International Journal of Knowledge and Language Processing
Volume **2**, Number 1, January **2011**
www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp. **40-52**

An Unsupervised Approach to Sentiment Word Extraction in Complex Sentiment Analysis

Lin Li[1,2], Yunqing Xia[1], Pengzhou Zhang[2]
[1]Department of Computer Science and Technology,
Tsinghua University,
Beijing 100084, China
[2]School of Computer,
Communication University of China,
Beijing 100024, China
lilin20081@sohu.com, yqxia@tsinghua.edu.cn, zhangpengzhou@cuc.edu.cn

Received November 2010; revised January 2011

ABSTRACT. *Complex sentiment analysis starts to attract research interests as subjectivity analysis becomes hot in natural language processing area. The challenging issues in complex sentiment analysis are three. Firstly, there exist many categorization models while it is still unknown to us which categorization model is feasible for computing. Secondly, the indispensable linguistic data and knowledge to automatic sentiment analysis are rare. Thirdly, research work on complex sentiment analysis is very limited. This ascribes mainly to limited linguistic data and knowledge and diverse categorization models.*

In this work, Plutchik's sentiment model is investigated and finally selected as the categorization model. We further come up with an unsupervised approach to extract sentiment words from large-scale news corpus using a small number human-compiled sentiment word seeds. Experiments show that the approach achieves more than 50% on accuracy, which is very promising to reduce manpower in sentiment lexicon construction.

Keywords: Complex sentiment, sentiment analysis, sentiment word extraction, sentiment model.

1. **Introduction.** Natural sentiment is much more complicated and multi-dimensional. Increasing amount of researchers have been dedicating to exploring an effective approach to recognizing subjectivity from text. Differently, this work targets at analysis of complex sentiment such as anger and fear with in natural language text. With the public getting used to express their concern and emotion about certain events, texts on the Internet such as blogs, newspapers and reviews tend to carry natural sentiment. Complex sentiment analysis thus has become the major concern in natural language processing (NLP). Sentiment analysis is a big challenge to data driven method and it can be applied to various fields including information extraction, question-answering system, product review classification,

marketing forecast, recommendation system, public opinion analysis, and etc. There is no doubt that enterprises or governments can take advantage from knowing public's sentiment about their products or policy, and then they can make reasonable decisions.

Three tasks must be handled in complex sentiment analysis. Firstly, a categorization model should be selected and agreed. In the past decades, efforts have been made by psychologists to design the model while no agreement is reached so far [1]. In sentiment analysis research, systems are designed to classify sentiment as positive or negative [2]. Although the system with polarity sentiment model obtained an outstanding performance, it is too rough to depict sentiment accurately. There is no doubt that sentiment is subtle and exquisite. However, there exist many categorization models and it is still unknown to us which categorization model is feasible for computing. This work investigates on various models in existence and finally selects a computation-friendly model. Secondly, linguistic data and knowledge are indispensable to assist the study of automatic sentiment analysis. For example, lexicons should be compiled and a collection of natural text should be annotated based on the selected categorization model. The lexicons provide keywords that indicate certain sentiment while the annotations offer materials for empirical study. Unfortunately, the linguistic data and knowledge are rare. Thirdly, computing methods should be designed to make use of the data and knowledge to perform automatic sentiment analysis. A large number of approaches are designed to address subjectivity analysis, some of which are knowledge-based [3] while others are data-driven [2]. For complex sentiment analysis, very limited work is published. Knowledge-based approaches usually involve intensive linguistic study, making it less feasible at this moment to achieve the goal. In contrast, data-driven approaches adopt machine learning algorithms to induce empirical patterns from data and can be promising to predict class labels.

In this work, Plutchik's sentiment model was chosen because the model is not only widely-accepted but also adaptable to computing. Plutchik's model was composed of eight categories and three dimensions, and the original idea of his theory was derived from the Darwinian's theory of evolution. Plutchik's model was more complex and involved in a number of psychological concepts, so constructing a sentiment lexicon based on his model is not as easy as before. Meanwhile, traditional lexicon construction methods show some disadvantages such as high complexity and low accuracy because of subjective factors of taggers. For this reason, a sentiment word automatic extraction method was proposed and implemented in this work.

The rest of this paper is organized as follows. In Section 2, related work is reviewed. In Section 3, fundamental theory is given. In Section 4, method on sentiment word extraction is presented. We present experiments and discussions in Section 5 and conclude this paper in Section 6.

2. **Related Work**. Automatic analysis on complex sentiment is related to two research areas. First, psychological study on complex sentiment provides this work strong evidence on designing a categorization model. Second, efforts on sentiment lexicon construction are enlightening to this work. Related work is summarized as follows.

