A Hybrid Approach to Identifying Sentiment Polarity for New Words

Yang Yang[1], Ruifan Li[1, 2], Yanquan Zhou[1, 2]
1, School of Computer Science
Beijing University of Posts and Telecommunications
2, Engineering Research Center of Information Networks
Ministry of Education
newyangyang@bupt.edu.cn, rfli@bupt.edu.cn, zhouyanquan@bupt.edu.cn

***Abstract*—Microblog is a typical form of heterogeneous information. For this information, identifying sentiment polarity of new words plays a fundamental role in sentiment analysis. In this paper, we proposed a hybrid approach using both statistic and syntax information to identifying the sentiment polarity of new words. We first filter the raw tweets out some noises and segment the clean data with POS tagging. Next, we collect new words by filtering rules. Then, we assign each new word with a polarity using both statistics and patterns information. We evaluate our approach on a real dataset from Sina Weibo, achieving a relatively high F-score of 0.241 compared with the baseline of 0.22.**

Keywords—dictionaries; new words; patterns; sentiment polarity

I. INTRODUCTION

Weibo, akin to Twitter, is the biggest micro-blogging website in Chinese. After being launched by Sina Company in 2009, Weibo has gained over 500 million registered users up to now. Here, about 100 million tweets are posted each day. Some tweets posted are full of sentiment.

Sentiment analysis tries to extract the subjective information from texts. As for general documents, the sentiment analysis of these tweets is meaningful. For example, analyzing tweets about some products could show some preferences. Similarly, mining tweets on some dishes could provide recommendations.

The polarity analysis of sentiment words, among the tasks of sentiment analysis, is fundamental. Sentiment words are those that convey positive or negative sentiment polarities. The sentiment polarity of a word indicates the direction in which the word deviates from the neutral. Besides, almost all the sentiment analysis tasks involve the use of sentiment lexicon. And the quality of the dictionary has a powerful influence on the performance of sentiment analysis system. Therefore, a sentiment lexicon system is essential for sentiment analysis.

However, it is difficult to maintain a universal sentiment lexicon for general cases. That is, the sentiment of many words is context-dependent [1]. For example, the word "pride" in "We feel pride in you" is a compliment, but in “Pride goes before a fall”, it is derogatory. Besides, the neologisms are growing in tweets and manual works are far from enough to deal with them. An urgent need is to automatically find the neologisms and give the polarities for them.

Sentiment analysis in tweets is different from general documents. Tweets in Weibo are informal with plenty of colloquial words and abundant neologisms (in our paper, neologisms are interchangeably used with new words). For these reasons, it is difficult to determine the sentiment polarity of these words. Furthermore, compared with the other words in one tweet, neologism is more crucial to the user's sentiment orientation.

In this paper, we try to find new words and determine their polarities in a hybrid approach. Our method first filters the raw tweets out some pre-defined noises and segments the clean data with POS tagging. Next, new words are collected by three successive filter rules. Finally, the new words are assigned with one of the three polarities: positive, negative, and neutral using statistics and patterns. We evaluate our method on a real dataset achieving a satisfactory result.

II. RELATED WORK

Sentiment analysis of text can be traced back to the 1990s. And the early work mainly focused on word semantic orientation calculation and the sentiment classification of document-level text. Then, researchers concentrated on fine-grained analysis, such as mining of product attributes, domain independent sentiment classification, and sentiment summary. Currently scholars have studied the word sentiment classification. And the main research methods can be summarized into two: the corpus-based method and the lexicon-based method[2].

The corpus-based method mainly utilizes the statistics of large corpus to make the determination. Some scholars found that there usually exists some relevance between the polarities of two adjectives connected by a conjunction; two adjectives like 'lovely and beautiful' connected by the conjunction 'and' have the same polarity; on the contrary, two adjectives connected by 'but' have different polarities. Based on this

This work was supported by the National Natural Science Foundation of China with grant No. 71231002 (Key Project) and 61100120, Fundamental Research Funds for the Central Universities, Project No. 2013RC0304.

observation, Hatzivassiloglou and McKeown [3] discovered a lot of opinionated adjectives from the large corpus of Wall Street Journal.

