International Journal of Hybrid Information Technology
Vol.8, No.9 (2015), pp.361-368
http://dx.doi.org/10.14257/ijhit.2015.8.9.33

Word Multi-Domain Semantic Polarity Algorithm Based on Domain Standard Word Set

Yun Sha and Shibo Zhang

School of Information Engineering, Beijing Institute of Petrochemical Technology
Beijing, China
School of Information Engineering, Beijing Institute of Petrochemical Technology
Beijing, China
E-mail: shayun@bipt.edu.cn

Abstract

Text orientation is one of the most important base work in web semantic calculation. The semantic polarity of words is the basis element of `text` *sentiment. But in traditional word orientation algorithm, each word gets orientation value based on its distance with these words in the standard word list. So the word polarity is always the same value. But in fact, the word may express different polarity in diverse domains. In this paper, word polarity is calculated based on domain standard word list according to its part of speech. The domain standard list is established by search engine result. The experimental result shows that the domain standard word set for word orientation can increase the accuracy rate of comments orientation calculate.*

Keywords: *Sentiment Classification; Word Semantic Orientation; Domain Standard Word List*

1. Introduction

The more we post on social sites like WeChat and Facebook, the larger the mounds of data we generate about our habits—what we like and dislike, interesting stories we've read, movies we've watched. How to use computer to acquire orientation of these evaluated texts, has been an urgent problem to be solved; And the word semantic polarity calculation is the base of text orientation calculation.

Usually, word orientation includes two aspects: polarity (positive or negative) and degree. For example: "very good" and "not bad" share the same polarity, but different degree. Word orientation calculation is to assign the word a signed real number, which sign express the polarity and value is the degree. There are two problems in word orientation calculation: 1). Words appear different polarity in different domain. For an instance, the word 'long'. In cellphone domain, if it is said 'stand-by time is long', the word 'long' is positive. While in the movie reviews, 'the movie is long', the word 'long' is a pejorative term. So lots of word should be put in a specific domain before we judge its orientation. 2).In many cases, a word has more than one part of speech in different context. And the polarity is different for the same word because the part of speech.

It is very necessary and significant to study word orientation in different domain and different part of speech. Usually most traditional methods of word semantic orientation calculation algorithms can be roughly classified into two directions: statistical method and knowledge library based method.

Statistical method is as following. Early in 1997, Hatzivassiloglou[1] made use of conjunctions to gain word orientation on adjectives which are joined by conjunction, and then cluster two kinds of adjectives by different emotional inclination. In 2003, Yu and Hatzivassiloglou counted concurrent probability between the word and a standard word

ISSN: 1738-9968 IJHIT

set of strong polarity to get the word's semantic orientation. In the same year, Turney and Litman[2] compute the word's similarity between labeled words to get the emotional polarity of words. Wang [3] focus on the role of orientation words in sentence context, instead of individual word only. A general analyze on the corpus proceeds to prove both prior orientation and the context impact are indispensable. But none of these algorithm considered the domain information when they calculated the word orientation.

While the method based on knowledge library calculate word similarity degree taking advantage of a knowledge library (such as HowNet or WordNet). In 2002, Kamps[4] used structure diagram of synonymous words to get semantic distance between the word to be estimated and selected standard word set to calculate the word's emotional polarity. In 2005, Zhu[5] presented two methods of word's semantic orientation based on meaning similarity grounded on HowNet and based on filed relating to semantics, and by comparing the two methods results, drew a conclusion that the method of word's semantic orientation based on meaning similarity grounded on HowNet had advantage over that based on filed relating to semantics and performed better on common words. Du[6] build an undirected graph in the use of word similarity computing technology first, and then partition the word-to-word graph by the idea of 'minimum-cut', thereby function optimization is adopted in this word semantic orientation computing framework and resolved by using simulated annealing algorithm.

The word polarity is calculated to judge the test orientation. There are two kinds of algorithms in traditional cross-domain sentiment classification: unsupervised [7][8] and semi-supervised [9 - 10].

Aiming to translate the classifier from the labeled domain to the unlabeled domain, these algorithms focus on the relationship between the domains. But the diversity of word orientation is not considered in them. In other words, the word level orientation for multi-domain is not considered in these cross-domain algorithms. On the other hand, the domain information is not considered in traditional word semantic orientation algorithms either.

