Determine the Polarity of Domain-specific Sentiment Words with Usage of Semantic Pattern of Sentences

YANG Li-gong
Department of Computer Science and Technology
Beijing Institute of Technology
Beijing ,China
e-mail: yyllgg@gmail.com

DENG Bo
Department of Computer Science and Technology
Beijing Institute of Technology
Beijing ,China
E-mail:dengbo-999@163.com

Abstract—**Text sentiment analysis is a new branch of computational linguistics which is widely concerned. In this paper, we present an approach to determine polarity of sentiment word based on context of sentence. We first change the context of sentence to semantic pattern vector, calculate the between different sentences, then compare sentences context indirectly by comparing similarity of their pattern vector, next we annotate polarity of sentiment word according to comparing result. Experiment shows that when the context of two sentences have high similarity, it is likely to have high precision in recognizing polarity of sentiment word. Our study shows it's feasible to use semantic pattern vector in representing context and judging polarity of sentiment words.**

Keywords—Sentiment word; Polarity; Semantic Pattern; Context

I INTRODUCTION

Text sentiment orientation is obviously influenced by the domain knowledge in this area. The context is also important to the orientation study. However, when we determine the polarity of sentiment words, there are not any effective mathematical models or methods that can be used to exactly describe the textual context of sentences now, and the establishment of domain knowledge base often requires a large amount of work. Hence, there are no effective methods using domain knowledge or context in determining the polarity of sentiment words now.

Yu[1] calculated the co-occurrence frequency of the new words and some words in seed words set by constructing the seed words set of sentiment words, but in this method, word pairs with identical polarity can only be determined through their high co-occurrence frequency, without considering the influences of text domain and context and possible changes of polarity of sentiment words caused by them. In fact, different words have different co-occurrence frequency in different domains, while as the context changes, the polarity of word varies significantly. Turney[2,3] adopted the statistical method of PMI-IR, which is still based on the idea of pointwise mutual information to study polarity of words by calculating the statistical dependency of some two words that are often used together in the text. This method is consistent with the basic idea in References [1], still belonging to the unsupervised learning method, and can deal with the polarity dependency among numerous sentiment words in aspect of statistics. However, Kamps [4,5] used the method based on the semantic dictionary, which calculated the semantic distance between candidate words and basic sentiment words mainly by using the synonymous structure diagram of WordNet, and then determined the semantic orientation of candidate words. This method uses the semantic knowledge of words to a certain extent, but it is not applicable for determining polarity of sentiment words under changing textual context as it does not consider the domain and textual context.

This paper provides a method for determining the polarity of sentiment words wholly from the aspect of textual context. The basic idea is to convert the textual context of sentences and represent it with semantic pattern vectors; indirectly study the degree of similarity of two sentences in the expression of textual context by comparing the degree of similarity of two sentences represented by semantic pattern vectors, and thus predict polarity of sentiment words contained therein.

II DETERMINE THE POLARITY OF DOMAIN-SPECIFIC SENTIMENT WORDS WITH USAGE OF THE SEMANTIC PATTERN VECTOR OF SENTENCES

The main difficulty to determine polarity of sentiment words through textual context of sentences lies in that the textual context is a relatively abstract and vague concept, with high subjectivity. Even for the same sentence, different readers will have different contextual understandings. But for the machine, as it has no appropriate background knowledge and adequate intelligence, it is less likely to understand the intrinsic semantic meaning as the mankind; in addition, there is no suitable algorithm or mathematic models that can be used to exactly depict and describe the semantic environment of sentences.

However, when we conduct the polarity discrimination of domain-specific sentiment words, though we cannot let the machine deeply understand the complex textual context of sentences, we can select a certain sentence A as the representative sentence with specific textual context. If there is another sentence B, which has the completely identical structure and expression with the said specific sentence A, that is, these two sentences are totally identical. Then they should have the same textual context, and their domain-specific sentiment words will surely have the same polarity. From this, it can be speculated that in a sentence C, if it's all words and the order of words in the sentence are identical with those of the specific sentence A and just the entity involved therein differs, then the sentence C surely

Published by Atlantis Press, Paris, France.

has the same context with sentence A and the meaning of expression of the former is almost identical with that of the latter. For example, sentence A: This little guy is so smart; and sentence C: This little boy is so smart. Comparing these two sentences, we can see that these two sentences are completely identical in all aspects except for the entity involved, which is "guy" in A while "boy" in C. Thus, it is known that these two sentences have the identical textual context, and the sentiment word "smart" in these two sentences has the same polarity, both positive.

Furthermore, if we take the expression of sentence A as a fixed semantic mode for expressing the specific textual context, and the polarity of sentiment words of A is definite, then comparing the semantic modes of sentence A and another sentence is just equal to compare the their textual context. After comparison, for the sentences are deemed with semantic modes similar to A, they will also have the similar textual context with A. Then it can be speculated that sentiment words of these sentences have the same polarity with those of sentence A. In such a way, polarity of sentiment words can be indirectly determined with help of the textual context.

