. This study focuses on unsupervised polarity identification of Chinese product reviews by using both the rich English knowledge and the limited Chinese knowledge.

The framework of our approach is illustrated in Figure 1. A Chinese review is translated into the corresponding English review using machine translation services, and then the Chinese review and the English review are analyzed based on Chinese resources and English resources, respectively. The analysis results are then combined to obtain more accurate results under the assumption that the individual sentiment analysis can complement each other. Note that in the framework, different machine translation services can be used to obtain different English reviews, and the analysis of English reviews translated by a specific machine translation service is conducted separately. For simplicity, we consider the English reviews translated by different machine translation services as reviews in different languages, despite the fact that in essence, they are still in English.

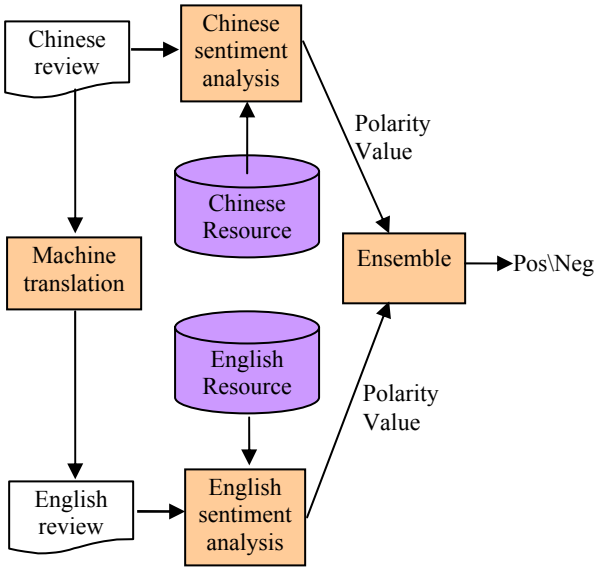

Figure 1. Framework of our approach

Formally, give a review rev^0 in the target language (i.e. Chinese), the corresponding review rev_i in the ith language is obtained by using a translation function: $rev^i = f^i_{Trans}(rev^0)$, where $1 \le i \le p$ and p is the total number of machine translation services. For each review rev^k in the kth language ($0 \le k \le p$), we employ the semantic oriented approach to assign a semantic orientation value $f^k_{SO}(rev^k)$ to the review, and the polarity orientation of the review can be simply predicated based on the value by using a threshold. Given a set of semantic orientation values $F_{SO} = \{f^k_{SO}(rev^k) \mid 0 \le k \le p\}$, the ensemble methods aim to derive a new semantic orientation value $f^{Ensemble}_{SO}(rev^0)$ based on the values in F_{SO}, which can be used to better classify the review as positive or negative.

The steps of review translation, individual semantic orientation value computation and ensemble combination are described in details in the next sections, respectively.

3.2 Review Translation

Translation of a Chinese review into an English review is the first step of the proposed approach. Manual translation is time-consuming and labor-intensive, and it is not feasible to manually translate a large amount of Chinese product reviews in real applications. Fortunately, machine translation techniques have been well developed in the NLP field, though the translation performance is far from satisfactory. A few commercial machine translation services can be publicly accessed. In this study, the following two commercial machine translation services and one baseline system are used to translate Chinese reviews into English reviews.

Google Translate [2] **(*GoogleTrans*)**: Google Translate is one of the state-of-the-art commercial machine translation systems used today. Google Translate applies statistical learning techniques to build a translation model based on both monolingual text in the target language and aligned text consisting of examples of human translations between the languages.

Yahoo Babel Fish [3] **(*YahooTrans*)**: Different from Google Translate, Yaho Babel Fish uses SYSTRAN's rule-based translation engine. SYSTRAN was one of the earliest developers of machine translation software. SYSTRAN applies complex sets of specific rules defined by linguists to analyze and then transfer the grammatical structure of the source language into the target language.

Baseline Translate (*DictTrans*): We simply develop a translation method based only on one-to-one term translation in a large Chinese-to-English

[2] http://translate.google.com/translate_t
[3] http://babelfish.yahoo.com/translate_txt

dictionary. Each term in a Chinese review is translated by the first corresponding term in the Chinese-to-English dictionary, without any other processing steps. In this study, we use the LDC_CE_DIC2.0[4] constructed by LDC as the dictionary for translation, which contains 128366 Chinese terms and their corresponding English terms.