Most of methods are based on a fixed standard word set without any domain information. The word orientation is calculated by the similarity of current word with each word in a fixed standard word set. Words in the fixed standard word set should achieve two principle: 1) the semantics of word in standard set must be clear; 2) the number of positive standard word and negative word should be balance.

But in these standard word sets, the domain and part of speech divergence is not considered. Moreover, a fixed standard word list can't adapt lingual diversity.

In this paper, a word orientation algorithm based on a domain standard word set is proposed. That is to calculate word orientation, different domain needs different standard word list.

The other part of the paper is organized as follow: the second section illustrates the method of domain standard word set construction; the third section describes in detail experiment by the method that is given in this paper and results are analyzed through comparing to the experimental results; the last section conclusion the work.

2. Construction of Domain Standard Word Set

The most problem of traditional word orientation algorithm lies in the accurate depends on the word number and quality in the standard word set. For one word may express different orientation in multi-domain, the standard word set can be a domain set according to specific domain.

The domain standard word set which are built based on domain data and the word's part of speech must follow a number of principles as follows:

1. Semantics of standard wordlist must express the current domain information and part of speech, which is ready for orientation judgment.

2. Domain standard word set must obtain from large-scale corpus and maintain two opposite balance sides.

3. Semantics of word in standard set must be enough clear (not having any ambiguity), and be marked by extreme intensity of emotions.

How to get the domain standard word set? In traditional algorithm, the words in the standard word set are selected manually. But the domain data can give us many information. In this paper, these domain standard words are selected according to their part of speech and domain specific feature.

In this paper, the word part of speech is also considered very important, for the adjective or verb in the text always show the text orientation. So two standard word sets should be established in one domain: adjective and verb.

In this paper, these words in domain standard word set are got by two source: 1) words which polarity are very clear. Such as ‘good’, ‘bad’, ‘excellent’ and etc. 2) domain feature word. Some word only has domain polarity. For instance, the word ‘long’, in cell phone domain, if it modify the ‘battery’, it is positive comment. But it may never occur as emotional word in some other domains.

The set D is domain text set:

$$D = \{D_1, \cdots, D_i, \cdots D_m \mid m \geq 1\} \quad (1)$$

The D_i is the i’th domain text set. $|D|$ is the number of all text in the data set, and the $|D_i|$ is the number of text in the i’th domain.

The domain standard word set is defined as:

$$W = \{(W_{1,adj}, W_{1,verb}), \cdots, (W_{i,adj}, W_{i,verb}), \cdots (W_{m,adj}, W_{m,verb}) \mid m \geq 1\} \quad (2)$$

In which, the subset $(W_{i,adj}, W_{i.verb})$ is the i’th domain standard word set. These words in $W_{i,adj}$ are the i’th domain standard adjective, while these words in $W_{i,verb}$ is the domain standard verb.

The domain feature word is constructed according to the ‘TFIDF’ value, which are adjective or verb. These adjective, verb are ordered by their ‘TFIDF’ value. These words have large ‘TFIDF’ value can be added to the domain feature word set (the domain standard list) separately. The k’th adjective in the i’th domain word is $w_{i,adj,k}$, which ‘TFIDF’ value is calculated as follows:

$$TFIDF_{i,adj,k} = TF_{i,adj,k} * IDF_{i,adj,k} \quad (3)$$

In which the $TF_{i,adj,k}$ is the $w_{i,adj,k}$ ’s Term Frequency in text. That is the times of the word occur in text. But the times of word occur in one text should be discounted when the text is very long. So the $TF_{i,adj,k}$ is calculated:

$$TF_{i,adj,k} = \frac{n_{i,adj,k}}{\sum_k n_{i,adj,k}} \quad (4)$$

In which the $n_{i,adj,k}$ is the k’th adjective occurring time in the i’th domain text. While the $\sum_k n_{i,adj,k}$ is the sum of all the adjective occurring time in the i’th domain.

$$IDF_{i,adj,k} = \log \frac{|D|}{1 + |t : w_k \in D_i|} \quad (5)$$

In which $|D|$ is the number of text in all domain. $|t : w_k \in D_i|$ is the number of text which $w_{i,adj,k}$ occurs in.

