Constructing Chinese Sentiment Lexicon Using Bilingual Information

Yan Su and Shoushan Li

Natural Language Processing Lab, Soochow University
1 Shizi Street, Suzhou, China 215006
{yansu.suda,shoushan.li}@gmail.com

Abstract. Currently, sentiment analysis has become a hot research topic in the natural language processing (NLP) field as it is highly valuable for many practical usages and theoretical studies. As a basic task in sentiment analysis, construction of sentiment lexicon aims to classify one word into positive, neutral or negative according to its sentiment orientation. However, when constructing a sentiment lexicon in Chinese, there are two major problems: 1) Chinese words are very ambiguous, which makes it hard to compute the sentiment orientation of a word; 2) Given the related research on sentiment analysis, available resource for constructing Chinese sentiment lexicons remains weak. Note that there are several corpus and lexicons in English sentiment analysis. In this study, we first use machine translation system with bilingual resources, i.e., English and Chinese information, then we get the sentiment orientation of Chinese words by computing the point-wise mutual information (PMI) values with English seed words. Experiment results from three domains demonstrate that the lexicon generated with our approach reaches an excellent precision and could cover domain information effectively.

Keywords: Sentiment Analysis, Sentiment Lexicon, Bilingual, PMI.

1 Introduction

The advancement of Web 2.0 technologies have led to the explosive growth of online opinion data. In order to automatically process these large-scale text information, sentiment analysis has recently received considerable interests in the Natural Language Processing (NLP) community [1-2]. In sentiment analysis, the task of sentiment lexicon construction is considered as a basic task which aims to detect whether a word is positive, negative or neutral, i.e., the semantic orientation of the word. This task is useful in many cases. First, a sentiment lexicon can provide important prior knowledge to improve the classification of higher-level text (document, sentence). For example, most unsupervised document-level sentiment classification methods are based on sentiment words [3]. Second, word-level sentiment analysis also has important significance for the word semantic understanding and disambiguation. As pointed by Wiebe [4], sentiment orientation can be associated with word's definition and contribute to the traditional word sense disambiguation task. Third, sentiment lexicon

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 322–331, 2013.

provides an important foundation for many real-life applications, such as text classification, automatic summarization, and text filtering.

However, up to now, no universally fine sentiment lexicon exists for Chinese sentiment analysis. Basically, it is difficult to build a good sentiment lexicon because many words are ambiguities when expressing the sentiment. That is to say, the polarities of words are sometimes sensitive to the topic domain or the context. Even worse, the same word may indicate different polarities with respect to different topic and context. For example, the Chinese word "圆滑" (yuan hua, slyness) has the meaning of "smooth" or "cunning". In sentence (1) as shown in the following, the word "圆滑" (yuan hua, slyness) is positive while being negative when appearing in sentence (2). In addition, traditional method of building sentiment lexicon is to expand by existing electronic dictionary or word knowledge base. Unfortunately, since the study on Chinese sentiment analysis research starts very late, the available resources on sentiment lexicon are extremely rare. Therefore, designing an efficient algorithm of constructing Chinese lexicon becomes a quite challenging and emergency task.

(1)：该笔记本电脑外形边角处理十分圆滑。(The notebook computer shape corner handling is very smooth.)

(2)：他变得圆滑，只能选择一再躲避现实。(He becomes cunning, and only selecst repeatedly to escape reality.)

In contrast, the study on English sentiment analysis starts much earlier and have got a number of related corpus and resources which could provid many seed words with correct sentimental categories. In this paper, we propose a novel method for computing Chinese word's sentiment orientation which combines bilingual corpus and English seed words to construct a Chinese sentiment lexicon. Our approach adopts machine translation system (Google Translate[1]) to eliminate the barriers between both Chinese and English languages. A source language and the corresponding translation language comment is seemed as an entire document, and then we compute point-wise mutual information (PMI) between each Chinese word and English positive (negative) seed words. Finally, we get a Chinese sentiment lexicon with semantic orientation value weights. Our method not only employs the polarity information in the English seed words, but also combines the bilingual constraint information in different context. Experiments across three domains show that our method could get a Chinese sentiment lexicon with a high precision and also could cover different context information.