2.1. **Complex Sentiment Categorization Model.**Most traditional sentiment analysis systems adopted a relatively simple or empirical sentiment model. As we mentioned before, polarity sentiment model has been applied widely. Since the polarity model discarded the complexity of sentiment, it cannot reflect the sentiment consisted in text. For this reason, some researchers have adopted a more complicated model in their work. Xu and Tao (2003) propose to classify sentiment into 24 categories [4], which is derived from the seven sentiment concept in traditional Chinese culture. Combining modern psychological theory with Xu's theory, Chen (2008) proposes a model which has seven categories and applied it to constructing a lexicon [5]. Xia et al. (2008) applied Thayer's two dimension model to lyrics classification [6]. Thayer believed that every sentiment is comprised of energy and stress dimension, therefore according to his theory sentiment can be classified into anxious, contentment, depression and exuberance.

Sentiment is as diverse as color, making it is difficult to summarize and specify sentiment for the unprofessional. Therefore, we have studied a large number of psychological theories to find a sentiment model, which can depict sentiment comprehensively and concisely. Psychologists have proposed many categorization models since nineteenth century. Even now, no agreement has been reached among the researchers. Sentiment categorization models can be classified into two camps: categorical models and dimensional models.

A. The categorical models

Researchers who support the categorical theory believe that people possess basic sentiment and complex sentiment. Basic sentiment is innate, has its particular produce-processed and performance, and people share basic sentiment with other animal. Besides of basic sentiment, other sentiment is named complex sentiment which is derivative of basic sentiments. Ekman (1986) proposed a categorical theory based on his research on facial expression [7]. Ekman's model includes six basic sentiments: surprise, joy, sadness, disgust, fear, and anger. Other sentiment is the combination of two or more basic sentiments.

B. The dimensional models

Dimensional theory suggests that a sentiment is similar to a vector in space and comprised of definite attributes named dimensions. Thayer made use of energy and stress dimension to classify different sentiment into four categories (see Figure 1).

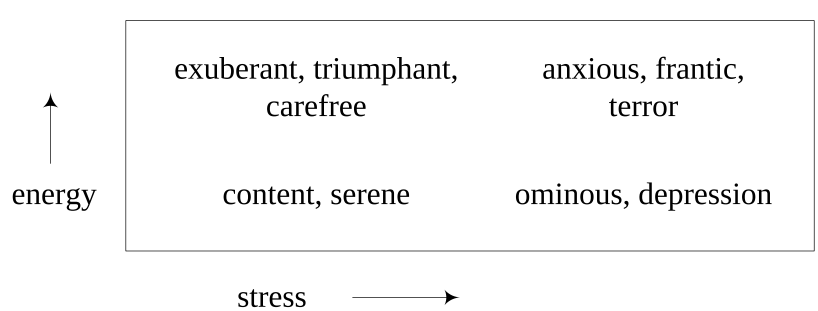

FIGURE 1. Thayer's sentiment model

Advantage of categorical theory includes that it is easier to be understood for its conforming to people's way of thinking. But sentiment is defined by descriptive approach, therefore different sentiments share common description. Moreover, the number of complex sentiment is undetermined. The overlapping sentiment definitions and undetermined quantity of sentiment categories lead to serious difficulties to computational linguistics. Comparing with categorical theory, dimensional theory is more suitable to quantitative sentiment analysis. Not only because the concise and informative definition of sentiment, but also less overlap among different sentiments. The use of abstract psychological terminology brings a few obstacles to researchers, but we believe that we can accept them gradually. We choose Plutchik's three-dimension model as the fundament of our work.

2.2. **Sentiment Lexicon Construction.** Psychologists suggest that the relation between sentiment and sentiment words is measurable. Hence sentiment lexicon is a useful resource to sentiment analysis. So far, sentiment lexicon automatic constructing usually is based on polarity sentiment model, in other words items are tagged as positive or negative. Wiebe et al. (1999) found that the distribution of sentiment words is uneven in different part-of-speech categories, and categorized sentiment of sentence by characters such as certain words, punctuation, and etc. [8]. Turney (2002) extracted certain phrase according to part-of-speech from customer review, and then evaluated its sentiment orientation by calculating co-occurrence [2]. Zhang (2007) examined Turney's theory in Chinese Xinhua corpus, and summarized part-of-speech patterns which probably contain sentiment words [9]. Chen et al. (2009) distinguished sentiment words by some Chinese grammar, and the accuracy of his experiment is 66.53% and the recall is 54.60% [10]. Zhu et al. (2005) calculate sentiment orientation of Chinese word by the semantic similarity and relevancy concept of HowNet [11]. Generally, constructing of sentiment lexicon can be divided into two main steps: (1) extract sentiment word from corpus; (2) estimate its sentiment.