The $IDF_{i,adj,k}$ is the inverse document frequency. For the word is selected from the domain, so the $|t : w_k \in D_i|$ will never be 0. Then formula (5) can be calculated as:

$$IDF_{i,adj,k} = \log \frac{|D|}{|t : w_k \in D_i|} \quad (6)$$

The method of semantic orientation calculation of words based on domain standard word set using semantic similarity degree between current word and corresponding standard word set according to word's semantic orientation in its domain and part of speech.

Semantic orientation of words under domain standard word set build on semantic similarity degree of words, so it is necessary for using form of structural to describe complicated semantic information. HowNet [11] is a common knowledge corpus, which takes word and their relationship in Chinese and concept attributes. Therefore we compute word semantic similarity degree based on the sum of distances [12].

3. Experiments

The difference between the method proposed in this paper and the traditional algorithm is that domain standard word set is used in this method, while the domain information is not considered in the traditional algorithm. First, we explain our test data set, and then give the result of standard word list without domain information. After that, the compare of the fixed word set and the domain standard word set.

3.1. Experimental Setup

We construct two domain data sets from KTV comments and restaurant comments from http://www.dianping.com/. Then we classified these comments into positive or negative comments artificially. There are 144 positive comments and 50 negative comments in KTV domain. And 101 positive comments and 96 negative ones in restaurant domain.

In the paper, the test set for construct standard word set is derived from positive (negative) emotion word list and positive (negative) in the training data set KTV and restaurant comment. For the adjective and verb, they are divided into adjective sets and verb sets and words with other part of speech are excluded. There are 78 adjective and 65 verb in the KTV set; and there are 83 adjective and 54 verb in the electronic comment.

3.2. Selection of Domain Standard Word Set

Since this domain standard word sets are composed by two part of words: common polarity words and the domain feature words.

1). Common polarity words which have no domain feature. By filtrating and sorting, the test set for standard word set construction contains 2619 positive adjectives, 2085 negative adjectives, 776 positive verbs and 859 negative verbs, in total, 4704 adjectives, 1635 verbs, 6339 words in test set. These words are all from the 'HowNet', which is labeled as desired or undesired. These adjectives and verbs which used most widely is used in the standard word set. We use the www.yahoo.cn to get the number of these word return results, and these words are ordered by these number. The larger value is considered used more widely.

The first 20 positive and 20 negative adjectives are selected for the adjective standard word separately. The first 20 positive and 20 negative verbs are also selected.

2). for these adjectives and verbs in these test data set, the 'TFIDF' value are calculated according to the formula (3). Then the same number of domain feature can be extracted.

3.2. Domain Standard Word Set Used in Different Domain

For evaluate the word orientation algorithm, the KTV and Restaurant comment orientation is calculated by add up all word orientation value in a comment, if the result is a positive number, the comment is positive, otherwise if it is a negative number, the comment is negative, otherwise it is semantic neutrality. In this algorithm, no syntax is considered.

Figure1 shows that the orientation accurate of the KTV and Restaurant comments based on domain standard word set compared with the traditional standard word set. In this experiment, the numbers of word in the standard list are all set as 40 pairs. In the figure 1, "KTV" means KTV data set, "Res" means restaurant data set. The result of the word orientation based on traditional standard word set is called FIXST. The result based on domain standard word set is called DST.

The accurate rate of KTV comments orientation analysis based on DST is 71.65%, while 57.84% is based on the FIXST. And 68.53%, vs. 39.54% is for the restaurant comments. Clearly, the domain standard set performs better than the fixed standard set.

Figure 2 shows the impact of word number in the standard word on accuracy in the proposed algorithm. From the figure, we can see for the two data sets, the effects of word number in the standard word list on the accuracy were found negligible. So the conclusion is the number of word in domain standard word set has little effects on word orientation accuracy.

