International Journal of Knowledge and Language Processing
Volume 5, Number 2, **2014**

www.ijklp.org
KLP International ©2014 ISSN 2191-2734
pp.20–31

An Approach to Sentiment Analysis for Chinese News Text Based on Topic Sentences Extraction

You Jianqing[1], ZhangYangsen[2] and Tong Yixuan[3]

[1] Institute of Intelligence Information Processing
Beijing Information Science and Technology University,
Computer Center
Beijing Information Science and Technology University
No. 12, East Qinghe-Xiaoying Road, Haidian District, Beijing, China
yjq@bistu.edu.cn

[2] Institute of Intelligence Information Processing
Beijing Information Science and Technology University, Beijing, China
zys@bistu.edu.cn

[3] Institute of Intelligence Information Processing
Beijing Information Science and Technology University, Beijing, China
tongyixuanjames@163.com

Selected Paper from Chinese Lexical Semantic Workshop 2014

ABSTRACT. *Text sentiment analysis has been a hotspot in Natural Language Processing area. Now many researchers are focusing on the goal about news text sentiment analysis. But how to analyze efficiently news text is still a hard task for being lacking in general discourse-analysis theories and universal process techniques. In this paper, we propose an approach to sentiment analysis for Chinese news text based on topic sentences extraction. Utilizing the structure peculiarity of Chinese news text, we extract three topic sentences per text based on the following four features: high-frequency word, news title, sentence location and tendentious-cue word which often conveys certain sentiment of the text author. And then according to two expression phenomena in Chinese news writing, we obtain news text sentiment by analyzing each extracted sentence sentiment and calculating their sentiment summation in same text. Finally, the method of this paper is shown about its feasibility and validity by experiments.*

Key words: Topic Sentences Extraction; Features, Chinese News Text; Sentiment Analysis

1. **Introduction.** With the development of information technology, kinds of online media are booming. Meanwhile those large portal sites and mainstream news organizations have been combing with each other so that to release and report news as soon as they could, which greatly facilitate the public to learn the latest information. But the more real-time contents and events are reported, the more details are exposed and the more texts are created. Thus, the current of latent public opinions changes faster and faster than ever.

How to utilize effectively news texts and mine their potential values, is attracting researchers' attention. Regarding it as our motivation and goal, we attempt to analyze the underlying sentiment of news texts. We extract three topic sentences per text and analyze their sentiment, then treat the summation sentiment as the text sentiment. In the rest parts of this paper, section 2 is the related work, section 3 presents topic sentences extracted model, section 4 introduces news text sentiment analysis, section 5 shows our experiments and discussions, section 6 concludes the paper and outlines the future work.

2. **Related Work.**

2.1. **Task Definition.** Although the principle "Objectivity and Impartiality" is generally accepted in news reported area, there still are kinds of deflections in news texts. In fact, even facing the same news/event, each reporter might hold different viewpoints because of various factors, such as his certain position (especially political standing), values, interests, hobbies, etc.[1]. Thus during the process of text writing, different reporter would express his mind with different scope and outlook. As a result, the news text would convey more or less its author's attitude, which is the literal sentiment of news text. How to analyze and achieve the sentiment is a part of our task.

Furthermore, some news texts only described certain objective facts which had happened, but we can discriminate the good facts from the bad ones. All these distinctions form the sentiment of news facts[2]. For example, A web News "The collapse of adobe buildings is the main reason for the casualties in earthquake"[1] (the details are omitted here) showed the earthquake had brought disaster such as collapse and casualties. Though it was an objective description about the earthquake loss, we regard it as a piece of negative fact news. In fact, as emotional creatures, people sometimes are hard to differentiate their subjective feelings (sentiment) from objective information (facts). That is, the sentiment of news facts is an ineluctability question for either sentiment analysis or sentiment annotation, so it is another part of our task.

Though there are disputes about the definition for news text sentiment analysis, the preceding two angles constitute our goal in this paper.

2.2. **Sentiment Analysis for News Text.** In Natural Language Processing (NLP) field, text sentiment analysis is also known as text opinion analysis or text polarity analysis and so on. Its prime work is to find and mine the text author's certain subjective information such as opinion, sentiment and attitude by NLP methods (here mainly Machine Learning

[1] http://news.ifeng.com/gundong/detail_2013_07/24/27863103_0.shtml.

techniques) [3]. The subjective information is generally classified into two polarity groups, Positive and Negative. Furthermore, in some researches, Objective (also called as none polarity) is treated as the third group.