Construction of complex sentiment lexicon mainly relies on manual work. Xu (2007) chose sentiment words manually from lexicons, and then classified those words according to their sentiment model [12]. Quality of complex sentiment lexicon is largely influenced by constructors, because everyone has their own judgment to sentiment. It is hard to guarantee consistent standards in lexicon construction so that the quality of lexicon is not stable and commonly used.

With a complex psychological model, we explore an automatic approach to sentiment lexicon construction. The major concerns are manpower and subjectivity in lexicon compiling. Our approach bases itself on a few seeds and seeks to extract sentiment words from a large collection of natural text.

3. **Theory.** The idea underlying our work is that we first carefully select a few seed of sentiment words for each category and attempt to extract more and more sentiment words from Gigaword using machine learning techniques.

3.1. **Assumption.** News report is released to the public to present the international or domestic events. It often accompanied with relevant comments [13]. Based on observations on a large volume of news reports, we come up with the following assumption

Assumption 1: *One piece of news focuses itself on one certain object and presents one* *particular sentiment. Hence the sentiment words that co-occur with the seed words reflect a same or similar sentiment.*

To guarantee validity of the assumption, we extract from our news collection the cases that probably do not comply with the assumption. Two observations are made.

(1) Background and examples are embedded in news text. To present the events more clearly, authors tend to spare text on background information and examples. However, the sentiment carried by the background information and examples might differ from that in the news. This brings noise to the above assumption.

(2) Rhetoric content is given to make the writing more attractive. In these cases, we find simile, metaphor, sarcasm and negation, which make sentiment of the words opposite to whole sentiment orientation.

Fortunately, the above cases occupy less than 5% percent. As expected, most news reports follow direct and definite writing style. Hence the exceptional cases give very limited influence on our approach. Based on above assumption, we propose to acquire sentiment of word using a statistical method.

3.2. **Sentiment Categorization Model.** Facing numerous psychological sentiment models, we think that an appropriate model should conform to two principles: (1) it is comprehensive so that it can cover all sentiment in Chinese (2) it defines sentiment clearly and definitely. Based on above principles we chose Plutchik's model.

Plutchik model is derived from evolutionism, which is a combined theory of dimensional and categorical theory. He believes that there exist eight basic sentiment categories including rage, terror, vigilance, amazement, ecstasy, grief, admiration, loathing. Meanwhile, orientation of a sentiment is decided by three dimensions: intensity, similarity and polarity. Plutchik model is shown in Figure 2.

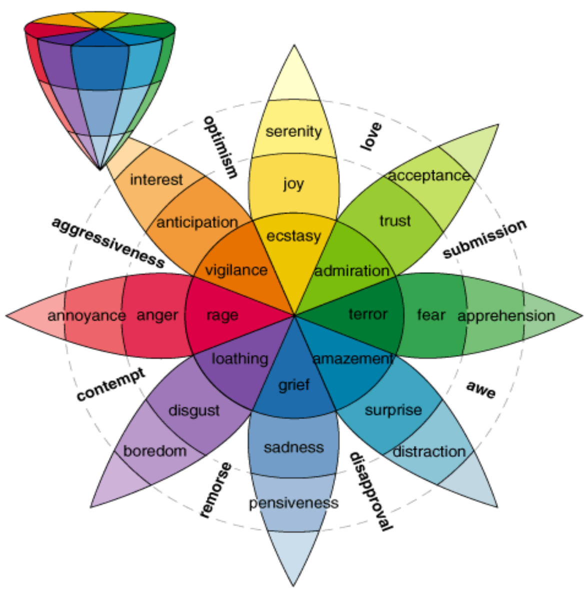

FIGURE 2. Plutchik's sentiment model

Comparing whit other models, Plutchik model has following advantages.

(1) Coverage

Plutchik believes that sentiment is expressed in different manners by different people, and complies with four main problems of survival, i.e., hierarchy, territory, identity and “temporal life”. Based on this belief, primary emotions are proposed: fear and anger, acceptance and disgust, sadness and joy, anticipation and surprise. As psycho-evolutionary theory of emotions is involved, complex sentiment can thus be further defined from perspectives of stimulation, cognition, subjectivity, behavior and function.

When coverage is concerned, Plutchik’s model can fully cover Ekman’s six-category theory [7], Chinese seven-category theory, Xu’s 20-category theory [4]. The sentiment class that other theories do not cover is anticipation, which is carried by keyword such as *anticipation*, *expectation*, *concern* and *interest*.

(2) Computing

Plutchik’s model involves three dimensions: intensity, similarity and polarity. For instance, *joy* is the polar opposite of *sorrow*, *terror* and *fear* are broadly similar, while *grief* and *pensiveness* vary in terms of intensity.

