Research on Building Chinese Blog Semantic Lexicon Based on Semantic Comprehension

Bin Wen1*, Huihui Dong1, Ling Ding2, Wenhua Dai2

1College of Information Engineering, Hubei Light Industry Technology Institute, Wuhan, Hubei, 430070, China
2Department of Computer Science, Hubei University of Science and Technology, Xianning, Hubei, 437100, China

vanbrian@hbust.com.cn

Abstract. Nowadays, investigators of phrases and text sentimental orientation identification mainly divided into two ways: semantic comprehension and machine learning. Machine learning needs a lot of training corpus and also cannot handle general field words effectively, meanwhile, semantic comprehension cannot get high scores at precision and recall. Therefore, in this paper, we propose a new fusion method based on our semantic comprehension for judging phrases’ polarity. In this paper, firstly, we modify the knowledge of Hownet, on the basis of its four primitives, the fifth primitive—sentimental primitive was proposed by our research, and annotated it into Hownet manually; secondly, we propose a new phrase sentimental similarity calculation method to compute word’s sentimental value; at last, we integrate transductive learning into this method to identify phrases’ sentimental orientation. The performance of experiment show that, compared with SVM or traditional semantic comprehension, it can get better results.

1. Introduction

With the rapid development of the World Wide Web, more and more users are willing to share their views and opinions on various platforms such as Blog, forums and BBS, it’s difficult to manage these such massive data manual. Then, it’s an urgent need for relevant natural language processing technology to deal with evaluation information. Especially of blogs, which has the special property of self-propagating, can take hot issues faster than other mediums. And each article also has a subjective point in whole document. So, this socialization medium tool has many challenges to process and attract many attentions nowadays. Take advantage of text sentimental orientation, investigators can get many useful information, on the basis of it, related organizations and enterprises can acquire users’ sentimental status, social phenomenon attitudes and merchandise reviews.

Recent researches reveal that, the relationship between vocabularies could be measured in a value; sentimental orientation of phrases or sentences is important for expressing sentiment orientation. Thus text sentimental orientation is based on sentimental orientation of phrase. Phrases’ sentimental orientation methods are divided into two categories: one is statistics method and the other is semantic comprehension. Machine learning(include deep machine learning) is the main method, which is usually used by investigators in nature language process.

Chinese blog is a few of differences from other platforms. Based on Web2.0, it is for conveying user-relationship for opinion sharing. In Chinese blog, the author expresses his views on various industries
and analyzes their ups and downs. It’s important to catch on author’s opinions by machine learning. It is urgent to discuss the blog in the phrase level. Thus, the other semantic comprehension method can be used to analysis blogs.

We are trying to establish a Chinese blog semantic dictionary for dividing Chinese-blogs into positive groups or negative groups. First of all, we discussed HowNet knowledge database, although it can express phrases’ semantic well, but it doesn’t contain phrases’ semantic similarity. Then, we improved the calculation of computing phrases’ value; meanwhile, emoticons also have semantic orientation in blog texts, it also need be discussed for getting author’s opinions by emotions; finally, with rapid development of network works, it is hard to discover new words by machine, so, new words should be discovered manually.

This paper is composed as follows: firstly, we introduce related works about sentimental classification of phrases, secondly, according to section II, we establish the semantic dictionary of Chinese phrases of blogs, and after that, we introduce our method about how to build semantic lexicon, and describe how it differed from others and effectively in section III. Fourthly, we make experiments for checking our procedure, finally, we gather work in a summary.

2. Background
Firstly, we introduce the previous work on phrase and text orientation. Phrase and text sentimental classification aims to divided phrases and texts into different sets: positive sets, negative sets or neutrality sets.

In the study of English, Turney[1,2] present a method named PMI-IR to classify words to different sets(positive set or negative set). It use “NEAR” by search engines(such as google.com) to count the correlation degrees of specific phrases and baseline phrases, and by this value we can judge the phrase’s orientation. It was proved effective in their experiments. Hatzivassiloglou[3] proposed a supervised learning method to judge the polarity of phrases. In their experiments, accuracy can get at 82 percent, on the basis of it, they integrate conjunctions across adjectives into their experiments, accuracy can up to 90 percent. Yu[4] establish adjective baseline sets which selected manually and each adjective also has strong sentimental orientation. New phrase’s sentimental orientation could be computed by co-occurrence probability of this word with benchmark phrases. Go[5] researched twitter’s text, which is very short, and divided into positive groups and negative groups take advantage of the method of distant supervision. Li[6] used a recurrent neural network method to classify Twitter text, which extract features by a filter without window size, Yin[7] proposed a convolutional neural network method in sentence modeling task based on attention mechanism. Liang[8] combined multi attention mechanism with convolutional neural network to deal with text emotion analysis, and achieve higher classification accuracy than single neural network model.