Now, there are two main directions about text sentiment analysis researches. One is that the text type is expanding from comment text to micro-blog text, news text etc., and the core level is moving from word level to sentence level and discourse level. The other is that more detailed sentiment is distinguished from text, that is, a finer-grained analysis. For example, in the 5th Chinese Opinion Analysis Evaluation (COAE2014)[2], task 1 was a discourse level goal and it aimed at extracting and analyzing the sentiment orientation sentences of News texts. While in the evaluation for the 2nd Conference on Natural Language Processing & Chinese Computing (NLP&CC, 2013)[3], the participants of task 2 were asked to recognize the latent emotion for each piece of Micro-blog text, which could be divided into anger, disgust, fear, happiness, like, sadness, surprise and objective (none attitude).

Although theories for discourse analysis have been proposed and discussed more than ten years, they are still in the exploratory stage. The common but most practical and popular resolution is to reduce discourse dimensionality to sentence dimensionality, even word dimensionality by kinds of text process techniques. Wang Fei et al.[4] presented that the sentiment was closely linked with the discourse structure so that some structure information, especially the inter-sentential relationship, may be used to improve the performance of sentiment analysis. Following this thought, they utilized some explicit associations (conjunctions) to forecast the relationships between sentences and then they discussed how these relationships influence the discourse sentiment. Zuo Weisong[5] presented that it was feasible to resolve the discourse sentiment question with discourse-to-sentence means. In the work, he adjusted all the sentences' weights which were depended largely on the location and the opinion target(s) of sentences, finally he calculated all sentences' sentiment according to each sentence's weight and regarded the result as the discourse sentiment.

In some researches, news text sentiment analysis also follows the above dimensionality reduction idea. When analyzing Web news text sentiment, Shen Xiaoye et al.[6] chose the top sentiment-intensity sentence as the key sentiment-sentence for each paragraph, and then scored for the key sentence's weight according to its sentence location and paragraph location in the whole text, thus he obtained the news text sentiment by all key sentences and their scores. As for news topics, Tao Fumin et al.[7]selected sentiment features which were based on certain news topics and applied these features into the corresponding topic sentiment analysis. In his experiments, the features-selected method could enhance the analysis results about news comments text, especially those comparatively decentralized topics. In the opinion summarization about "Animal Clone", Lun Weiku et al.[8] proposed that an opinion sentence was the smallest semantic unit which could be extracted and sentiment words should be considered as cues when extracting opinion sentences and

[2] http://www.liip.cn/CCIR2014/pc.html.
[3] http://tcci.ccf.org.cn/conference/2013/index.html.

determining their polarities. Thus, they designed algorithms to detect sentiment words and extract those sentences which contained sentiment words, and then identified the polarities of sentences and finally finished documents summarizations.

3. **Model for Topic Sentences Extraction.** As previously mentioned, we discuss how to conduct topic sentences extraction model according to Chinese news text structure peculiarity and obtain the text sentiment by analyzing the extracted sentences sentiment. In this section, using the corpus of COAE2014 task 1, we introduce the method to create topic sentences extraction model and the following four features are taken into account.

3.1. **Feature for High-frequency Word.** In the field of traditional automatic abstraction, most of the related researches are managed to avoid all kinds of redundant and duplicate information which are chiefly due to High-frequency Word (HFW). But when analyzing news text sentiment, we present that the duplicate information plays the role of emphasis on news topic and text sentiment. Thus we need to focus on the questions that which HFW would be choose and what degree a certain HFW could make contribution to news topic sentences extraction.

After the preprocessing steps for all the corpus texts, such as word segmentation (the NLPIR/ICTCLAS APIs[4] provided by Dr. Zhang Huaping), stop-word removal and word-frequency statistics etc., we obtain the HFWs (marked every HFW as HFW Ⅰ and the corresponding vocabulary set as HFW Set Ⅰ). In fact, we are more eager to find those HFWs (marked every HFW as HFW Ⅱ and the corresponding vocabulary set as HFW Set Ⅱ) which are better at highlighting news content(s) than HFW Ⅰ. Usually, HFW Ⅱ can't be found from the common dictionary(ies), but are new words or new phrases which are created/combined/refined by text author(s) according to the target which would be reported, thus these HFW Ⅱ might be used frequently in the texts.