Janyce Wiebe *et al.* [4] followed a similar approach. They used a word clustering method on a large corpus to obtain the opinionated adjectives. However, the methods mentioned above only had the adjectives evaluated ignoring the opinionated words of other part-of-speech. Riloff *et al.* [5] developed some templates manually evaluate and select seed words, and utilized an iterative method to collect the opinionated nouns. Turney and Littman [6] proposed a method of point mutual information to determine whether a word is an opinionated one. This method is applicable to identify the opinionated words of various parts-of-speech, but it largely depends on sentiment seed lexicon. The advantage of corpus-based approach is simple; the disadvantage is that the corpora of comments are difficult to obtain.

The semantic relations between the words in the dictionary are mainly used to extract the sentiment words by the lexicon-based method. The dictionary mentioned above generally refers to WordNet, HowNet and so on. This method is simple, but depends on the number and quality of seed words, and it would bring in noise caused by the ambiguity of words. To avoid the ambiguity of words, some scholars used the annotation of words to boost the recognition of sentiment word [7-8]. Turney *et al.* proved that we could obtain higher accuracy by using paradigm words with low sensitivity. In addition, some scholars continued to use the point-wise mutual information method proposed by Turney to identify the opinionated words by calculating the degree of association with all the adjectives in WordNet. The advantage of this method is that we can build a large dictionary of opinionated words, but polysemy is very common to many words, the dictionaries we built often contain lots of ambiguous words.

Assume that word and its synonyms have the same polarity. Based on this assumption, we can determine the polarity of a word by analyzing its synonyms' polarity. However, not all languages have rich sentiment words. In Chinese, Yuen *et al.* [9] proposed the use of the characteristics of a single Chinese character to determine the sentiment polarity. Also, some scholars enrich the sentiment lexicon by translating the language with rich sentiment words, such as translating the English dictionary into Chinese. But experiments showed that the polarity of some words had been changed after translation, which also proves what J. Wiebe has pointed out in [10]: "Meaning of words and their polarities have a certain relationship, but words with similar meaning do not necessarily have the same polarity". The lexicon-based methods are unable to find domain dependent sentiment words since many entries in dictionaries are domain independent.

III. METHODS

In our approach, we first clean the raw data. Then the cleaned corpus is segmented with POS tagging. Next, new words are got by filtering. At last, the new words are classified to three kinds of polarities: positive, negative, and neutral.

A. Weibo Cleaning

The content of raw Weibo data has elements that are redundant in our task and they may cause certain problems in processing. Thus we clean out these elements to make data computer-friendly. The elements removed are: URLs, “shared from somewhere” expressions, meaningless symbols, author separators, author IDs, Weibo hashtags, repeated punctuation symbols, strings of digits.

B. Segmentation

Segmentation is pre-processing step in natural language processing tasks in Chinese. In our setting, whether a new word can be detected depends on this word can be segmented correctly or not. Thus we need a segmentation tool that has a good reputation in finding new words in large amount of text. What we use in our case is NLPIR (Natural Language Processing and Information Retrieval platform), previously known in the name of ICTCLAS, a tool that integrates functions like extracting new words, segmentation, and POS tagging. NLPIR is freely available at http://ictclas.nlpir.org/.

Firstly, we use NLPIR to collect new words in the corpus; here “new” means these words are new to NLPIR. Secondly, add these new words to user dictionary (the dictionary to which user can add unknown words) to boost the performance of segmentation. Thirdly, segment the cleaned corpus with POS tagging. Most of the new words are assigned with POS if NLPIR can identify their POS according to context; if not, a POS of n_new is assigned, meaning this is a new word with an uncertain POS.

C. New Words Selection

As mentioned above, NLPIR can extract new words with respect to itself, but the new words we concerned are defined by not listed in the given old dictionary. So selection method mixing statistic and morphology information is applied.

Firstly, we generate the vocabulary of corpus using simple word counting. The words that contain non-Chinese characters or only have one Chinese character are filtered. Secondly, the new words list is obtained by kicking out of vocabulary the words appear in the old dictionary and the words that satisfy filtering rules illustrated below.

Two filtering rules are conducted.

1. Words that occur less than a certain number of times should be filtered. By inspecting the corpus vocabulary, we find that words with very few occurrences are usually not real words, many of them are names or just meaningless character strings. In practice, we make the threshold number 10.

2. Words with certain POSs should be filtered. The goal of our task is to find the sentiment polarity of new words. But in the vocabulary we obtained above, many of the words do not have sentiment polarities, such as names of people and those of places. These words degrade the performance of sentiment classification. POS tagging gives us convenience to filter out them. The POSs filtered out in our system include proper noun, localizer, time word, verb noun, adverb, pronoun, quantifier, numeral, preposition, conjunction, mimetic word etc.