The Chinese-to-English translation performances of the two commercial systems are deemed much better than the weak baseline system. Google Translate has achieved very good results on the Chinese-to-English translation tracks of NIST open machine translation test (MT)[5] and it ranks the first on most tracks. In the Chinese-to-English task of MT2005, the BLEU-4 score of Google Translate is 0.3531, and the BLEU-4 score of SYSTRAN is 0.1471. We can deduce that Google Translate is better than Yahoo Babel Fish, without considering the recent improvements of the two systems.

Here are two running example of Chinese reviews and the translated English reviews (*HumanTrans* refers to human translation):

Positive Example: 优点很多,外形也很好。
HumanTrans: Many advantages and very good shape.
GoogleTrans: Many advantages, the shape is also very good.
YahooTrans: Merit very many, the contour very is also good.
DictTrans: merit very many figure also very good
Negative example: 内存太小不支持红外。
HumanTrans: The memory is too small to support IR.
GoogleTrans: Memory is too small not to support IR.
YahooTrans:The memory too is small does not support infrared.
DictTrans: memory highest small negative not to be in favor of ir.

3.3 Individual Semantic Orientation Value Computation

For any specific language, we employ the semantic orientated approach (Kennedy and Inkpen, 2006) to compute the semantic orientation value of a review. The unsupervised approach is quite straightforward and it makes use of the following sentiment lexicons: **positive Lexicon (*Positive_Dic*)** including terms expressing positive polarity, **Negative Lexicon (*Negative_Dic*)** including terms expressing negative polarity, **Negation Lexicon (*Negation_Dic*)** including terms that are used to reverse the semantic polarity of a particular term, and **Intensifier Lexicon (*Intensifier_Dic*)** including terms that are used to change the degree to which a term is positive or negative. In this study, we conduct our experiments within two languages, and we collect and use the following popular and available Chinese and English sentiment lexicons[6], without any further filtering and labeling:

1) Chinese lexicons

***Positive_Dic^{cn}*:** 3730 Chinese positive terms were collected from the Chinese Vocabulary for Sentiment Analysis (VSA)[7] released by HOWNET.

***Negative_Dic^{cn}*:** 3116 Chinese negative terms were collected from Chinese Vocabulary for Sentiment Analysis (VSA) released by HOWNET.

***Negation_Dic^{cn}*:** 13 negation terms were collected from related papers.

***Intensifier_Dic^{cn}*:** 148 intensifier terms were collected from Chinese Vocabulary for Sentiment Analysis (VSA) released by HOWNET.

2) English lexicons

***Positive_Dic^{en}*:** 2718 English positive terms were collected from the feature file *subjclueslen1-HLTEMNLP05.tff* [8] containing the subjectivity clues used in the work (Wilson et al., 2005a; Wilson et al., 2005b). The clues in this file were collected from a number of sources. Some were culled from manually developed resources, e.g. *general inquirer*[9] (Stone et al., 1966). Others were identified automatically using both annotated and unannotated data. A majority of the clues were collected as part of work reported in Riloff and Wiebe (2003).

***Negative_Dic^{en}*:** 4910 English negative terms were collected from the same file described above.

***Negation_Dic^{en}*:** 88 negation terms were collected from the feature file *valenceshifters.tff* used in the work (Wilson et al., 2005a; Wilson et al., 2005b).

***Intensifier_Dic^{en}*:** 244 intensifier terms were collected from the feature file *intensifiers2.tff* used in the work (Wilson et al., 2005a; Wilson et al., 2005b).

[4] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm
[5] http://www.nist.gov/speech/tests/mt/
[6] In this study, we focus on using a few popular resources in both Chinese and English for comparative study, instead of trying to collect and use all available resources.
[7] http://www.keenage.com/html/e_index.html
[8] http://www.cs.pitt.edu/mpqa/
[9] http://www.wjh.harvard.edu/~inquirer/homecat.htm

The semantic orientation value $f^k_{SO}(rev^k)$ for rev^k is computed by summing the polarity values of all words in the review, making use of both the word polarity defined in the positive and negative lexicons and the contextual valence shifters defined in the negation and intensifier lexicons. The algorithm is illustrated in Figure 2.