Quantification is a challenging task in complex sentiment analysis because not only class labels but also the relations between the labels must be defined. Plutchik’s model provides a natural approach to quantify the sentiment types. Considering every sentiment as one thee-dimension vector, sentiment of an unknown word can be computed by measuring similarity between the word and the sentiment lexicon.

To summarize, the Plutchik’s model shows a broader coverage and computing friendly. Thus we adopt the Plutchik’s model in this work as sentiment model.

4. **Sentiment Word Extraction.**

4.1 **The Idea.** The idea for sentiment word extraction is given below. We first segment each news article into paragraphs and sentences. Second, we segment each paragraph and sentence into words and assign each word a tag part-of-speech. Third, potential sentiment words are extracted based on statistical statistics. At last, we assign each word a sentiment label by calculating sentiment similarity between the word and the lexicon. The workflow of the method is given in Figure 3.

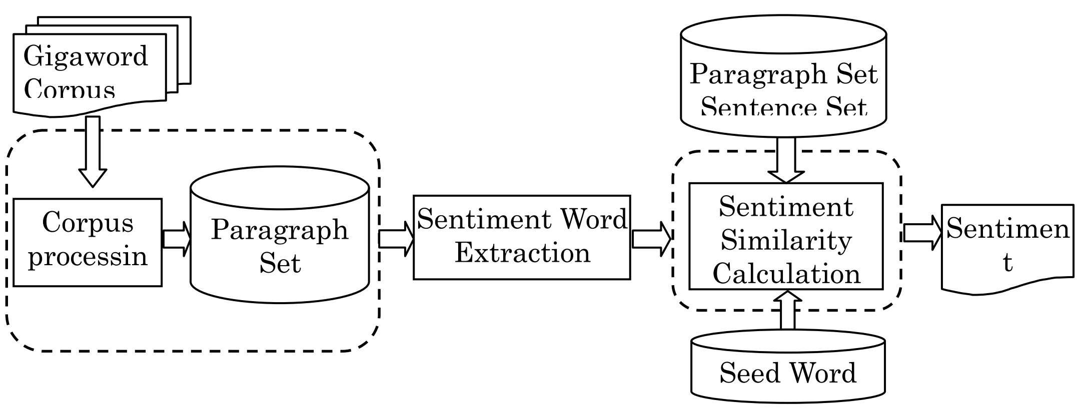

FIGURE 3. Workflow of sentiment extraction method

4.2. **Sentiment Word Seeds.** It is essential that we should choose the seed word which is unambiguous and can expresses keen and definite sentiment. In this work, we combined Plutchik's word list which contains 142 items with other lexicon resources like *HowNet* and *Modern Chinese Dictionary* to build sentiment word sets. For ensuring a high frequency of the seed words, we ranked the words according to their occurrence in *Gigaword*. Finally, we chose 100 seed words to start the sentiment word extraction. Some typical seed words are given in Table 1.

TABLE 1. Sample sentiment words

Category	Sentiment words
grief	悲伤 悲痛 伤心 难过 忧伤
amazement	惊奇 惊愕 惊诧 惊讶 意外
terror	惊恐 恐惧 惊骇 害怕 惧怕
ecstasy	狂喜 陶醉 高兴 愉快 喜悦
vigilance	警觉 警惕 谨慎 警戒 期盼
loathing	憎恨 憎恶 仇恨 厌恶 憎恶
rage	暴怒 狂怒 恼怒 生气 气愤
admiration	钦佩 欣赏 羡慕 信任 信赖

4.3. **Sentiment Similarity.** We define sentiment similarity to evaluate the similarity between a potential sentiment word w and all seed words. Let $K_i(i = 1, ..., 8)$ represents the number of seed words in each sentiment category, $s_{ij}(j = 1, ..., K_i)$ the *j*-th sentiment word, and $Sim(w, s_{ij})$ the sentimental similarity between word w and sentiment word s_{ij}. We obtain $Sim(w, s_{ij})$ based on two statistical methods: PMI (Point-wise Mutual Information) and LR (Likelihood Ratio).

(1) Improved PMI formula

$$PMI(w, s_{ij}) = (1 + \alpha) log \frac{c(w, s_{ij})}{c(w)c(s_{ij})} \quad \textbf{(1)}$$

in which $c(w, s_{ij})$ denotes the number of co-occurrences of w and s_{ij} in the corpus. $c(w)$ indicates the number of occurrences of w in the corpus while $c(s_{ij})$ that of s_{ij}. α is weight of the phrase pattern that contains word w. To model the context, two kinds of co-occurred ranges are adopted: sentence and paragraph.