D. Determining the Sentiment Polarity

One of the most popular approaches to classify word sentiment is to utilize the statistic information of context words. (Wang *et al*, 2012) proposed an approach of building smiley emotion lexicon. The basic idea is to determine the sentiment polarity by inspecting the occurrence of positive and negative words[11]. The underlying hypothesis here is that sentiment words tend to occur in context that have more words with the same polarity rather than those with opposite polarities. This hypothesis is in accordance with intuition, but things are not always the case. Without considering the syntactic structure, some information like antonym is lost.

To overcome the shortcomings, we propose a hybrid method considering both statistic and syntax information. To bring in syntax information, phrase patterns are introduced.

Despite the statistic information, syntax can also provide useful clues. Like most languages, Chinese have some fixed expressions for the basic conjunctive relationships: progressive, coordination, and adversative. These relationships show strong clues about the sentiment polarity of words involved. In progressive relationship, a sentiment word is followed by a word with the same sentiment polarity, and the strength is usually stronger. In coordination relationship, the two words involved have the same sentiment polarity, if any. In adversative relationship, the two words have contrary polarities. In this paper, we use phrase patterns to exploit this information; the antonym words are considered. Table I shows a part of the patterns we used.

The flowchart of the sentiment classification is shown in Fig. 1. Given a specific new word to be considered, the system scans the corpus. Once the new word appears, its context sentence is taken out to be analyzed. Firstly, according to the given sentiment dictionary, the system counts how many positive and negative words are there in the sentence, adding the numbers to the counters. Secondly, the pattern matcher searches the sentence to see which patterns the sentence can match. If one pattern can be matched and the new word is involved in the conjunctive relationship, we can infer the sentiment polarity of the new word by looking up the polarity of the other word. If successful, we call it a positive match or a negative one, adding to its counter.

After the procedure is done, the system assigns a sentiment polarity to the new word according to the sentimental neighbor counts, the pattern matches, and the lexicon information of the new word. The counts of positive and negative pattern matches are denoted as M_p and M_n respectively.

TABLE I EXAMPLES OF PHRASE PATTERNS

Phrase Patterns	Approximate Translation	Type
不仅 而且， 不仅 还， 不但 而且， 不但 还	not only but also, even more	progressive
又 又， 既 又， 又，并且	and, also	coordinating
但是，却， 不过，只是	but, however	adversative

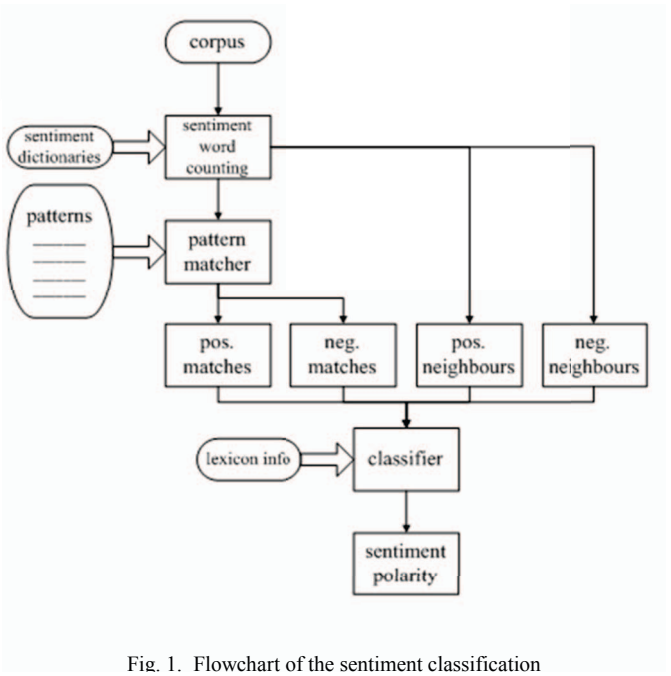

Fig. 1. Flowchart of the sentiment classification

Step 1. Get the possible polarity i with (1). Check this polarity first, but the other one will be checked, too.

$$i = \begin{cases} 1, & M_p > M_n \\ -1, & otherwise \end{cases} \tag{1}$$

Step 2. Calculate the strength of i with (2). In the formula, M_1 means M_p and M_{-1} means M_n; CT_i is the threshold for pattern count and RT_i is the threshold for the ratio of the two kinds of match-count. CT_p and RT_p are for positive candidates; CT_n and RT_n are for negative candidates. P stands for the penalty for nouns since we find that nouns are more likely to be neutral; its value is 1 if the new word is not a noun. If strength $S > 0$, we get the word's sentiment polarity with (3). The strength values between different words can show the relative confidence.