According to the assumption above, we propose a method to indirectly determine polarity of sentiment words and polarity confidence level by calculating the similarity degree of semantic pattern vector of sentences.

The said method of forming semantic pattern vectors of sentences mainly adopts that used in References [6]. But in this paper, in order to achieve more accurate comparison results of the textual context of sentences, on the basis of referring to the similarity comparison method of semantic pattern vector used in References [6], we further orient and limit the comparison method. The comparison of semantic modes of sentences can be divided into three levels. The first level is to compare the syntactic structure of sentences, namely, the sentences should have the same language expression structure; the second level is to compare the quantity of identical words in the same position of the two sentences; and the third level is to compare the quantity of dissimilar words in the same position of any two sentences, as well as the similarity degree of such dissimilar words in aspect of semantics. The calculation method of semantic similarity of words still adopts that used in References [6], which is based on Synonym Dictionary [7]. Through comparison in these three levels, the similarity of textual context of the said two sentences can be accurately gained, and the matching process of specific semantic pattern vectors is as follows:

Step 1: By conducting conventional work like word segmentation, part-of-speech tagging, entity and type of entity tagging, convert sentences A and B into semantic pattern vectors, with the general form as follows:

$$t \ =(a_1,a_2,...a_r,entity1,b_1,b_2,...b_s,entity2,\ c_1,c_2,...c_t) \quad （1）$$

Wherein, the letter t means the semantic pattern vector of sentences, and a_r, b_s, and c_t mean the words respectively preceding, middle, and following the entity in the vector.

Step 2:First, compare structures of t_A and t_B (the semantic pattern vectors of sentences A and B), which is to mainly compare the entity type and entity position, and based on this, the similarity of these two sentences in this aspect can be calculated out, with the calculation formula as follows:

$$S_0(t_A,t_B)=\begin{cases}1 & t_A,t_B \text{with identical entity type and entity position} \\ 0 & t_A,t_B \text{with dissimilar entity type and entity position}\end{cases} \quad （2）$$

Step 3: In case that the calculated result of Step 2 is 1, respectively count the quantity of identical words of the three sub-vectors sited at the beginning, middle, and end of the two semantic pattern vectors, as well as their positions, in such way, we can obtain the similarity of the said sentences caused by identical words of sub-vectors.

If $f(t_A,t_B)$ represents the quantity of identical words in semantic pattern vectors of sentences A and B, then the calculation formula is as follows:

$$S_1(t_A,t_B)=\frac{f(t_A,t_B)}{r+s+t} \quad （3）$$

Step 4: Based on the calculated result of Step 3, respectively count the quantity of dissimilar words of the three sub-vectors sited at the beginning, middle, and end of the two semantic pattern vectors, as well as their positions, and with usage of Synonym Dictionary and the similarity calculation formula used in References [6], we can calculate out the similarity of such dissimilar words and further calculate out the similarity of the two semantic pattern vectors caused by dissimilar words.

According to the formula used in References [6], if $Sim(w_{tA},w_{tB})$ represents the similarity of dissimilar words appearing in t_A and t_B, then the calculation formula is as follows:

$$S_2(t_A,t_B)=\frac{\sum Sim(w_{tA},w_{tB})}{r+s+t} \quad （4）$$

Wherein, w_{tA} and w_{tB} respectively represent the dissimilar words appearing in semantic pattern vectors of two sentences.

Step 5:The results of Step 3 and 4 can be integrated into the total similarity of such two semantic pattern vectors, with the calculation formula as follows:

$$S(t_A,t_B) = S_1(t_A,t_B) + S_2(t_A,t_B) = \frac{1}{r+s+t}[f(t_A,t_B) + \sum Sim(w_{tA},w_{tB}) \quad （5）$$

With the help of the formulas above, we can calculate the similarity of semantic pattern vectors through judging the similarity of their structures.

On the basis of the above comparing similarity of two sentences' semantic pattern vector, we can define a threshold, such as, 0.5. When the similarity of two pattern vectors exceeds the threshold, it can be determined that these two patterns have very similar semantic environment thus it can be inferred that polarity of sentiment words appearing in sentences are identical. And the credibility of sentiment words polarity discrimination can be measured by the similarity degree of two pattern vectors. Namely, the similarity can be regarded as confidence level of polarity discrimination.