(3) Improved LR formula

$$LR(w, s_{ij} = (1 + \alpha)\Big(log \frac{b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)}{b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)}\Big) \quad \textbf{(2)}$$

$$L(k, n, x) = x^k(1 - x)^{(n-k)} \quad \textbf{(3)}$$

in which c_1 and c_2 are number of occurrences of w and s_{ij} in corpus, respectively. c_{12}

is number of the co-occurrences of w and s_{ij}. N is total of words in corpus. α is weight of the phrase pattern that contains w. p, p_1, and p_2 are calculated as follows.

$$p=\frac{c_2}{N}, p_1=\frac{c_{12}}{c_1}, p_2=\frac{c_2-c_{12}}{N-c_1} \tag{4}$$

4.4. **Sentiment Strength.** Sentiment strength of a word is defined as sum of the sentiment similarity between the word and every seed word. Let $ES_i(w)(i=1,...,8)$ denote the sentiment strength of word w in the i-th sentimental category, which is calculated as follows.

$$ES_i(w)=\sum_{j=1}^{K_i} S(w,s_{ij}) \tag{5}$$

It appears that a word with higher sentiment strength usually expresses stronger sentiment. By observing the results, however, we find that many adverbs are mistakenly considered as sentiment words due to high co-occurrence statistics. For example, *"非常* (very much)*","极其*(extremely)*"*often co-occur with sentiment words. But they are not actually sentiment words. To filter the noise, an improved TF-IDF (Term Frequency-Inverse Document Frequency) is adopted as follows.

$$ES_i^*(w)=\frac{log\left(1+\frac{ES_i(w)}{\sum_{w\in W} ES_i(w)}\right)\times\sqrt{log(1+\frac{N}{DF(w)})}}{\sum_{i=1}^{8}\left[log\left(1+\frac{ES_i(w)}{\sum_{w\in W} ES_i(w)}\right)\times\sqrt{log\left(1+\frac{N}{DF(w)}\right)}\right]^2} \tag{6}$$

In Equation (5), $DF(w)$ denotes the total number of sentiment categories which contain w. This formula decreased sentiment strength of the noise so that the accuracy of our approach is improved.

Finally, we sort the potential sentiment words and select items from top to expand the sentiment lexicon.

5. Evaluation.

5.1. Setup.

Dataset

The Simplified Chinese Gigaword (Version 2.0) is selected as development dataset in our experiments, which is derived from the Xinhua News Agency. To evaluate the proposed method, we manually compiled five hundred sentiment words.

Evaluation metrics

The purpose of this work is to extract sentiment words from a large collection of text. So we favor accuracy rather than recall. Accuracy is calculated using the following formula.

$$Accuracy=\frac{count\,of\ sentiment\,words\,correctly\,extracted}{count\,of\ sentiment\,words\,extracted} \tag{7}$$

Evaluation method

Three parameters influence performance of our method:

1) Similarity measure: the measure can be PMI or LR.

2) Co-occurrence range for similarity measure: sentence or paragraph.

3) Number of seeds: 50 or 100.

We thus design eight experiments described in Table 2.

TABLE 2. Eight experiments that combine three parameters

Experiment Variables	**Experiment Methods**	**EX1**	**EX2**	**EX3**	**EX4**	**EX5**	**EX6**	**EX7**	**EX8**
Sentiment Similarity Calculating	PMI	❍	❍	❍	❍				
	LR					❍	❍	❍	❍
Number of Seed Words	50	❍	❍			❍	❍		
	100			❍	❍			❍	❍
Co-occurrence Range	Sentence	❍		❍		❍		❍	
	Paragraph		❍		❍		❍		❍

5.2. **Results.** Results of eight experiments are given in Table 3.

TABLE 3. Experimental results

	Vigilance	**Terror**	**Amazement**	**Rage**	**Loathing**	**Grief**	**Ecstasy**	**Admiration**
EX1	31.2%	15.2%	12.2%	33.8%	41.6%	22.2%	**35.8%**	45.2%
EX2	29.6%	15.0%	12.4%	30.4%	39.2%	19.0%	32.6%	39.2%
EX3	**37.6%**	**18.8%**	**19.4%**	**40.0%**	**41.8%**	**23.2%**	34.8%	**53.8%**
EX4	23.8%	14.2%	16.6%	34.2%	33.4%	17.2%	28.6%	39.2%
EX5	25.6%	17.0%	16.0%	30.4%	27.0%	18.6%	34.8%	42.4%
EX6	11.2%	8.0%	5.0%	19.0%	15.8%	5.4%	15.0%	18.2%
EX7	25.8%	13.6%	13.8%	25.2%	24.2%	13.6%	23.4%	34.2%
EX8	11.8%	10.2%	11.8%	22.4%	19.4%	9.2%	15.6%	22.2%

5.3. **Discussions.** We discuss how the parameters influence performance od the method as follows.