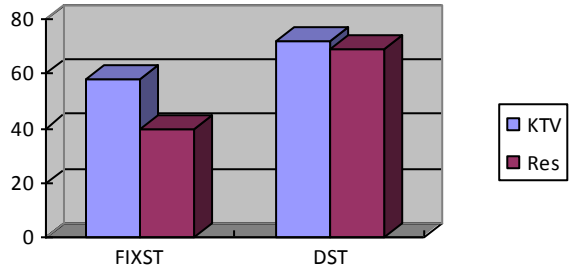

Figure 1: The Accurate Rate Based on Fixed and Domain Standard Word Set

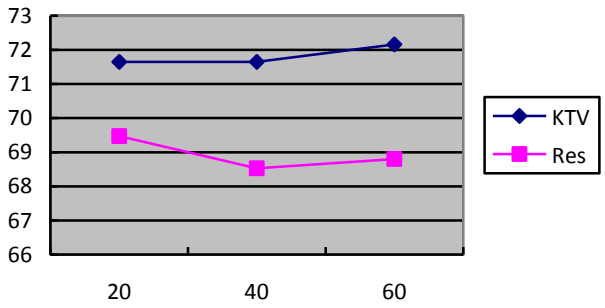

Figure 2: The Impact of Word Number in the Standard Word on Accuracy

4. Conclusion

Word orientation is the bases technology on text orientation. In this paper, an algorithm of word orientation calculation is proposed based on the part of speech and domain information. In experiment, a comparative study of the method of semantic orientation calculation of words based on different standard reveals that: the method of semantic orientation calculation of words based on domain standard word set performs better than that based on fixed word set. By contrasting the work on adjective test set and verb test set, it is found that: regardless of the method of semantic orientation calculation of words based on domain benchmark set or based on none domain benchmark word set, they show more effectively on adjectives than verbs.

In the paper, only adjective and verb are considered for standard word set. In future work, more parts of speech will be included in the domain standard word set.

Acknowledgements

This work is supported by the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions: 13031821005 and Beijing Municipal Commission of Education: KM201210017006.

References

[1] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and EACL'97, Eighth Conference of the European Chapter of Association for Computational Linguistics: pp. 174-181
[2] Peter D. Turney. Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, 2002.
[3] Gen WANG and Jun ZHAO, Orientation analysis of Chinese word. (In Chinese) http://nlpr-web.ia.ac.cn/2006papers/gnhy/nh4.pdf
[4] Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten D. Rijke. Using WordNet to Measure Semantic Orientations of Adjectives. LREC'04, Proceedings of the 4th International Conference on Language Resources and Evaluation, Vol. IV: 1115-1118.
[5] ZHU Yan-lan, MIN Jin, ZHOU Ya-qian, HUANG Xuan-jing and WU Li-de, Semantic orientation computing based on HowNet. (In Chinese) Journal of Chinese Information Processing, Vol.20 No.1, pp: 14-20.
[6] Du Weifu, Tan Songbo, Yun Xiaochun and Cheng Xueqi, A new method to compute semantic orientation, Journal of Computer Research and Development, 2009,46(10): 1713-1720.
[7] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen and Qiang Yang. Transfer Learning with Graph Co-Regularization. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12). Toronto, Ontario, Canada. July 22-26, 2012.
[8] WEI Xian-Hui, ZHANG Shao-Wu, YANG Liang and LIN Hong-Fei. Cross-Domain Sentiment Analysis Based on Weighted SimRank. (In Chinese) Pattern Recognition and Artificial Intelligence, Vol. 26, No.11, Nov.1004-1009, 2013.
[9] Dinko Lambov, Gael Dias and Veska Noncheva. Sentiment Classification across Domains. Progress in Artificial Intelligence, Lecture Notes in Computer Science, 2009 Vol.5816: 622-633.
[10] Shoushan LI and Chengqing ZONG. Multi-domain Adaptation for Sentiment Classification: using Multiple Classifier Combining Methods. NLP-KE'08. International Conference on Natural Language Processing and Knowledge Engineering: 1-8.
[11] Zhendong DONG and Qiang DONG.: 'HowNet', http://www.keenage.com
[12] Ming XIA, Yun SHA, Xiaohua WANG, Huina JIANG, Semantic orientation calculation of words with different part of speech, ICNC'11, The 7th International Conference on Natural Computation, pages:939-943, 2011.

Authors

Yun Sha

Female, 1976- , Gain doctor degree in 2004. Her research interest in the nature language process and Artificial Intelligent.

Shibo Zhang

Male, 1977- , His research interest in the nature language process and Machine Learning.