The remainder of this paper is organized as follows. Section 2 overviews the related work in constructing sentiment lexicon. Section 3 presents our approach of combining bilingual resource and English seeds words to generate a Chinese sentiment lexicon. Section 4 evaluates the experimental results. Finally, Section 5 draws the conclusion and outlines the future work.

[1] www.google.com

2 Related Work

According to the granularities of the concerned text, the tasks of sentiment analysis can be broadly divided into three main groups: document-level [5-7], sentence-level [8], word-level [9-11]. Among these tasks, the word-level sentiment analysis has been considered as one basic task for sentiment analysis. The main objective of such tasks is to automatically construct a sentiment lexicon. Generally, the main methods for constructing a sentiment lexicon can be categorized into three types: (1) Use existing electronic dictionary or word knowledge data base to generate sentiment lexicon. In English study, the resource of Word Net is popularly used [12-13], while in Chinese study, the resource of HowNet is often used [14]. The main idea of this kind of approach is to find sentiment words in dictionary which have the similar semantics with the unknown word and then infer the sentiment orientation of the unknown words. However, this method cannot cover the context of the words. (2) Apply unsupervised machine learning method: the co-occurrence frequency is often employed in corpus to infer the word's close connection with some kind of polarity categories. Some representative papers include: [15-17]. For these approaches, the initial seed words play a central role in the success of constructing a high-precision sentiment lexicon. (3) Use human-annotated corpus: It inferred the sentiment tendency of one word according the co-occurrence relationships or semantic relations on the basis of annotate sentiment classification corpus. This kind of method needs lager amount manual annotated corpus. Some representative papers include [18-20].

Unlike all above studies, the proposed method in this study do not rely on the Chinese sentiment seed words, but make use of bilingual corpus and the English seed words when build a Chinese sentiment lexicon. This is a first attempt to construct a sentiment lexicon by using bilingual information. Our approach combines the English seed words and bilingual statistics resource to get a Chinese lexicon which could cover better contextual information in a specific domain.

3 Construction Method for Chinese Sentiment Lexicon with Bilingual Resources

3.1 Combine English Seed Words and Bilingual Constrain Information

Traditional methods of constructing a sentiment lexicon includes: one is to expand lexicon by existing electronic dictionary or word knowledge base, this method is obviously dependent on the number of Chinese seed words. The other is based on manual annotated corpus, this need large-scale Chinese corpus. However, Chinese sentiment analysis starts late, the available Chinese corpus and Chinese seed words resources are rather limited. In contrast, English sentiment analysis related resources are rich and easy to get. Therefore, in our approach, we collect bilingual corpus across

three domains and some English seed words to construct Chinese sentiment lexicon. Specifically, with the help of Google Translate, we translate the English comments into the Chinese comments and also translate the Chinese Comments into the English reviews. Therefore, the contents of each comment is consist of the source language text and the corresponding translated text, so Chinese sentiment words and English sentiment words with the same meaning will appear in a single document. Then we calculate the mutual information (PMI) between each unknown Chinese word and the English positive (or negative) seed words. Finally we get a lexicon with confidence weights according to the PMI value of positive (negative) category. The proposed method can take full advantage of the bilingual constraint information. The framework of the entire algorithm is illustrated in Figure 1.