Input: *a review rev^k in the kth language. Four lexicons in the kth language: Positive_Dic^k, Negative_Dic^k, Negation_Dic^k, Intensifier_Dic^k, which are either Chinese or English lexicons;*

Output: *Polarity Value $f^k_{SO}(rev^k)$;*

Algorithm Compute_SO:

1. *Tokenize review rev_k into sentence set S and each sentence $s \in S$ is tokenized into word set W_s;*
2. *For any word w in a sentence $s \in S$, compute its SO value SO(w) as follows:*
   1) *if $w \in$Positive_Dic^k, SO(w)=PosValue;*
   2) *If $w \in$Negative_Dic^k, SO(w)=NegValue;*
   3) *Otherwise, SO(w)=0;*
   4) *Within the window of q words previous to w, if there is a term $w' \in$Negation_Dic^k, SO(w)= –SO(w);*
   5) *Within the window of q words previous to w, if there is a term $w' \in$Intensifier_Dic^k, SO(w) =$\rho \times$SO(w);*
3. $f^k_{SO}(rev^k) = \sum_{s \in S} \sum_{w \in W_s} SO(w)$;

Figure 2. The algorithm for semantic orientation value computation

In the above algorithm, *PosValue* and *NegValue* are the polarity values for positive words and negative words respectively. We empirically set *PosValue*=1 and *NegValue*= –2 because negative words usually contribute more to the overall semantic orientation of the review than positive words, according to our empirical analysis. ρ>1 aims to intensify the polarity value and we simply set ρ=2. *q* is the parameter controlling the window size within which the negation terms and intensifier terms have influence on the polarity words and here *q* is set to 2 words. Note that the above parameters are tuned only for Chinese sentiment analysis, and they are used for sentiment analysis in the English language without further tuning. The tokenization of Chinese reviews involves Chinese word segmentation.

Usually, 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.

3.4 Ensemble Combination

After obtaining the set of semantic orientation values $F_{SO}=\{f^k_{SO}(rev^k) \mid 0 \le k \le p\}$ by using the semantic oriented approach, where *p* is the number of English translations for each Chinese review, we exploit the following ensemble methods for deriving a new semantic orientation value $f^{Ensemble}_{SO}(rev^0)$:

1) Average

It is the most intuitive combination method and the new value is the average of the values in F_{SO}:

$$f^{Ensemble}_{SO}(rev^0) = \frac{\sum_{k=0}^{p} f^k_{SO}(rev^k)}{p+1}$$

Note that after the new value of a review is obtained, the polarity tag of the review is assigned in the same way as described in Section 3.3.

2) Weighted Average

This combination method improves the average combination method by associating each individual value with a weight, indicating the relative confidence in the value.

$$f^{Ensemble}_{SO}(rev^0) = \sum_{k=0}^{p} \lambda_k f^k_{SO}(rev^k)$$

where $\lambda_k \in [0, 1]$ is the weight associated with $f^k_{SO}(rev^k)$. The weights can be set in the following two ways:

Weighting Scheme1: The weight of $f^k_{SO}(rev^k)$ is set to the accuracy of the individual analysis in the *k*th language.

Weighting Scheme2: The weight of $f^k_{SO}(rev^k)$ is set to be the maximal correlation coefficient between the analysis results in the *k*th language and the analysis results in any other language. The underlying idea is that if the analysis results in one language are highly consistent with the analysis results in another language, the results are deemed to be more reliable. Given two lists of semantic values for all reviews, we use the Pearson's correlation coefficient to measure the correlation between them. The weight associated with function $f^k_{SO}(rev^k)$ is then defined as the maximal Pearson's correlation coefficient between the reviews' values in the *k*th language and the reviews' values in any other language.

3) Max

The new value is the maximum value in F_{SO}:

$$f_{SO}^{Ensemble}(rev^0) = \max\left\{f_{SO}^k(rev^k) \mid 0 \le k \le p\right\}$$

4) Min

The new value is the minimum value in F_{SO}:

$$f_{SO}^{Ensemble}(rev^0) = \min\left\{f_{SO}^k(rev^k) \mid 0 \le k \le p\right\}$$

5) Average Max&Min

The new value is the average of the maximum value and the minimum value in F_{SO}:

$$f_{SO}^{Ensemble}(rev^0) = \frac{\max\left\{f_{SO}^k(rev^k) \mid 0 \le k \le p\right\} + \min\left\{f_{SO}^k(rev^k) \mid 0 \le k \le p\right\}}{2}$$

6) Majority Voting

This combination method relies on the final polarity tags, instead of the semantic orientation values. A review can obtain p+1 polarity tags based on the individual analysis results in the p+1 languages. The polarity tag receiving more votes is chosen as the final polarity tag of the review.