$$S(i) = \begin{cases} \frac{M_i - CT_i}{M_p + M_n}\left(\frac{max(M_p, M_n)}{min(M_p, M_n)} - RT_i \cdot P\right), & M_i > CT_i \\ 0, & otherwise \end{cases} \tag{2}$$

$$\text{polarity} = \begin{cases} positve & i = 1 \\ negative & i = -1 \end{cases} \tag{3}$$

Step 3. If step 2 fails, we check the count of positive and negative neighbor words in (4); they are denoted as N_p and N_n. Similar to step 2, N_1 denotes N_p, N_{-1} denotes N_n, NCT_i is the threshold for neighbor word count and NRT_i is the threshold for the ratio of the two kinds of neighbor word count. NCT_p and NRT_p are for positive candidates; NCT_n and NRT_n are for negative candidates. P has the same meaning as the one in step

2. Also similar to step 2, we make i 1 first if $N_p > N_n$, -1 otherwise.

$$S(i) = \begin{cases} \frac{N_i - NCT_i}{N_p + N_n}\left(\frac{max(N_p, N_n)}{min(N_p, N_n)} - NRT_i \cdot P\right), & N_i > NCT_i \\ 0, & otherwise \end{cases} \quad (4)$$

If the strength value is greater than 0, we compute the word's polarity with (3); otherwise, the sentiment polarity of the word is neutral. Since all the new words can share the one-time scan of corpus, the algorithm complexity is O (n).

IV. Experiments

A. Experimental Setup

The corpus we used contains 10 million Sina Weibo tweets which are downloaded raw tweets containing elements that will be processed later. The corpus covers almost all the topics, making it a good dataset for experiments and comparisons. The old dictionary contains 77457 regular words but no newly invented ones. The sentiment dictionary has 8038 positive Chinese words and 6769 negative ones; this is a relatively complete sentiment dictionary.

B. Experiment Analysis

After the tweets are cleaned, NLPIR finds out the out-of-vocabulary words that it does not contain. Eventually, 3008 words are collected. Some examples of the words are:

非诚勿扰(a TV show), 雅诗兰黛(a makeup brand),

淘宝(an e-commerce website), 团购(group purchase),

情何以堪(cyberword, means "hard to bear")

With the help of the 3008 words, NLPIR has a satisfactory performance in segmentation and POS tagging. We pick out the new words using method in the previous section. Table II shows the distribution of words after every filtering. We observe that most of the low frequency words are nouns. And people tend to invent more nouns than verbs or adjectives.

We labeled the new words for evaluation. Every word is labeled by three persons and its polarity is determined by voting. Among the 6048 new words, we have 5143 neutral ones, 453 positive ones and 452 negative ones. The positive and negative words are called sentiment words. We evaluate the result by F-score, as defined in (5). Accuracy is the number of correctly classified sentiment words divided by number of words the system classified as sentimental. Recall is the number of correctly classified sentiment words divided by number of sentiment words in labeled data.

TABLE II
DISTRIBUTION OF WORDS AFTER EVERY FILTERING

Word class	original	Word count filtering	Old dic. filtering	POS filtering
total	335,760	86,853	33,331	6,048
noun	288,351	55,988	29,887	4,990
verb	24,574	20,166	1,162	903
adj	4,482	3,862	89	83
others	18,353	6,837	2,193	72

$$F\text{-}score = \frac{2 \times accuracy \times recall}{accuracy + recall} \quad (5)$$

As to baseline, we did not use Step 2, so only statistic information was used. Other parameters are the same as the run which achieves the best performance below. We obtain an F-score of 0.22 for baseline. Fig. 2 and Fig. 3 illustrate how penalty, NRT_p and RT_p affect the F-score. Some factors are fixed in the runs.

CT and *NCT* and fixed because they are simply thresholds, once set to proper values, they are not important. RT_n and NRT_n are fixed because experiment shows they have small influence. We think the reason is that this method is not very proper to handle negative words; unlike positive words, negative words appear in various contexts, making them difficult to discriminate. To handle negative words, more complicated method should be proposed.