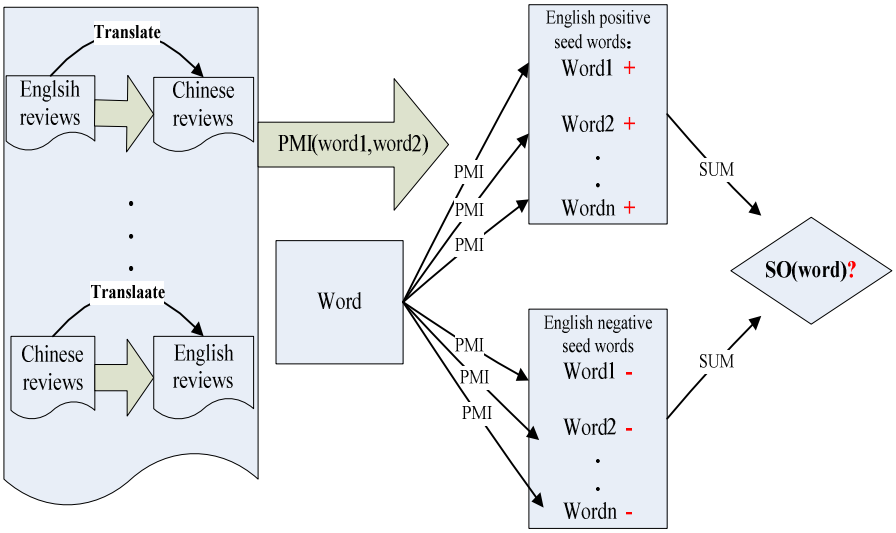

Fig. 1. The framework of Chinese sentiment lexicon construction based on bilingual resources Similarity Computation between Each Two Words

3.2 Similarity Computation between Each Two Words

The PMI-IR algorithm uses mutual information as a measure of the strength of semantic association between two words. In our approach, the point-wise mutual information (PMI) are utilized to compute the similarity between Chinese word and the English positive (or negative) seed words word. The PMI values from both the positive and negative sides can be used to determine the word's sentiment polarity. The computation formula of PMI is shown in Equation (1) [16]:

$$PMI(w_1, w_2) = \log_2 \left(\frac{p(w_1 \,\&\, w_2)}{p(w_1)p(w_2)} \right) \tag{1}$$

Where $p(w_1 \& w_2)$ denotes the co-occurrence probability of the two terms, and $p(w_1)$, $p(w_2)$ denote the occurrence probability of two words respectively. The log of this ratio is the amount of information that we acquire about the presence of one of the words in the same document. Here, co-occurrence means the two words appear in the same document.

Given the English seed words, we first calculate the PMI value of the Chinese word in corpus with all English positive seed words and then calculate the PMI value of the Chinese word with all English negative seed words. Finally, we utilize the discrepancy of two PMI values to determine the polarity of the Chinese word. The specific computation formula is given as follows:

$$SO(w_{Chinese}) = POS_j \times \left[\lambda \times \sum_{i=0}^{N+} PMI(w_{Chinese}, w^i_{English+}) - \sum_{k=0}^{N-} PMI(w_{Chinese}, w^k_{English-}) \right] \quad (2)$$

Where $w^i_{English+}$ is a word in the set of positive seed words in English, $w^k_{English-}$ is a word in the set of negative seed words in English. *N*+ is the size of English positive seed word set, and *N*- is the size of English negative seed word set. The parameter λ is related with the size of two seed word set. In our experiment, we find that the size of positive seed word set is smaller than the size of negative seed word. Thus, the parameter λ is set to be larger than 1, exactly to be 1.2 in this study. In addition, since the sentiment words are more likely to be adjectives and verbs, we set the parameters POS_j of these terms with corresponding priority. In particular, adjectives, verbs, and other word, corresponding priority is set to 3, 2, and 1 respectively. Assuming that the word "good", "excellent", "perfect" as positive seed words, 'bad "," poor "," disappointed "as negative seed words. More specifically, a Chinese word is assigned a numerical rating *SO* by taking the mutual information between the given context and the word see like “excellent” and subtracting the mutual information between the given context and the word set like “poor”. In addition to determining the direction of the word’s semantic orientation (positive or negative, based on the sign of the rating), this numerical rating also indicates the strength of the semantic orientation.

3.3 Our Algorithm

In our approach, we first utilize machine translation system translate Chinese (or English) comments into another language. Consequently, each comment is represented by two languages. Then, we use the English seed words to calculate PMI value between each unknown Chinese word and English positive (negative) seed words according to formula (1). Finally, formula (2) is employed to get a Chinese sentiment lexicon with polarity weight. The detailed algorithm is given in Figure .