In our experiments, we find that each HFW Ⅱ is, in general, consist of a certain number of HFW Ⅰ (often 2-5) which occur consecutively, so we can obtain HFW Set Ⅱ from HFW Set Ⅰ. But at the same time, the size of HFW Set Ⅱ can influence the size of HFW Set Ⅰ in reverse. Thus after obtaining HFW Set Ⅰ, we must manage to get HFW Set Ⅱ and update HFW Set Ⅰ, and then merge two vocabulary sets into one as the final high-frequency vocabulary. The algorithm for the generation of high-frequency vocabulary is shown as follow.

[Algorithm] The generation of high-frequency vocabulary

Input: the whole news texts corpus

Output: the high-frequency vocabulary

Step 1: the preprocess stage for all the news text in corpus. The mainly work is word segmentation and stop-word removal.

[4] http://ictclas.nlpir.org/.

Step 2: the HFW Set Ⅰobtaining stage. After word-frequency statistics, we choose any word whose frequency (marked as tf_I) is greater than *2* as HFW Ⅰcandidate, so we can obtain HFW Set Ⅰby traversing the corpus.

Step 3: the HFW Set Ⅱ obtaining stage.

for each sentence of every text in the corpus

for each HFW Ⅰin sentence

try to combine the HFW Ⅰ with its following word(s) (1 to 4 word) into a new word and count its frequency, if the frequency is above threshold (it is closely related to the number of the sentences in the text and the details are shown in Table 1), mark the new word as HFW Ⅱ and its frequency as tf_{II}, then put the new word into HFW Set Ⅱ.

Step 4: the update stage for HFW Set Ⅰ.

for each HFW Ⅱ

for each HFW Ⅰwhich constitutes of HFW Ⅱ

if ($tf_{II} < tf_I$) then update tf_I, New $tf_I = tf_I - tf_{II}$.

else remove the HFW Ⅰfrom HFW Set Ⅰ.

Step 5: the merger stage for HFW Set Ⅰand HFW Set Ⅱ. We finally select 20 high frequency words and the selection principle is that HFW Set Ⅱ is priority and HFW Set Ⅰ is secondary only when the size of HFW Set Ⅱ is less than 20, according to the descending order of word-frequency.

TABLE 1. THRESHOLD FOR TF_{II} WHEN OBTAINING HFW Ⅱ

Number of sentences in text	＜10	10~29	30~69	≥70
Threshold	2	3	4	5

After HFWs selection, we need to assign proper weight for each word so that we can estimate its contribution to the sentence. In news topics expressive function, we assume that the HFWs appeared in news title are more valuable than those appeared in the beginning and end paragraph, also the HFWs appeared in beginning and end paragraph are more valuable than those appeared in other location. Being spired by the hypothesis [9] and considering that there isn't any paragraph information but only sentence ID in the corpus of COAE2014 task 1, we assume that the first or last *l* sentences should constitute the beginning or end paragraph and the HFWs weight setting is shown in Table 2. Here, the range of the parameter *l* is same as the threshold in table 1.

TABLE 2. WEIGHT FOR HIGH FREQUENCY WORD

Position of high frequency word	Appeared in news title	Appeared in *l* sentence in beginning or end	Appeared in other position
Weight	5	3	1

Now, we can extract topic sentences according to high-frequency words and their contributions by the equation (1).

$$f_1(s_i) = \frac{\sum_1^{n_i} \left(word_{ij} \times weight_{ij} \right)}{\sum_1^m \sum_1^{n_i} \left(word_{ij} \times weight_{ij} \right)} \qquad (1)$$

Where s_i is the *i* sentence in the text, n_i is the total HFW of the sentence s_i contains, $word_{ij}$ is the *j* high-frequency word in the sentence s_i, $weight_{ij}$ is the word weight, *m* is the total sentences in the news text.

3.2. **Feature for News Title.** News title is the soul for its refinement. Also it is the most important and the most expressive part in the text. Thus news title would greatly influence the effect of topic sentences extraction. In this paper, we regard news title as a sentence so that we can calculate the similarities between the title and each sentence to evaluate their relationship. In other words, if the sentence similarity between the title and certain sentence is higher, the relationship between the sentence and the text is closer and the sentence may express more information about the news topic, and vice versa.