The penalty is very effective for it reduces many errors in nouns. The figure shows that NRT_p and RT_p are useful; analysis shows the ratios should be a little high because even in neutral context there are much more positive words than negative ones, a high threshold can keep the balance. When NRT_p and RT_p are 3 and 5 respectively, the best F-score of 0.241 is achieved.

V. Conclusion and Future Work

In this paper, a hybrid approach considering both statistic and syntax information to discriminating the sentiment polarity of new words is proposed. We first filter the raw tweets out

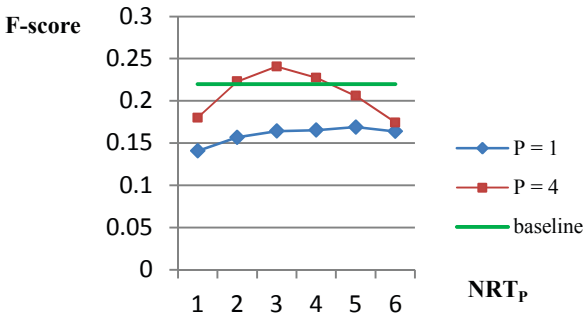

Fig. 2. F-score under different NRT_p
CT = 5 RT_p=6 RT_n=1 NCT=5 NRT_n=1

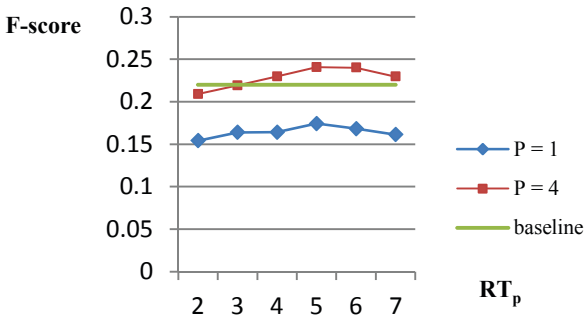

Fig. 3. F-score under different RT_p
CT = 5 RT_n=1 NCT=5 NRT_p=3 NRT_n=1

some noises and segment the clean data with POS tagging. Next, we collect new words by three successive filter rules. Then, we assign each new word with a polarity using both statistics and patterns information.

We evaluate our method on a real dataset achieving a relatively high F-score of 0.241 compared with the baseline of 0.22. From our observations, we conclude that the difficulties of this task comprise two points. One is that words with low frequency have little available context information, but most of the new words are in this case. Another is that the negative words appear in various contexts resulting in that many of the errors are caused by negative words.

As analyzed above, the result could still be improved in the future. For one thing, how to employ the insufficient context information is valuable to investigate. And the other, how to discriminate negative words is crucial to sentiment analysis.

REFERENCES

[1] D. Peter and Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in Proc. of ACL'02, 2002, pp. 417-424.

[2] D. Rao and D. Ravichandran, "Semi-Supervised polarity lexicon induction". in Lascarides A, ed. Proc. of the EACL, Morristown, 2009, pp. 675−682.

[3] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proc. of the EACL'97, Morristown, 1997, pp. 174-182.

[4] J. Wiebe, "Learning subjective adjectives from corpora," in Schultz AC, ed. Proc. of the AAAI, Menlo Park, 2000, pp. 735−740.

[5] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in Collins M, Steedman M, eds. Proc. of the EMNLP 2003, Morristown, 2003, pp. 105-112.

[6] P. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," in ACM Trans. on Information Systems, 2003, pp. 315−346.

[7] A. Andreevskaia and S. Bergler, "Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses," in McCarthy D, Wintner S, eds. Proc. of the European Chapter of the Association for Computational Linguistics (EACL), Morristown, 2006, pp. 209-216.

[8] F. Su and K. Markert, "Subjectivity recognition on word senses via semi-supervised mincuts," in Ostendorf M, ed. Proc. of the NAACL 2009, Morristown, 2009, pp. 1-9.

[9] Yuen and W. M. Raymond et al, "Morpheme-based derivation of bipolar semantic orientation of Chinese words," in Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 1008-1014.

[10] J. Wiebe and R. Mihalcea, "Word sense and subjectivity," in Dale R, Paris C, eds. Proc. of the Conf. on Computational Linguistics/Association for Computational Linguistics (COLING/ACL), Morristown, 2006, pp. 1065-1072.

[11] W. Y. Wang and D. L. Wang and S. Feng and R. F. Li and L. Wang, "An Approach of Building Microblog Smiley Emotion Lexicon and Its Application for Sentiment Analysis," Compter & digital engineering, vol. 40, no. 11, pp. 6-9, 2012.