In Chinese, more and more investigators have great enthusiasm on blog application, and it’s in a trend of rapid growth. On the lever of phrase, Zhu. etc[9] put forward that, each phrase has a sentimental orientation weigh, which can be counted by the benchmark set which have very strong orientation and was picked up from HowNet and Internet manually. If phrase was determined into positive sets, it means it’s much more correlation with positive baseline set than negative set or neutrality set. In the same principle, it is negative or neutrality. Xie [10] experimented three methods, the first method based on emoticon, the second method based on the sentiment lexicon and the third method based on mix approach over hierarchical structure using SVM. And experiments proved that the third method can get the best results among these methods.

3. Methodology
Due to authors want to express their opinions in their blogs; so sentimental orientation of blog should be researched in many ways, especially in the method of semantic comprehension, which discussed in different fields about phrases, emoticons and sentences, respectively. Therefore, we attempt to build a Chinese blog semantic dictionary for promoting the results of sentimental orientation in blogs. In the following part of this paper, we will take over how to build Chinese blog semantic dictionary. And we
also want to prove how it effectively. As show in figure 1, our researches divided into two parts: phrases analysis and emotions analysis, Then, put them together.

3.1 Our Phrase Sentimental Orientation Algorithm
HowNet knowledge is an authoritative comprehension dictionary in Chinese. In this knowledge, each phrase have one meanings corresponding one concept. So, if phrase has different meanings, it may has the same concepts. The semantic similarity between different phrases means while two phrases have been exchanged in their environment of the context, how the original text altered without changing sentence’s meanings. In HowNet knowledge, there has four main primitives: the first is “first basic primitive”, the second is “other basic primitive”, the third is “relationship primitive” and the fourth is “symbol primitive”, in our research, we established a set of primitives named “desired/positive” and “undesired/negative”, which will take very important role in our phrase classification, and hasn’t got follow with interest before. It likes to reduce the importance of phrase’ sentimental similarity in computation. Then, we set this primitive as a new main primitive, called the fifth primitive “Sentiment Primitive” (SP), represented by $SP$ and counted by formula 1.

$SP(Phrase_1, Phrase_2) = \begin{cases} 1 & SM_{phrase_1} = SM_{phrase_2} \\ 0 & \text{others} \end{cases}$

Then, formula 2 was proposed for computing the new phrase semantic similarity (PhraseSS) between phrases.

$PhraseSS(Phrase_1, Phrase_2) = \sum_{i=1}^{12} \alpha_i \prod_{j=1}^{4} \left( Sim(P_i, P_j) + \alpha_i \cdot SP(Phrase_1, Phrase_2) \right)$

In our experiments, we set parameters $\alpha_1$ and $\alpha_2$ as 0.7 and 0.3 by which experiments can be stable. $Sim(P_i, P_j)$ and “first basic primitive”, “other basic primitive”, “relationship primitive” and “symbol primitive” denote the “first basic primitive”, “other basic primitive”, “relationship primitive” and “symbol primitive”. $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ must meet the conditions: $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ and $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \alpha_4$.

We get twenty-pair baseline phrases, which must take intense orientation in our experiments. Therefore, the phrases’ Sentimental Orientation Value of PhraseSOV is counted by the following formula.

$PhraseSOV(tp) = \frac{\sum_{i=1}^{n} WordSS(bm - p_i, tp)}{n} - \frac{\sum_{i=1}^{m} WordSS(bm - n_i, tp)}{m}$
Where “tp” is testing-phrase, \(bmp - p_i\) denotes positive group of phrases and \(bmn - n_i\) denotes negative group of phrases, and table 1 lists results before and after improved method of semantic similarity. “Value” is the method of before without the fifth primitive; and “improved-Value” is counted by our method.

| Unclassified Phrase | Polarity     | Value       | Improve-value |
|---------------------|--------------|-------------|---------------|
| Good                | Positive     | 1.235210    | 7.789160      |
| Pure                | Positive     | 0.361980    | 6.936698      |
| Hot                 | Neutrality   | 3.451588    | 3.451588      |
| International       | Neutrality   | 0.027563    | 0.027563      |
| Corruption          | Negative     | -0.86464    | -6.92650      |
| Bad                 | Negative     | -1.16433    | -7.11834      |

In our experiments, phrase “good” changed from 1.235210 to 7.789160, “corruption” also improved from -0.86464 to -6.92650. Accompanied by this obvious promotion, it is simple to find out phrases with polarities. PhaseSOV(tp) set threshold-value at +4.0 and -4.0 for making a distinction between positive phrase and negative phrase, which can achieve the best test results in our experiments.