Our Algorithm

Input:

English reviews U_{en}, Chinese reviews U_{cn};

English seed words L_{en};

Output:

A Chinese sentiment lexicon with polarity weight L_{cn};

Process:

1. Initialize the bilingual reviews $U = \varnothing$;
2. Translate every comment in U_{en} into Chinese comment, combine the English review and translated review as an entire bilingual comment, add the bilingual comment into U;
3. Translate every comment in U_{cn} into Chinese comment, combine the English review and translated review as an entire bilingual comment, add the bilingual comment into U;
4. In the feature vector set of U to calculate for each Chinese word between positive seed word and negative seed word by mutual information (PMI);
5. According to formula (2) to calculate the polarity of Chinese word, the symbol can represent the polarity of the word, while the absolute value represents the intensity.
6. Sort the lexicon according to the polarity strength of word;

4 Experiment

In this section, we systematically evaluate our approach on the data collection as described in Section 4.

4.1 Experimental Setting

The data collection contains three domains: Electronic, Beauty, and Software. In our approach, we collect both English corpus and Chinese corpus from http://www.amazon.cn/. Table 1 reports the distribution of English documents and Chinese documents in each domain. English seed words come from the lexicon provided by [21]. This lexicon contains 2000 positive and 4000 negative seed words. For the Chinese text, we use the Chinese text segmentation tool ICTCLAS to obtain the word list.

Table 1. The number of English and Chinese documents in three domains

	Electronic	Beauty	Software
English (positive)	2000	1000	1000
English (negative)	2000	1000	1000
Chinese (positive)	2000	1000	1000
Chinese (negative)	1850	1000	700

4.2 Experimental Results

Note that too many words exist in our corpus and manual annotating all the word is too time-consuming. Thus, in the preliminary experiment, we only get part of words with high scores for manual annotation.

In our approach, we adopt precision to evaluate the effect of classification, which is defined as follows:

$$Precision = \frac{number\ of\ correctly\ classified\ words}{total\ number\ of\ all\ labeled\ words} \tag{3}$$

Table 2 shows the precision of the Top 100, 200 words with highest scores and some instances of positive, negative Chinese word. From this table, the precision in top 100 words is very high and only a very small number of words are miss-predicted. In detail, there are 5, 3, and 5 words are not correctly predicted in Electronic, Beauty and Software respectively. As far as the top 200 words are concerned, the precisions of predicting the sentiment category in three domains are 84.5%, 83.5%, 89.0% respectively. The good performance of our approach is mainly due to its both utilizing the polarity of the English seed words and taking full advantage of the bilingual constraints information to sentiment word in the context environment. Therefore, the obtained Chinese sentiment lexicon not only covers the context information of the words in the special environment, but also achieves higher precision.

To check the results, we find that the errors mainly occur in the following two cases: 1) some words appear in the context of a particular sentimental category but contain no sentiment. For example, the Chinese word “造成”(zao cheng, cause) is neutral in the general sense, but it often appear in negative comment sentences. Our method depends on the context information of the words, which makes the word is predicted as a negative word. Another example is the word “加剧”(jia ju, exacerbate) which also often appear in a negative context. 2) Polarity shifting is another popular phenomenon to lead classification error for the polarity detection. For example, “帮助”(bang zhu, help) is positive in general sense, but it also appear in polarity shifting sentence like “There is not too much help for me”. As a result, our approach calculates the negative weight of the word will be much more possible than positive.

Table 2. Classification precision of the Top 100, 200 words with highest score and some positive, negative instances

	Electronic	Beauty	Software
Precision (Top 100)	95.0%	97.0%	95.0%
Precision (Top 200)	84.5%	83.5%	89.0%
Positive instances	柔和、清晰	精致、优雅	简洁、神奇
Negative instances	糟糕、退款	气愤、坏	浪费、沮丧

In addition, the framework we proposed is quite general and applicable for sentiment lexicon construction in any domains. It is capable of incorporating different sources of available information for the automatic construction of a domain- oriented sentiment lexicon. For instance, the Chinese word "圆滑"(yuan hua, slyness) has the same meaning of "smooth" and "cunning". It seems to be positive when is used to descript a computer, but it's considered to negative when decrypting a person. However, in our approach we make full advantage of the bilingual constraint information. "圆滑"(yuan hua, slyness) often appears in the same sentence with its translation "smooth" in electronic domain, so it is more likely to be inferred as a positive word. While the word often appears in a negative context with its translation words "cunning", so it is more likely to be inferred with a negative semantic orientation in this context.