After a great deal of analysis for news texts, we find that those nouns play a finishing touch on the role for news and the other POS (part of speech) words detail the content. Therefore, the nouns similarity and other POS words similarity are separately taken into account in the process for sentence similarity calculation[10], which is shown with the equation (2).

$$f_2(s_i) = \alpha sim_n(s_i, t) + \beta sim_o(s_i, t) \qquad (2)$$

Where s_i is the *i* sentence in the text, $sim_n(s_i, t)$ is the nouns similarity between sentence s_i and news title, $sim_o(s_i, t)$ is the other POS words similarity between sentence s_i and the title, α and β is the adjustment coefficients which meet the equation $\alpha + \beta = 1$. In particular, $sim_n(s_i, t)$ and $sim_o(s_i, t)$ are shown with the equation (3) and (4).

$$sim_n(s_i, t) = \frac{Nouns_{same}}{Noun_{s_i} + Nouns_t - Nouns_{same}} \qquad (3)$$

Where $Noun_{s_i}$ is the number of nouns in the sentence s_i, $Nouns_t$ is the number of nouns in news title, $Nouns_{same}$ is the number of nouns which are appeared in sentence s_i and news title simultaneously. As a note, we regard all the HFW Set Ⅱ (the details are shown in section 3.1) as nouns.

$$sim_o(s_i, t) = \frac{\sum o_{1i} \times o_{2i}}{\sqrt{o_{1i}^2} \times \sqrt{o_{2i}^2}} \qquad (4)$$

Before calculating equation (4), we need to combine all the other POS words, which are appeared in the sentence s_i and news title, into a collection so that we can get two vectors

$V_s(o_{11}, o_{12}, ..., o_{1n})$ and $V_t(o_{21}, o_{22}, ..., o_{2n})$ which presents respectively the vector for other POS words of the sentence s_i and news title. Thus the cosine similarity could be used to obtain the similarity of other POS words between the sentence s_i and news title according to equation (4).

Of course, if there are negative words appeared in sentence and/or News title, the negative effect for sentence similarity should be considered. In this paper, we calculate the absolute difference by the subtraction for the two numbers of negative words which are appeared respectively in sentence and news title. Then according to the absolute difference and the result of equation (4) with the principle "the similarity is 0 (zero) if the difference is odd" and "the similarity remains if the difference is even", we can get the other POS words similarity between every sentence and news title.

3.3. **Feature for Sentence Location.** In most news texts, the guide-reading paragraph, the beginning paragraph and the end paragraph may contain some conclusive sentences which help readers gain outstanding information about the reported news. To some extent, the conclusive sentences can be regarded as topic sentences and help the work for topic sentences extraction. But in the corpus of COAE2014 task 1, there is no any paragraph information which could be used to estimate those conclusive sentences. We assume that all those sentences, the first or last *l* sentences in text, might be conclusive sentences and could be used for topic sentences extraction.

Furthermore, as we all know, every sentence with difference location might make different contribution to information expression. Besides, in News domain there is a special structure, the so-called "inverted pyramid" structure, which is distinctive from other types of text. More specifically, in News text structure, the more important part/sentence is nearer the text beginning location than the less important part/sentence. So we can use certain decreasing function to evaluate the location information for news text structure.

Thus, supposing that the total number of news sentences is *m*, we can differentiate the sentence importance according to its location[11] by equation (5). The threshold for parameter *l* is same as the threshold in table 1.

$$f_3(s_i) = \begin{cases} 1, & i \le l \text{ or } i \ge m\text{-}l \\ 0, & i \in other\ location \end{cases} \quad (5)$$

3.4. **Feature for Tendentious-cue Word.** Language expression custom, writing standardization and content intelligibility lead to a universal phenomenon that many news texts often end with recapitulative or summative sentence, named as cue-sentence in this paper. The cue-sentence can be divided into two categories.

One is that there is a common conclusive word in news cue-sentence, for examples, *"总之"* (in short, in a word, all in all etc.) and *"综上"*(in conclusion, above all etc.). Such kind of sentence is named as conclusive-sentence in this paper and it also appears widely in other fields texts. But it doesn't help the work for sentiment analysis.