In order to build Chinese blog semantic lexicon, we just considered the part-of-speech which may take sentimental orientation, such as adjectives, nouns and verbs, so, these three kinds of phrases is our research focus.

3.2 Emoticon Analysis
In Chinese blogs, authors may show their opinions by the usage of emoticons, which can make blogs much more directly and vividly. Now, emoticons are accepted and used by more and more authors in different platforms. In order to build the semantic lexicon, we get all of the emotions and divided them into positive, negative and neutrality manually. Then, we establish a emoticons lexicon named EL.

3.3 New Phrases Analysis
Because of the characteristics of Chinese phrases, a large amount of new phrases producing and spreading on Internet every year, even many of the new phrases have been put into the Xinhua Dictionary. We divided new phrases source into three categories by different reasons. One phrase category is the phrase without any orientation before, but after one hot issue, it gets semantic orientation. Such as “ni xing zhe”, “ye qing hui”. Another category, which always written by authors in wrongly written on purpose, also can take polarities. These phrases can make sentences or texts looks like vivid, therefore, many young authors are high on it. For example: “up” denotes hold on, and “xi fan” means like even love. The third category is the phrases which never emerged before in Chinese Xinhua Dictionary. Authors create them with means, and they were quickly accepted by public users because it's easy to see the meaning from the words. Such as “jiong” means embarrassed because of it looks like a people’s embarrassed face, “bu ming jue li” means we don’t know the theory but just feel it’s fierce, these phrases never appeared before in Xinhua Dictionary.

3.4 Evaluation Performance
In natural language processing, Recall-rate, Precision, comprehensive index value-F are used for assessing our experiments. As shown in the following formulas.

\[
R = \frac{\sum_{c_j \in C} \text{right}(c_j)}{\sum_{c_j \in C} \text{res}(c_j)}
\]
\[ P = \frac{\sum_{j \in c} \text{right}(c_j)}{\sum_{j \in c} \text{sen}(c_j)} \]  
\[ F = \frac{2 \cdot R \cdot P}{R + P} \]

Where, the \( \text{right}(c_j) \) denotes the blog texts which classified in the correct groups; \( \text{sen}(c_j) \) denotes the texts which classified to \( c_j \); \( \text{res}(c_j) \) references the texts which belong to the group of \( c_j \).

4. Experiment and Evaluation
A large number of blogs are available on internet; we got vast of blogs from blog.sina.com and blog.sohu.com, which are the two most popular platforms in Chinese. There are 5000 blogs downloaded from these two blog platforms in our experiments.

Our experiment procedures, for building blog semantic lexicon (we called it BSL), are took over as follows:

a) Firstly, we use ICTCLAS to parse and tag the part-of-speech, then we delete the stop words because of stop words have no orientations.

b) For adjectives, nouns and verbs may have semantic orientations, so we just extract these phrases, then sort them on the basis of their frequency of usage and deleted the words which are obtained incorrect obviously.

c) Calculating \( \text{PhraseSOV} \) by our method which introduced in paragraph 3 section 3.1 to determine the phrases into positive set, negative set or neutrality set. And then take all of the phrases into BSL with their orientations.

d) According to the regulation of paragraph 3 section 3.2, emoticons are bring into the semantic lexicon. Because emoticons’ orientation cannot be computed, so we complete it manually.

e) New phrases also need to obtained, we divided them into corresponding groups and then been joined into BSL.

f) If there has new words join into BSL, we go back to the step c); else exit processing.

Experiments are carry out by the steps mentioned above and after finish the experiments, experiments results are shown in the tables below.

| Orientation | Part-of-Speech | Total |
|-------------|----------------|-------|
|             | Adjective      | Noun  |       |
| Positive    | 3568           | 2897  | 6465  |
| Negative    | 1526           | 658   | 2184  |

| Source       | Positive | Negative |
|--------------|----------|----------|
| Blog.sina.com| 45       | 38       |
| Blog.sohu.com| 55       | 82       |

| Words        | Positive | Negative |
|--------------|----------|----------|
| New Phrases  | 48       | 67       |

Table 2 detailed our method’s results and it has been proved effectively; table 3 detailed emoticons with semantic orientation; and table 4 conveys new phrases with orientation implications. The data of
table 3 and 4 are need to be obtained by us. In BSL, phrases are in high frequency usage, but after our experiments, the words which are not often use with intense orientation also can join in our lexicon if the PhraseSOV is between +4.0 and -4.0 after transductive computing. Meanwhile, our method for computing phrases’ orientation is not consummate, several phrases maybe classified into the wrong group.