III EXPERIMENT AND EVALUATION

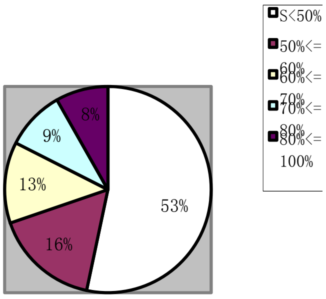
To compare the similarity of semantic pattern vector, we choose corpus supplied by the Third Chinese Opinion Analysis Evaluation (COAE2011) as corpus in our experiment. We selected five sentences from this corpus to form a basic pattern vector set. From this corpus we selected 600 sentences that are similar to the sentences in the basic set. These 600 sentences composed testing dataset of sentence to be determined. The next preprocessing works included word segmentation,part-of-speech tagging, entity recognition, sentiment word tagging for all the sentences and manual annotation of the polarity of all the sentiment words for later precision comparison. We used methods in above section to compare semantic pattern vector of sentences in basic set and sentences in testing set. By comparing similarity, we made judgment to polarity of sentiment words in testing sentences. Finally, we compared these identified sentiment words from 600 sentences in testing dataset with those manually annotated sentiment words. The relationship between sentences similarity and precision of sentiment words polarity discrimination is as follows:

TABLE 1: THE RELATIONSHIP BETWEEN SIMILARITY OF PATTERN VECTORS AND PRECISION OF SENTIMENT WORDS POLARITY DISCRIMINATION

	50%≤S＜ 60%	*60%≤S＜ 70%*	*70%≤S＜ 80%*	*80%≤S＜ 100%*
Sentence 1	0.426	0.551	0.764	0.852
Sentence 2	0.385	0.574	0.812	0.869
Sentence 3	0.463	0.485	0.836	0.920
Sentence 4	0.519	0.506	0.791	0.838
Sentence 5	0.367	0.394	0.685	0.759
Average	0.432	0.502	0.7776	0.8476

When similarity is compared, the number of each testing sentences occupies obviously different proportion under different similarity. Generally speaking, the higher the similarity is, the smaller proportion occupied by sentences is. We draw the following pie chart to show this relationship：

Figure1：the relationship of proportion occupied by testing sentences under different similarity



From the above figures we can see that, when semantic pattern vectors of 5 different sentences were compared with those of other testing sentences, the accuracy rates of sentiment words polarity discrimination are greatly different under circumstance of different similarities. Due to the strict requirements when performing fine matching of semantic pattern vectors, there are few sentences reaching very high similarity. Actually in this experiment, similarities of most testing sentences compared with basic pattern vectors are lower than 60%. Figure 1 shows the proportion of each part under circumstance of different similarity when compared with basic pattern vector. According to this figure, although testing sentences are selected manually, there are still more than half sentences' similarities are lower than 50%. It proves that it is rather difficult to find sentences with high similar context. On one hand it is because the number of sentences in pattern vector set we built is small, on the other hand, it is because the forms of expressions for sentences are flexible and diverse; therefore the matching of semantic pattern vector cannot reach high similarity.

In addition, from table 1 we can see that when sentences in testing dataset highly match with victors in basic set, generally, when the matching degree exceeds 80%, the precision of sentiment words polarity discrimination is higher. This shows that it is feasible to describe context of sentences by semantic pattern vectors of sentences and then determine the polarity of sentiment words.

Because of the limitation of human resources, we only use a small-scale of testing sentences to make this experiment. Obviously, it is necessary to promote such an experiment to a large scale of testing corpus so that the result will have universal significance.

IV CONCLUSIONS AND FUTURE WORK

Polarity determination of sentiment words is preliminary but important work in text sentiment analysis. The polarity of sentiment words differs greatly according to different fields of text and the context. This is the reason why in many applications we cannot dynamically determine sentiment words polarity to fit actual situation. In this paper we use the form of semantic pattern vector of sentences as the tool of describing the context of the sentence and determine polarity of sentiment words indirectly by comparing context of sentences. Although the current sentence scale is small and the precision of experiment result was not high, this paper originally proposes the method of using context to determine sentiment words polarity. The next work is to improve multivariate information contained by semantic pattern vector so as to express the context more accurately. Furthermore, we can adopt more appropriate matching algorithms to calculate the similarity of semantic pattern vectors.

REFERENCES

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

[2] P.Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the ACL, 417-424, 2002

[3] Turney, Peter D. and Michael L. Littman. Measuring praise and

criticism: inference of semantic orientation from association. ACM Transactions on Information Systems 2003. 21: 315-346.

[4]J.Kamps and M.Marx. words with attitude. In Proceedings of the First International Conference on Global WordNet, 2002,332-341

[5]J.Kamps and M.Marx, R.Mokken, and M.de Rijke. Using WordNet to measure semantic orientations of adjectives. In Proceedings of LREC, 2004

[6] Deng Bo, Fan Xiaozhong, Yang Ligong. Entity Relation Extraction Method Using Semantic Mode [J]. Computer Engineering, 2007, 10:212-214

[7] Mei Jiaju, Gao Yunqi, Li Yiming. Synonym Dictionary [M]. Shanghai: Shanghai Dictionary Publishing Press.1983