The other is that in some news cue-sentences there is certain tendentious-cue word, for

examples, *"试问"*(it could be asked that) and *"可以预见"*(it could be foreseen that). The sentence which contains tendentious-cue word often conveys some sentiment of the text's author and it has an outstanding role in news text sentiment analysis, named as tendentious-sentence in this paper. We regard the tendentious-sentence as one kind of topic sentences and extract it by tendentious-cue word matching.

We collect 22 tendentious-cue words which appear frequently in news texts and apply them into the work of topic sentences extraction according to the equation (6).

$$f_4(s_i) = \begin{cases} 1, & s_i \ contains \ certain \ tendentious\text{-}cue \ word \\ 0, & s_i \ don't \ contains \ certain \ tendentious\text{-}cue \ word \end{cases} \quad (6)$$

3.5. **Model for Topic Sentences Extraction Based on Multi-feature.** When finishing features selection, we apply the preceding four features to conducting topic sentences extraction model, shown with the equation (7).

$$f_{fin}(s_i) = \sum_k w_k \times f_k(s_i) \quad (7)$$

Where the parameter k is 1 to 4, $f_k(s_i)$ stand the four selected features, w_k stand their corresponding weights.

Before our experiments, we have postulated that topic sentences extraction would mainly depend on the repeated information and the key information thus these two kinds of information should be given higher weights than other factors. While in news texts, the repeated information corresponds to the feature for high-frequency word and the key information corresponds to the feature for news title, so the weights of the two features might higher than other features. In fact, this hypothesis is demonstrated by the experiment result that the weights (w_k) of the four features are respectively 0.4, 0.4, 0.1 and 0.1.

4. **News Text Sentiment Analysis Based on Topic Sentences.** When analyzing Chinese news text sentiment, we apply two obvious expression phenomena in Chinese news expression which would correspond to two procedures in the analysis process.

The first is that it is very strict and pretty standard in the diction procedure of Chinese news writing. Especially, The usages *"褒词贬用"*(a commendatory word is used as a derogatory one) and *"贬词包用"*(a derogatory word is used as a commendatory one) are extremely rare, even forbidden. Thus the expression regulation could be used to evaluate the sentence sentiment by calculating the number of sentiment words in sentence. More specially, if the number of positive words is more than the number of negative words in a sentence, it would be regarded as a positive sentence, and vice versa.

The second is that when sentence construction it is consistent about the expression logic, intonation and tone in whole context. The consistency makes it possible to determine the news text sentiment by figuring the number of positive sentences and the number of negative sentences. More specially, if the number of positive sentences is more than the number of negative sentences, the news text would be regarded as a positive text, and vice versa.

Thus we construct a two-part sentiment dictionary for news text sentiment analysis. The first part is mainly used for the sentiment of news facts (introduced in section 2.1) and it is made up of those sentiment words based on news domain, named as domain sentiment dictionary. There are total 71 positive words, such as *"捐资"*(donations) and *"正能量"* (the positive energy) etc., and 214 negative word, such as *"垮塌"*(collapse) and *"伤亡"*(casualties) etc. The second part is mainly used for the sentiment of News text and it is made up of those common sentiment words, named as common sentiment dictionary. There are total 21192 words which are mainly from *"情感词汇本体"*(Chinese Emotion Word Ontology)[12] by Dalian University of Technology.

5. **Experiments.** In the corpus of COAE 2014 task 1, each news text includes several parts, such as the http URL, news title, document ID, source, the reported time and detailed content etc. In particular, the content part is composed of sentences and their corresponding sentence IDs without any paragraph information.

We select 150 news texts as our training set. In this set, we annotate the sentiment for every text, the topic sentences and their sentiment polarities. Besides, we choose 150 texts as our test set and only annotate the sentiment for every text. Table 3 shows some detailed information about the two set.

TABLE 3. COMPARISON BETWEEN TRAINING SET AND TEST SET

Data set	Number of positive texts	Number of negative texts	Number of neutral texts	Total numbers of sentences
Training set	54	58	38	3335
Test set	59	56	35	3230

We have discussed that the factor of news facts sentiment would greatly influence the annotated results. And indeed, different people might have different views about the same news text, which leads to an abnormal phenomenon that the number of neutral texts is lower than positive texts and negative texts in table 3, but not the common sense that the number of neutral texts should higher than other two texts in news domain.