5.3.1. **Sentiment Similarity Measures.** This observation investigates effects of sentiment similarity measures (i.e., PMI and LR). According to Table 3, PMI outperforms LR significantly by 10.17% on average. Therefore, we conclude that PMI can reflect sentiment similarity more precisely.

To make the comparison between similarity measures more distinct, we compare results in experiment EX1 and EX5 in Figure 4. Both experiments use 50 seed words and sentence as co-occurrence range while they adopt different similarity measures. Seen from Figure 4, PMI outperforms LR in vigilance, rage, loathing, grief, ecstasy and admiration. The exception happens in terror and amazement, in which very low accuracy is achieved. This is because journalists tend less likely to use terror and amazement in their writings. It leads to a data sparse problem in sentiment similarity calculation. As sparse data cause more trouble in PMI than that in LR, it is unsurprising that a lower accuracy is achieved in these two cases.

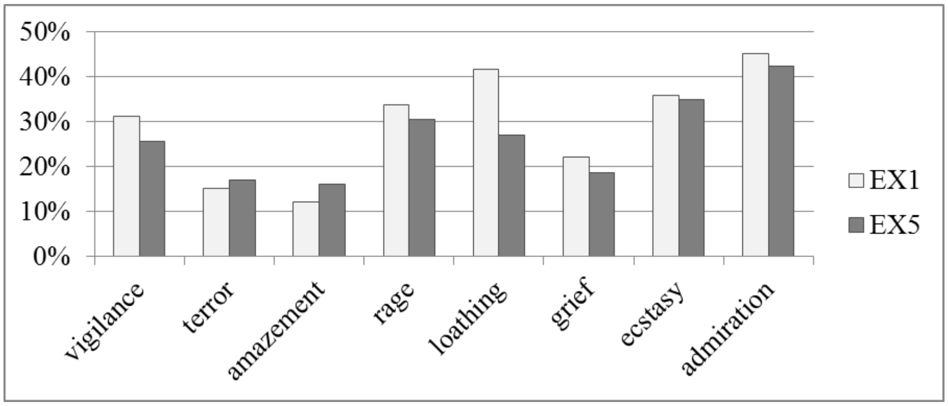

FIGURE 4. Comparison between similarity measures PMI and LR

5.2.2. **Co-occurrence Range.** We believe that the performance of our method with a smaller co-occurrence range would be better. Experiment results supported our assumption. Seen from Table 3, accuracy has been improved 7.73% in average.

To disclose influence of co-occurrence range, we present results in experiments EX1 and EX2 in Figure 5. Both experiments use 50 seed words and adopt PMI as similarity measure while they use different co-occurrence range. Seen from Figure 5, the method with sentence as co-occurrence range outperforms that with paragraph in all sentiment categories except *amazement*. We find the inferiority is very small, which can be viewed as noise. It can be safely concluded that sentence is a better range in finding co-occurring sentiment words. This accords to the normal observation that sentence is a minimum meaning unit while paragraph usually brings more noise in co-occurrence analysis.

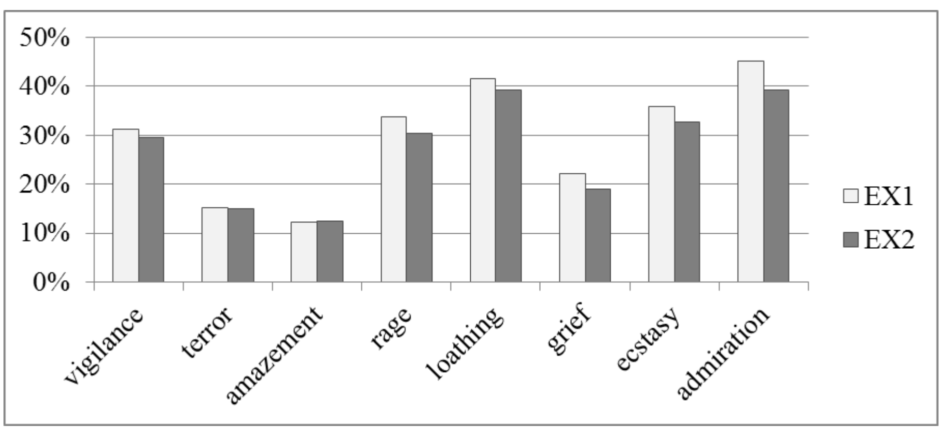

FIGURE 5. Comparison between co-occurrence range of sentence and paragraph

5.2.3. **Number of Seed Words.** We also believe that collecting more seed words should be helpful. Results in Table 3 show that the average improvement is not huge, i.e. 0.28%. Comparing with Zhu's work [9], the reason that we do not acquire significant improvement by increasing number of seeds is that the Plutchik's model is more complex. When number of seeds increases from 50 to 100, accuracy in category *rage* and *amazement* has been improved by 2.05% and 4.00%, respectively. Thus, we can conclude that increasing number of seeds is positive but its influence varies.

