Opinion Sentence Extraction and Sentiment Analysis for Chinese Microblogs

Hanxiao Shi, Wei Chen, and Xiaojun Li

School of Computer Science and Information Engineering, Zhejiang GongShong University, Hangzhou 310018
hxshory@foxmail.com

Abstract. Sentiment analysis of Chinese microblogs is important for scientific research in public opinion supervision, personalized recommendation and social computing. By studying the evaluation task of NLP&CC'2012, we mainly implement two tasks, namely the extraction of opinion sentence and the determination of sentiment orientation for microblogs. First, we manually label the sample of microblog corpus supplied by the organization, and expand the sentiment lexicon by introducing the Internet sentiment words; second, we construct the different feature sets based on the analysis of the characteristic of Chinese microblogs. Finally, we use SVM classifier to generate a model based on training corpus, and implement the predication of test corpus. Evaluation results show our work has good performance on two tasks.

1 Introduction

In the context of Web 2.0, microblog, as an important way for people to communicate with each other, has the characteristics of large amount and fast update. It is a reliable source to explore people's view and sentiment orientation. And the research on the natural language processing technology for microblogs has currently become a new research hotspot, of which sentiment analysis is an important topic.

The aim of sentiment analysis or opinion mining is to know people's opinion and sentiment orientation. At present the main technology is divided into two categories: one is to classify emotions according to the number of positive sentimental words and negative sentimental words in the text, with the method of combining sentiment lexicon with rules; the other is to adopt machine learning by which features in text are discovered and classifiers like Naive Bayes, Max Entropy and Support Vector Machine are applied.

At present English microblog research has made progress in sentiment analysis, for example, sentiment classification research for emoticons and hashtag as the features, while sentiment analysis research for Chinese microblogs relatively lagged behind either in resources or in methods.

By studying the evaluation task of NLP&CC'2012, we mainly implement two tasks, namely the extraction of opinion sentence and the determination of sentiment orientation for Chinese microblogs supplied by the evaluation organization. Experiment results show our work has good performance on two tasks.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 417–423, 2013.

2 Related Work

In the last decade, sentiment analysis attracted the attention of many researchers and has become a hot research topic in the field of information retrieval and natural language processing. As evaluation task is mainly about extraction of opinion sentence and determination of sentiment orientation, the following will introduce the related research from two aspects.

2.1 Subjective Text Recognition

Subjective text is the main object of text sentiment analysis. Therefore, it is very important to identify a large amount of subjective and objective web texts in advance to effectively narrow the analysis scope and reduce interference [1]. In order to effectively identify and extract factual information in an extraction task, Riloff et al. [2] designed a variety of filtering methods, using subjective classifiers to filter out subjective text, then extracted information, and tested in MUC-4 terrorism data sets; Toprak and Gurevych [3] finished a subjectivity classification experiment of English and French documents in DEFT'2009 Text Mining Challenge. In addition, Finn et al. [4] concluded that feature selection method based on POS tagging can get better effect than that on bag-of-words by studying subjective and objective sentence classification.

Above all, subjectivity text recognition is mainly based on sentiment words, using various methods of text feature expression and classifiers to classify. This method is clearly defined, and the essential problem is the selection of features.

2.2 Sentiment Orientation Analysis

In sentiment orientation analysis, a subjectivity text is generally divided into two or three types, namely positive, negative and neutral, to measure the preference of evaluated object. According to the different knowledge sources such as sentiment lexicon and corpus, and to different methods of classification, sentiment orientation analysis can be divided into the two kinds of methods.

The first method is based on sentiment lexicon and rules. It is to separately calculate the number of mixed sentimental words in text. If the number of positive word is greater than that of negative word, text is for positive sense, otherwise for negative sense, if equal for neutral. As the method relies on the quality of sentiment lexicon, the construction of sentiment lexicon should be focused on. Its main idea is: based on the collection of words with the known polarity as sentiment lexicon seed, new words with unknown sentiment polarity can be predicted by using some correlated method in the sentiment seed lexicon to find out similar words or similar semantic words, and then to calculate the new sentiment orientation through the polarity of these words. This method requires high coverage of the seed word. At present, the most commonly used lexicons are WordNet in English and HowNet in Chinese.

The second method is based on machine learning. With sentiment words and phrases, syntactic dependency and theme-related characteristics as classification features, a classification model is generated by classifier training. By using the constructed model, classification of test documents can be implemented in order to realize sentiment orientation discrimination of test documents. Pang et al. [5], by the bag-of-words technology, used Naive Bayes, Maximum Entropy and Support Vector Machines to do text sentiment polarity research, and made a comparative analysis of these three methods. Through experiments, the conclusion is that the performance of SVM is better, with the highest accuracy 80%.