4 Empirical Evaluation

4.1 Dataset and Evaluation Metrics

In order to assess the performance of the proposed approach, we collected 1000 product reviews from a popular Chinese IT product web site-IT168[10]. The reviews were posted by users and they focused on such products as mp3 players, mobile phones, digital camera and laptop computers. Users usually selected for each review an icon indicating "postive" or "negative". The reviews were first categorized into positive and negative classes according to the associated icon. The polarity labels for the reviews were then checked by subjects. Finally, the dataset contained 886 product reviews with accurate polarity labels. All the 886 reviews were used as test set.

We used the standard precision, recall and F-measure to measure the performance of positive and negative class, respectively, and used the MacroF measure and accuracy metric to measure the overall performance of the system. The metrics are defined the same as in general text categorization.

4.2 Individual Analysis Results

In this section, we investigate the following individual sentiment analysis results in each specified language:

CN: This method uses only Chinese lexicons to analyze Chinese reviews;

GoogleEN: This method uses only English lexicons to analyze English reviews translated by *GoogleTrans*;

YahooEN: This method uses only English lexicons to analyze English reviews translated by *YahooTrans*;

DictEN: This method uses only English lexicons to analyze English reviews translated by *DictTrans*;

In addition to the above methods for using English resources, the lexicon-based method investigated in Mihalcea et al. (2007) can also use English resources by directly projecting English lexicons into Chinese lexicons. We use a large English-to-Chinese dictionary - LDC_EC_DIC2.0[11] with 110834 entries for projecting English lexicons into Chinese lexicons via one-to-one translation. Based on the generated Chinese lexicons, two other individual methods are investigated in the experiments:

CN2: This method uses only the generated Chinese Resources to analyze Chinese reviews.

CN3: This method combines the original Chinese lexicons and the generated Chinese lexicons and uses the extended lexicons to analyze Chinese reviews.

Table 1 provides the performance values of all the above individual methods. Seen from the table, the performances of **GoogleEN** and **YahooEN** are much better than the baseline **CN** method, and even the **DictEN** performs as well as **CN**. The results demonstrate that the use of English resources for sentiment analysis of translated English reviews is an effective way for Chinese sentiment analysis. We can also see that the English sentiment analysis performance relies positively on the translation performance, and **GoogleEN** performs the best while **DictEN** performs the worst, which is consistent with the fact the *GoogleTrans* is deemed the best of the three machine translation systems, while *DictTrans* is the weakest one.

Furthermore, the **CN** method outperforms the **CN2** and **CN3** methods, and the **CN2** method performs the worst, which shows that the generated Chinese lexicons do not give any contributions to the performance of Chinese sentiment analysis. We explain the results by the fact that the term-based one-to-one translation is inaccurate and the generated Chinese lexicons are not reliable. Overall, the

[10] http://www.it168.com

[11] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

approach through cross-lingual lexicon translation does not work well for Chinese sentiment analysis in our experiments.

4.3 Ensemble Results

In this section, we first use the simple average ensemble method to combine different individual analysis results. Table 2 provides the performance values of the average ensemble results based on different individual methods.

Seen from Tables 1 and 2, almost all of the average ensembles outperforms the baseline **CN** method and the corresponding individual methods, which shows that each individual methods have their own evidences for sentiment analysis, and thus fusing the evidences together can improve performance. For the methods of **CN+GoogleEN, CN+YahooEN** and **CN+DictEN**, we can see the ensemble performance is not positively relying on the translation performance: **CN+YahooEN** performs better than **CN+GoogleEN**, and even **CN+DictEN** performs as well as **CN+GoogleEN**. The results show that the individual methods in the ensembles can complement each other, and even the combination of two weak individual methods can achieve good performance. However, the **DictEN** method is not effective when the ensemble methods have already included **GoogleEN** and **YahooEN**. Overall, the performances of the ensemble methods rely on the performances of the most effective constituent individual methods: the methods including both **GoogleEN** and **YahooEN** perform much better than other methods, and **CN+GoogleEN+YahooEN** performs the best out of all the methods.