To reveal influence of seed number, we present results in experiments EX1 and EX3 in Figure 6. Both experiments use PMI as similarity measure and sentence as co-occurrence range while they differ from each other in number of seed words. Seen from Figure 6, the method using 100 seed words outperforms that with 50 seed words in all sentiment categories except *ecstasy*. We notice that the inferiority is very small, which can be viewed as noise. We thus can safely conclude that more seed words may bring us higher accuracy in sentiment word extraction.

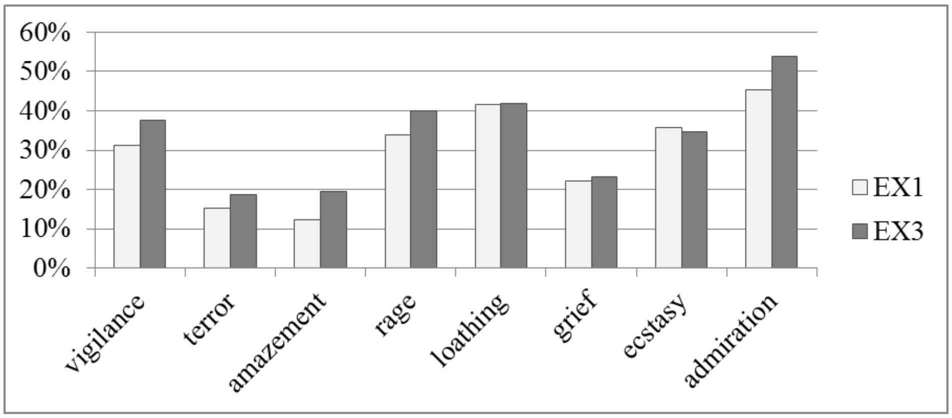

FIGURE 6. Comparison between number of seed words 50 and 100

5.2.4. **Sentiment Category.** Accuracy of the method varies significantly in different sentiment categories. To make the comparison, we present average accuracy of the method in either sentiment categories in Figure 7.

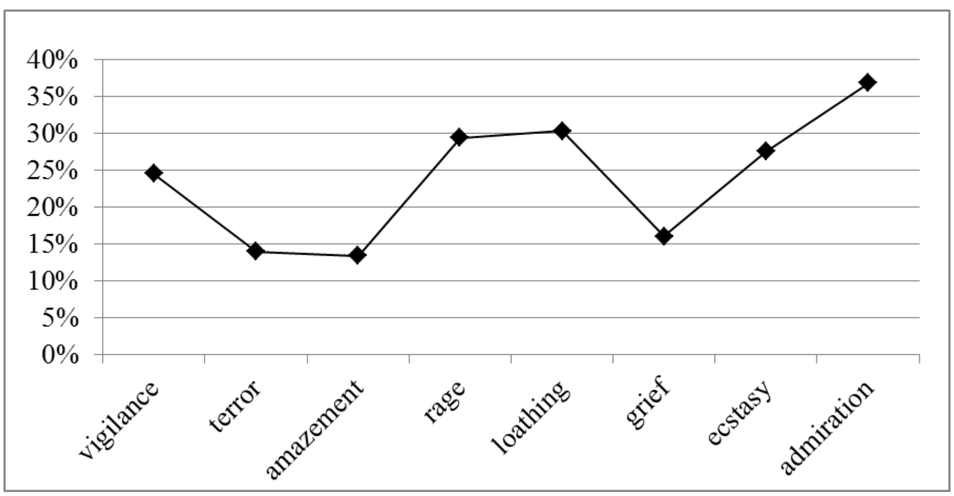

FIGURE 7. Comparison between sentiment categories on average accuracy

Seen from Figure 7, average accuracy in admiration is 36.80%, while the method yields about 15% in terror, amazement, and grief is approximate 15%. Observation shows that distribution of sentiment is not balanced/ For example, terror, amazement and grief appear much less than other sentiment categories within news corpus. Hence, the sentiment words carrying the three types of sentiment are less than that of other categories. It is the low co-occurrence that results in low accuracy.

5.3. **Error Analysis.** By analyzing results in our experiments, we find the following typical sources of errors.