Given an English seed lexicon, a simple method of building a Chinese lexicon might be to translate directly from English lexicon by machine translation systems. However, the direct translation has a lot of ambiguity, and cannot cover a lot of sentiment words in a specific domain. We calculate and compare the coverage ratio of top 100, 200 words with high score in the translated Chinese lexicon across three domains.

Table 3. Coverage ratio of top 100,200 words with high score in the translated Chinese lexicon across three domains

	Electronic	Beauty	Software
Coverage (100)	66.0%	60.0%	65.0%
Coverage (200)	57.0%	51.0%	54.0%

It can be seen from Table 3, the Chinese lexicon obtained by the direct translation has a low coverage of true sentiment words in corpus. Therefore, only use translation cannot get the emotional polarity of the many words in particular domain. Instead, our approach could provide a good supplement to capture many other Chinese sentiment words in a specific domain.

5 Conclusion

In this paper, we employ both the bilingual corpus and English seed words to construct a Chinese sentiment lexicon. In particular, we calculate PMI values between each unknown Chinese word and English positive (negative) seed words. The proposed framework is general and is applicable for lexicon construction in any domain.

It is capable of incorporating different sources of available information for the automatic construction of a context-aware sentiment lexicon. Experiment results from three domains demonstrate that the lexicon generated with our approach reach an excellent precision and could get many sentiment words in a special domain.

For the future work, we will consider the label of annotation corpus and polarity shifting phenomenon to improve the performance of sentiment lexicon construction. Furthermore, we plan to apply our approach to construct sentiment lexicon in other languages.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP, pp. 79–86 (2002)
2. Li, S., Huang, C., Zong, C.: Multi-Domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology, 25–33 (2011)
3. Kennedy, A., Inkpen, D.: Sentiment Classification of Movie Reviews using Contextual Valence Shifters. Computational Intelligence 22(2), 110–125 (2006)
4. Wiebe, J., Mihalcea, R.: Word Sense and Subjectivity. In: Proceeding of ACL-COLING, pp. 1065–1072 (2006)
5. Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL, pp. 174–181 (1997)
6. Wiebe, J.: Learning Subjective Adjectives from Corpora. In: Proceedings of AAAI, pp. 735–740 (2000)
7. Cui, H., Mittal, V., Datar, M.: Comparative Experiments on Sentiment Classification for Online Product Reviews. In: Proceedings of AAAI, pp. 1265–1270 (2006)
8. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In: Proceedings of ACL, pp. 271–278 (2004)
9. Kim, S., Hovy, E.: Determining the Sentiment of Opinions. In: Proceedings of COLING, pp. 1367–1373 (2004)
10. Li, S., Huang, C., Zhou, G., Lee, S.: Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In: Proceedings of ACL, pp. 414–423 (2010)
11. Li, S., Wang, Z., Zhou, G., Lee, S.: Semi-Supervised Learning for Imbalanced Sentiment Classification. In: Proceedings of IJCAI, pp. 1826–1831 (2011)
12. Andrea, E.: Determining the Semantic Orientation of Terms through Gloss Classification. In: Proceedings of CIKM, pp. 617–624 (2005)
13. Hassan, A., Radev, D.: Identifying Text Polarity Using Random Walks. In: Proceedings of ACL, pp. 395–403 (2010)
14. Zhu, Y., Min, J., Zhou, Y., Huang, X., Wu, L.: Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing, 14–20 (2006)
15. Hatzivassiloglou, V., Wiebe, J.: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: Proceedings of ACL, pp. 299–304 (2000)
16. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL, pp. 417–424 (2002)
17. Popescu, A., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of HLT/EMNLP, pp. 339–346

18. Akkaya, C., Wiebe, J., Mihalcea, R.: Subjectivity Word Sense Disambiguation. In: Proceeding of EMNLP, pp. 190–199 (2009)
19. Li, S., Huang, C.: Word Sentiment Orientation Computing with Feature Selection Methods. In: CLSW (2009)
20. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. In: Proceedings of WWW, pp. 347–356 (2011)
21. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of HLT/EMNLP, pp. 347–354