We further show the results of four typical average ensembles by varying the combination weights. The combination weights are respectively specified as λ·**CN**+(1-λ)·**GoogleEN**, λ·**CN**+(1-λ)·**YahooEN**, λ·**CN**+(1-λ)·**DictEN**, λ_1·**CN**+λ_2·**GoogleEN**+(1-λ_1-λ_2)·**YahooEN**. The results over the MacroF metric are shown in Figures 3 and 4 respectively. We can see from the figures that **GoogleEN** and **YahooEN** are dominant factors in the ensemble methods.

We then investigate to use other ensemble methods introduced in Section 3.4 to combine the **CN**, **GoogleEN** and **YahooEN** methods. Table 3 gives the comparison results. The methods of "**Weighted Average1**" and "**Weighted Average2**" are two weighted average ensembles using the two weighing schemes, respectively. We can see that all the ensemble methods outperform the constituent individual method, while the two weighted average ensembles perform the best. The results further demonstrate the good effectiveness of the ensemble combination of individual analysis results for Chinese sentiment analysis.

Individual Method	Positive			Negative			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	MacroF	Accuracy
CN	0.681	0.929	0.786	0.882	0.549	0.677	0.732	0.743
CN2	0.615	0.772	0.684	0.678	0.499	0.575	0.630	0.638
CN3	0.702	0.836	0.763	0.788	0.632	0.702	0.732	0.736
GoogleEN	0.764	0.914	0.832	0.888	0.708	0.787	**0.810**	**0.813**
YahooEN	0.763	0.871	0.814	0.844	0.720	0.777	0.795	0.797
DictEN	0.738	0.761	0.749	0.743	0.720	0.731	0.740	0.740

Table 1. Individual analysis results

Average Ensemble	Positive			Negative			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	MacroF	Accuracy
GoogleEN+YahooEN	0.820	0.900	0.858	0.885	0.795	0.838	0.848	0.848
GoogleEN+YahooEN +DictEN	0.841	0.845	0.843	0.838	0.834	0.836	0.840	0.840
CN+GoogleEN	0.754	0.949	0.840	0.928	0.678	0.784	0.812	0.816
CN+YahooEN	0.784	0.925	0.848	0.904	0.736	0.811	0.830	0.832
CN+DictEN	0.790	0.867	0.827	0.847	0.761	0.801	0.814	0.815
CN+GoogleEN +YahooEN	0.813	0.927	0.866	0.911	0.779	0.840	**0.853**	**0.854**
CN+GoogleEN+ YahooEN+DictEN	0.831	0.891	0.860	0.878	0.811	0.843	0.852	0.852

Table 2. Average combination results

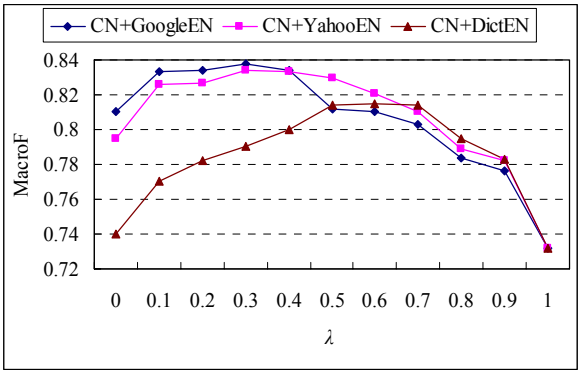

Figure 3. Ensemble performance vs. weight λ for λ*CN+(1-λ)*GoogleEN/YahooEN/DictEN

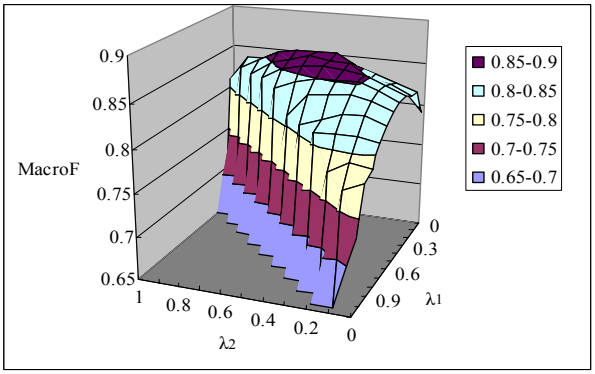

Figure 4. Ensemble performance vs. weights λ_1 and λ_2 for λ_1*CN+λ_2*GoogleEN+(1-λ_1-λ_2) *YahooEN