1) Part-of-speech patterns

As of language is abundant and dynamic, a number of non-sentiment words will be included by the sentiment lexicon if we depend merely on part-of-speech patterns. For instance,

> *这项贷款由工商银行发放，对象包括省内所有开发型科研院所，主要用于新产品的研制、开发。*
>
> (Provided by ICBC, the loan targets at all province-wide institute s and should be used mainly in new product research and development.)

Phrase *"主要 用于*(used mainly in)*"* complies with *"adverb verb"* pattern. But neither is sentiment word. Research efforts will be conducted on refining part-of-speech patterns.

2) Sentiment similarity calculating methods

Neither PMI nor LR is able to estimate semantic relation between co-occurrence words. In fact, some of them are collocations rather than sentiment similarity. For instance, "*感到*(feel)"、 "*深感*(clearly feel)", and "*意味着*(mean)" are not sentiment words while they are usually collocated with sentiment words. But according to PMI and LR their sentiment strength value is very big, even applying TF-IDF cannot exclude them. Therefore, finding other calculation methods is our future work.

3) Context relevance of sentiment words

Polysemy is rather ordinary in Chinese. For instance, *"发作*(break out)*"* means either disease breaks out or temper or anger flares up. These words bring huge difficulty to computer processing. Besides, negation is another challenge in our work.

6. **Conclusion.** Complex sentiment analysis is an important topic to natural language processing area. This paper discusses on sentiment categorization models and selects the Plutchik's model for computing purpose. To construct sentiment lexicon, an unsupervised sentiment words extraction method is proposed, which makes use a small number of seed words and expands the lexicon using statistical approaches. Experimental results show that 1) PMI is superior to LR in sentiment similarity calculating, 2) higher accuracy is achieved with more seeds and smaller co-occurrence range. The best accuracy in this experiment is about 50%, which is promising to reduce manpower in sentiment lexicon construction.

Future work is planned as follows. First, we will try to refine part-of-speech patterns. Second, semantic information will be incorporated into the method. Third, more context information will be used in appropriate manners. The ultimate goal of this work is the complex sentiment analysis tool that is able to understand complex sentiment carried by a given piece of text. With the sentiment lexicon, research will be carried out to achieve this goal.

Acknowledgements. This work is partially supported by NSFC (60703051) and MOST of China (2009DFA12970). We sincerely thank the reviewers for the valuable comments.

REFERENCES

[1] Robert Plutchik, *Emotions and Life: perspectives from psychology, biology, and evolution*, American Psychological Association, 2003.

[2] Turney Peter, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [C], *Proc. of the 40th Annual Meeting of the Association for Computation Linguistics* pp. 417-424, 2002.

[3] Linhong Xu, Hongfei Lin and Zhihao Yang, Text Orientation Identification Based on Semantic Comprehension, *Journal of Chinese Information Processing*, vol.21, no.1, pp.96-100, 2007.

[4] Xiaoying Xu and Jianhua Tao, Research on sentiment definition in China sentiment systems, *Proc. of The 1st Conference on Effect Computing and Intelligent Interaction*, pp.199-205, 2003.

[5] Jianmei Chen, *Construction of Chinese Sentiment Ontology and Applications*, Master Thesis, Dalian Institute of Technology, 2008.

[6] Yunqing Xia, Linlin Wang and Kam-Fai Wong, Sentiment Vector Space Model for Lyric-based Song Sentiment Classification, *International Journal of Computer Processing of Oriental Languages*, vol. 21, no. 4, pp. 331-345, 2008.

[7] K. T. Strongman, The psychology of emotion, *theories of emotion in perspective*, 4th ed. Chichester, Wiley, 1996.

[8] Janyce Wiebe, Rebecca Bruce and Thomas O'Hara, Development and use of a gold standard dataset for subjectivity classifications [A], *Proceedings of the 37th Annual Meeting of the Association for Computation Linguistics*, pp. 246-253,1999.

[9] Ziqiong Zhang, *Towards Word Combination Patterns in Chinese Sentiment Analysis*, Master Thesis, Harbin Institute of Technology, 2007.

[10] Jianmei Chen, Hongfei Lin and Zhihao Yang, Automatic Acquisition of Emotional Vocabulary based on Syntax, *CAAI Transactions on Intelligent Systems*, pp.100-106, 2009.

[11] Yanlan, Zhu, Jin Min, Yaqian Zhao, Xuanjing Huang, and Lide Wu, Semantic Orientation based on HowNet, *Journal of Chinese Information Processing*, vol.20, no. 1, pp. 14-20, 2005.

[12] Linhong Xu, *The Text Affective Computing Based on Semantic Resource*, Master Thesis, Dalian Institute of Technology, 2007.

[13] Yuning Dong, *Linguistic Research on News Review*, Fudan University, 2007.