The focus of the current Chinese sentiment analysis research is in the field of product reviews, news comments and film reviews. As for microblogs, it develops in recent years as a new type of social media while sentiment analysis research for microblogs is relatively less. A lot of people are working at Twitter sentiment analysis research in foreign countries. Go [6] and others used classifiers such as NB, ME and SVM to carry out the Twitter sentiment classification and the results show that the performance of SVM is better than other two classifiers. As to feature selection, features can be selected by the use of Unigram and Bigram models with the combination of parts-of–speech (POS). Barbosa and Feng [7], according to grammar characteristics of Twitter, observed its influence on the sentiment classification, considering forwarding messages, labels, links, punctuation, and exclamation marks, etc. Joshi [8] and others devised a sentiment analyzer to analyze and calculate microblog sentiment. Domestically, the study of microblog has emerged, but lack of further and advanced research. Basically it still adopts the traditional text analysis methods and it is short of depth analysis of features of micro blog, meanwhile, theory system is still not standard.

With the rapid development of microblog and the sharp increase of microblog users, the sentiment analysis for microblog should become a research hotspot with business, economic and cultural value.

3 Evaluation Task Analysis

Evaluation object is the micro blog provided by the organizers, including 20 topics and each topic has about 1000 items. Task 1 is to determine each sentence in micro blog is an opinion sentence or a non-opinion one. On the basis of task 1, task 2 is to judge sentiment orientation of the opinion sentence in task 1, including positive, negative and neutral. According to these two tasks, we first analyzed the sample text, and implemented the specific task decomposition.

3.1 Opinion Sentence Extraction

First we can consider extraction task of opinion sentences as a machine learning task, which is on how to make use of the existing micro blog sample to do manual label and feature extraction, construct model with the corresponding machine learning method, and then use the generated model to predict the test set.

According to the rules in NLP&CC '2012, opinion sentence does not include the sentences of sentiment self-expression, for example, "I am very happy.", defined by evaluation, this sentence is sentimental but not an opinion sentence. Opinion sentence defined by evaluation is confined to the evaluation of other objects, not including the inner self sentiment. So we have to design the corresponding rules to filter out this kind of sentence patterns. Based on the characteristics of micro blog corpus, we have carried out three steps to pre-process micro blog corpus. The first step is to remove the topic labels. The second step is to determine Chinese word segmentation. The last step is to filter self sentiment sentences by rules. Then as for the filtered sentences, we extract features, use the training set to train SVM model and predict the test set, and get the results of opinion sentences.

This article uses the word segmentation cloud services (research version) supplied by HYLANDA corp. Due to the input and output of results as XML format, it is very suitable for evaluating the processing of corpus. After word segmentation, we first filter subjective expressive sentence from its results in line with the requirements of opinion sentence identification.

Finally, we extract nine features as a feature set to judge opinion sentences, such as whether to contain an emoticon, a sentiment word, number of sentimental words, an exclamation point or a question mark, a consecutive exclamation mark or a question mark, inversion words, a degree adverb, a modal particle, and the network language, etc.

3.2 Sentiment Orientation Analysis

Task 2 is to determine sentiment orientation of micro blog, including positive, negative and neutral. Evaluation of this task is based on task 1.

We had a feature selection of sentiment classification mainly on the basis of the feature of sentiment words in combination with the characteristics of the micro blog. There are seven features mainly considered, such as the number of positive and negative emoticons, number of positive and negative sentiment words, the existence and non-exsistence of inversion word (the premise: the word should be in front of sentiment words), question mark and continuous question marks, etc.

First is manual labeling training. Sentiment polarities are divided into positive, negative and neutral. It is observed that the considered neutral micro blogs generally do not contain the sentiment words. Second, by using SVM classifier for training micro blog of three sentimental polarities, the corresponding model is generated. Finally, with the classification model, the public test set provided by the organization can be predicted.

4 Experimental Results

On the basis of some existing sentiment lexicons (such as the sentiment analysis words set in "Hownet", "Tongyici Cilin" provided by HIT IR-Lab), we extract a basic polarity vocabulary of 7926 words, 1993 words for positive sentiment and 5936

words for negative sentiment. In addition, we establish a degree adverb lexicon and a polarity shifting word lexicon. The adverb lexicon mainly collects 219 Chinese degree level words in the sentiment analysis words set in "Hownet". Due to the relatively limited number of polarity shifting words, the polarity shifting word lexicon is constructed mainly through manual collection and develops with the help of "Hownet" and "Tongyici Cilin ".

Based on this, we also extend the existing sentiment lexicons, such as increasing the cyberword: 51 for positive and 405 for negative, and increasing expression symbol lexicon: 84 for positive and 46 for negative, so as to solve the problem that the current word segmentation system can't distinguish some network language and emoticons. Moreover, there are modal particles and subjective word lexicons, the former is given that the tone of the users would affect the judgment of microblog opinion sentence; and the latter can help to filter out the sentences that self-sentiment expressing sentences do not belong to the opinion sentences, according to the evaluation sets.

4.1 Opinion Sentence Extraction Experiment

(1) Training set construction and training model generation

Training set is constructed by extracting the features from microblog corpus about 1219 microblogs labeled manually, and then converting them into SVM training format, including the following steps:

- Extraction of the features of network language emoticons

The part of the work is to extract network language and emoticons from the manual labeled corpus of microblogs. The present word segmentation technology is not able to identify the network language and emoticons, therefore it should be done before word segmentation. The way to extract is to do text matching consulting the content of the network language lexicon and network emoticons lexicon, recording the corresponding features whether being contained or not, and the number of occurrences.