Ensemble Method	Positive			Negative			Total	
	Precision	Recall	F-measure	Precision	Recall	F-measure	MacroF	Accuracy
Average	0.813	0.927	0.866	0.911	0.779	0.840	0.853	0.854
Weighted Average1	0.825	0.922	0.871	0.908	0.798	0.849	**0.860**	**0.861**
Weighted Average2	0.822	0.922	0.869	0.908	0.793	0.847	0.858	0.859
Max	0.765	0.940	0.844	0.919	0.701	0.795	0.820	0.823
Min	0.901	0.787	0.840	0.805	0.910	0.854	0.847	0.848
Average Max&Min	0.793	0.936	0.859	0.918	0.747	0.824	0.841	0.843
Majority Voting	0.765	0.940	0.844	0.919	0.701	0.795	0.820	0.823

Table 3. Ensemble results for **CN & GoogleEN & YahooEN**

5 Conclusion and Future Work

This paper proposes a novel approach to use English sentiment resources for Chinese sentiment analysis by employing machine translation and ensemble techniques. Chinese reviews are translated into English reviews and the analysis results of both Chinese reviews and English reviews are combined to improve the overall accuracy. Experimental results demonstrate the encouraging performance of the proposed approach.

In future work, more additional English resources will be used to further improve the results. We will also apply the idea to supervised Chinese sentiment analysis.

Acknowledgments

This work was supported by the National Science Foundation of China (No.60703064), the Research Fund for the Doctoral Program of Higher Education of China (No.20070001059) and the National High Technology Research and Development Program of China (No.2008AA01Z421). We also thank the anonymous reviewers for their useful comments.

References

A. Andreevskaia and S. Bergler. 2008. When specialists and generalists work together: overcoming domain dependence in sentiment tagging. In *Proceedings of ACL-08: HLT*.

J. Blitzer, M. Dredze and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In *Proceedings of ACL2007*.

Y. Choi, E. Breck, and C. Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proc. EMNLP*.

A. Devitt and K. Ahmad. 2007. Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of ACL2007*.

K. Eguchi and V. Lavrenko. 2006. Sentiment retrieval using generative models. In *Proceedings of EMNLP*.

V. Hatzivassiloglou and K. R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of EACL*.

V. Hatzivassiloglon and J. M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*.

K. Hiroshi, N. Tetsuya and W. Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of COLING*.

N. Kaji and M. Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of EMNLP-CONLL*.

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

S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING*.

L.-W. Ku, Y.-T. Liang and H.-H. Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI*.

J. Li and M. Sun. 2007. Experimental study on sentiment classification of Chinese review using machine learning techniques. In *Proceeding of IEEE-NLPKE2007*.

B. Liu, M. Hu and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of WWW*.

R. McDonald, K. Hannan, T. Neylon, M. Wells and J. Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of ACL2007*.

R. Mihalcea, C. Banea and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL*.

T. Mullen and N. Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*.

B. Pang, L. Lee and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

B. Pang and L. Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

A. –M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of EMNLP*.

J. Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of ACL*.

E. Riloff and J. Wiebe 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP2003*.

P. J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie and associates. 1966. The General Inquirer: a computer approach to content analysis. The MIT Press.

H. Takamura, T. Inui and M. Okumura. 2005. Extracting semantic orientation of words using spin model. In *Proceedings of ACL*.

I. Titov and R. McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*.

B. K. Y. Tsou, R. W. M. Yuen, O. Y. Kwong, T. B. Y. La and W. L. Wong. 2005. Polarity classification of celebrity coverage in the Chinese press. In *Proceedings of International Conference on Intelligence Analysis*.

P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL.

S. Wang, Y. Wei, D. Li, W. Zhang and W. Li. 2007. A hybrid method of feature selection for Chinese text sentiment classification. In *Proceeding of IEEE-FSKD2007*.

T. Wilson, J. Wiebe and P. Hoffmann. 2005a. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT/EMNLP2005*, Vancouver, Canada.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan. 2005b. OpinionFinder: a system for subjectivity analysis. In *Proceedings of HLP/EMNLP on Interactive Demonstrations*.

J. Yao, G. Wu, J. Liu and Y. Zheng. 2006. Using bilingual lexicon to judge sentiment orientation of Chinese words. In *Proceedings of IEEE CIT2006*.

Q. Ye, W. Shi and Y. Li. 2006. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach. In *Proceedings of 39th Hawaii International Conference on System Sciences*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP2003*.