Additional, Chinese new phrases are emerge continually year by year, our semantic lexicon also needs to follow the tendency; so, we always need to add new words into our lexicon. therefore, new words which haven’t existed in lexicon, we must count its PhraseSOV value and bring them into positive group, negative group or neutrality group.

For proving our lexicon, we take 3000 blogs from blog.sina.com and blog.sohu.com, and classified them into positive sets, negative sets and neutrality sets. Detailed data is shown in the following Table 5.

| Corpus Type       | Orientation | Total |
|-------------------|-------------|-------|
|                   | Positive    | Negative | neutrality |
| Blog.sina.com     | 428         | 253    | 319        |1000 |
| Blog.sohu.com     | 521         | 218    | 261        |1000 |

Blogs text in Chinese need preconditioning, then we divided the text into positive group while the positive phrases are more than negative phrases; otherwise it may be put into negative group. Emoticons were regarded as phrases if they have orientations. We also considered the adverbs (included degree adverbs and negative adverbs) during judging the phrases’ semantic orientation.

| Blogs categories | Source       | P      | R      | F      |
|------------------|--------------|--------|--------|--------|
| positive         | Blog.sina.com| 73.2   | 80.5   | 76.5   |
|                  | Blog.sohu.com| 78.2   | 72.4   | 74.8   |
| negative         | Blog.sina.com| 84.9   | 78.5   | 82.8   |
|                  | Blog.sohu.com| 80.8   | 79.8   | 81.4   |

According to the Recall rate, Precision and F-value, we proved that it can get ideal results, the best results can up to 84.9% when mentioned to negative Chinese blogs about blog.sina.com, and the best average results of Recall rate is up to 80.5%. In the meantime, Recall rate and F-value are also much more better while comparing with other methods, finally, taking advantage of our Chinese semantic lexicon, it attested blogs can be classified into positive sets and negative sets more effectively.

5. Conclusion and future work

The discussion of blogs has caused may attention in nowadays; and blog sentimental orientation is becoming a challenging problem because of characters of Chinese. We put forward a method to build a Chinese blog semantic lexicon (called BSL), and with this lexicon, we classified blogs, which downloaded from blog.sina.com and blog.sohu.com, into positive group and negative group. Experiment results proved that it can get ideal results by our method. However, BSL is based on the frequency of usage, it may unsuitable for BBS, micro-blogs or other platforms. In the further work, we will improve our method to build an semantic dictionary with greater coverage.

References
[1] Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proceeding of the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., 2002.
[2] Vasileios Hatzivassiloglou, Kathleen R. McKeown. Predicting the Semantic Orientation of Adjectives [A]. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European.
[3] Turney Peter, Littman Michael. Measuring Praise and Criticism: Inference of Semantic Orientation from Association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[4] Yu H, Hatzivassiloglou V. Towards answering opinion question: separating facts from opinions and identifying the polarity of opinion sentences[c]//M. Collins and M. Steedman(eds): Proc. of the EMNLP-03: The 8th Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, July, 11-12. 2003:129-136

[5] Go A, Huang L, Bhayani R. Twitter sentiment classification using distant supervision[C]//CS224N Project Report, Standford, 2009.

[6] Li Song-ru, Chen Duan-sheng. Recurrent neural network using attention model for sentiment analysis [J]. Journal of Huaqiao University(Natural Science), 2018,39(2):252-255.

[7] Yin w, Schitze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4(10):259-272.

[8] Liang Bin, Liu Quan, Xu Jin, et al. A spect-based sentiment analysis based on multi-attention[J]. Journal of Computer Research and Development, 2017, 54(8):1724-1735.

[9] ZHU Yan-lan, MIN Jin, ZHOU Ya-qian, HUANG Xuan-jing, WU Li-de. Semantic Orientation Computing Based on HowNet[J]. Journal of Chinese Information Processing. 2006, 20(1): 14-20.

[10] XIE Lixing, ZHOU Ming, SUN Maosong. Hierarchical Structure Based Hybrid Approach to Sentiment Analysis of Chinese Micro Blog and Its Feature Extraction[J]. Journal of Chinese Information Processing. 2012, 26(1): 73-83.