- Word segmentation

Word segmentation includes the following steps:

1) to convert the corpus that needs to be segmented into XML format in accordance with HYLANDA corp.'s cloud segmentation rules, and to set the corresponding parameters;

2) to request HYLANDA corp.'s cloud segmentation word API to operate on segmentation;

3) to obtain word segmentation results and save them.

- Extraction of features from the segmentation results, and training set generation

This part of the work is to use the XML parsing technology to analyze segmentation results, extract features from segmentation results according to the existing sentiment lexicon, and convert them into the format that can be trained by SVM. We label 0 and 1 for the features whether being contained or not, here 0 for non-opinion sentences and 1for opinion sentences. At this point, we get a training set containing 1219 data.

◆ Training model generation

By using SVM classifier to train the acquired training set, the training model is generated.

(2)Test set construction and results

Test set construction is similar to the previous process. First of all, the predicted microblogs are preprocessed. In addition, to predict microblog corpus also needs word segmentation. It converts into the corresponding format, and then predicts it by using the previously generated training model. We respectively predict 20 themes of microblog and 20000 sentences, and the results obtained are shown in table 1.

Table 1. The evaluation results of task 1

Micro-average			Macro-average		
Precision	Recall	F-Measure	Precision	Recall	F-Measure
0.645	0.959	0.772	0.649	0.960	0.770

The evaluation results of the task in NLP&CC'2012 rank in top 3 in all evaluation teams, the performance on the Recall is particularly great. The reason of our good performance is that we carried out effective sentiment lexicon expansion as well as the relevant preprocessing, which make the results more comprehensive.

4.2 Sentiment Orientation Analysis Experiment on Microblogs

This task is to determine sentiment orientation of microblogs, including positive, negative and neutral. Evaluation of this task is based on a task 1, namely to analyze sentiment orientation of the opinion sentences. So before task 2, according to the result of task 1, we need to extract opinion sentences in microblogs as the test set in task 2.

(1) Training set construction and training model generation

In this part, extract 451 microblogs labeled opinion sentences from 1219 microblogs labeled manually as the training corpus. Specific process is familiar with the training set construction in task 1, also including extraction of network language emoticons, word segmentation, feature extraction, and training model generation. At this point, we get the training set containing 451 data, and its model after training.

(2) Test set construction and the results

The part of the work uses the microblogs predicted as opinion sentence in task 1, with reference to the previous method of training set construction to build the test set. We evaluate 20 themes of microblog for sentiment orientation and the results are shown in table 2.

Table 2. The evaluation results of task 2

Micro-average			Macro-average		
Precision	Recall	F-Measure	Precision	Recall	F-Measure
0.804	0.771	0.787	0.809	0.778	0.793

The evaluation results of the task in NLP&CC'2012 rank in top 2 in all evaluation teams. Compared with the results from other units, the Recall is still our advantage. We carried out rich features as well as sentiment sources, which make the results more comprehensive.

5 Conclusion

The social network represented by microblog in recent years made rapid development and sentiment analysis task in microblog attracted people's attention. In the evaluation task, our team makes full use of natural language processing technology and machine learning, as well as optimize feature design and process scheme. At last we obtained the ideal results in the evaluation task 1 and task 2 of NLP&CC'2012. In our future work, we will continue to study especially on social events in microblog discussion, on how to effectively analyze the orientation of events, emotional state recognition and degree of emotion recognition, etc. As time goes on, we can even study the changing process of netizen's sentiment orientation of a social event in the process of spread.

Acknowledgement. This paper was supported by the Zhejiang Provincial Natural Science Foundation of China (grant no. LY13F020007, LY13F020010, Z1110551), the Humanity and Social Science on Young Fund of the Ministry of Education (grant no. 12YJC630170), and the Science and Technology Department of Zhejiang Province of China (grant no. 2011C23075).

References

1. Yu, H., Hatzivassiloglou, V.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: Proceedings of EMNLP 2003, pp. 129–136 (2003)
2. Riloff, E., Wiebe, J., Phillips, W.: Exploiting Subjectivity Classification to Improve Information Extraction. In: Proceedings of AAAI 2005, pp. 1106–1111 (2005)
3. Toprak, C., Gurevych, I.: Document Level Subjectivity Classification Experiments in DEFT'09 Challenge. In: Proceedings of the DEFT 2009 Text Mining Challenge, pp. 89–97 (2009)
4. Finn, A., Kushmerick, N., Smyth, B.: Genre Classification and Domain Transfer for Information Filtering. In: Proceedings of the 24th BCS-IRSG European Colloquium on Information Retrieval Research: Advances in Information Retrieval, pp. 353–362 (2002)
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002, pp. 79–86 (2002)
6. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical report, Stanford (2009)
7. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36–44 (2010)
8. Joshi, A., Balamurali, A.R., Bhattacharyya, P., Mohanty, R.: C-Feel-It: A Sentiment Analyzer for Micro-blogs. In: Proceedings of ACL 2011, pp. 127–132 (2011)