5.2. **Analysis for Topic Sentences Extraction.** According to topic sentences extraction model, we get the extraction sentences and sort them with descending order by their weights. In order to verify the feasibility and validity of the model, we choose the top 5 extracted topic sentences (respectively denoted with 1 to 5) per text in training set and compare them with the annotated topic sentence(s). The comparison regulation is that if any annotated sentence matches any sentence of the top 5 ones, the former would be labeled with the latter's order. The statistics for comparison result is shown in table 4. It should be specially explained that some news texts contain two topic sentences, thus there are totally 166 annotated sentences for 150 texts in table 4.

TABLE 4. COMPARISON BETWEEN ANNOTATED SENTENCES AND EXTRACTED SENTENCES

The labeled order that the annotated sentences matches the extracted ones	1	2	3	4	5	Not Top 5
Number of sentences	99	40	16	6	1	4
Percentage	59.64%	24.1%	9.64%	3.61%	0.6%	2.41%

From table 4, we can find that the annotated sentences are mostly the top 3 order of the extracted sentences rank. More specially, in the 166 annotated topic sentences, there are 155 sentences in the top 3 and the percentage is 93.38%. It shows that the conducted model for topic sentences extraction does well for our goal. Also, the underlying conclusion is that we might extract *three* topic sentences per text to do other related researches and we would apply the conclusion to news text sentiment analysis.

5.3. **Sentiment Analysis for Chinese News Text.** According to the implicit conclusion in section 5.2, we extract three sentences per text in test set by the topic sentences extraction model, totally 450 sentences. And then using the two expression phenomena in Chinese news writing and their corresponding procedures for sentiment analysis, which are mentioned in section 4, we get a group of analysis results (marked as Method 1). In order to compare our method, we treat each text as a unit by Unigram feature and use Naive Bayes Classifier (NBC) to train in the training set, and then test all the news texts in the test set, thus achieve another group of analysis results (marked as Method 2). Table 5 shows the comparison between two groups of results.

TABLE 5. COMPARISON BETWEEN TWO GROUPS OF RESULTS.

Analysis results	Total documents	Correctly classified documents	The correct percentage	Incorrectly classified documents	The incorrect percentage
Method 1	150	107	71.3%	43	28.7%
Method 2	150	94	62.7%	56	37.3%

Though our training set and test set is relatively small, our approach is better than NBC in news text sentiment analysis and the precision (the correct percentage in table 5) is obviously higher. We discuss three reasonable interpretations.

Firstly, from the perspective of Machine Learning, fine-grained means would perform better than coarse-grained means in classification experiments. In this paper, our approach begins with word level, sentence level and finally discourse level, while the NBC is based on discourse level. The different level determines the fine-grained analysis would bring better results, as expected.

Secondly, we collect 285 domain sentiment words to analyze the sentiment of news facts. That is, the construction of domain sentiment dictionary enhances the precision for news text sentiment analysis.

Thirdly, as document [2] had discussed, news text sentiment may also include the audiences' sentiment about the news text (mainly their attitudes, thoughts after reading the news). The audiences' sentiment has greatly influence our work, such as the construction of

domain sentiment dictionary, the sentiment annotation for training set and test set. And the performance of NBC method, to some extent, reduce the impact of human factors, also it lower the accuracy which is based on the manual annotation.

6. **Conclusion and Future Work.** In this paper, we discuss the task about Chinese news text sentiment analysis and the feasibility to achieve the goal by the work of sentences (such as topic sentences) sentiment analysis. Taking advantage of four features which include high-frequency word, news title, sentence location and tendentious-cue word, we conduct topic sentences extraction model and apply it to extract *three* topic sentences per text. Then we analyze each sentiment for the three extracted sentences in the same text and regard their summation as the text sentiment. We find that (1) in Chinese news domain, *three* sentences per text could be well performed for the research of topic extraction and its related work, (2) word-match pattern is still a practical method in some sentiment analysis researches, such as news sentiment analysis, (3) discourse level sentiment analysis can be resolved by dimensionality reduction according to text features.

Our future work would mainly focus on some improvements about topic sentences extraction model and sentiment analysis techniques for news texts. For example, our sentiment analysis only achieves an accuracy of 71.3% which is at the medium level in other sentiment analysis researches. So it is necessary to discuss further and optimize the model and analysis techniques by more angles and experiments. Besides, the rest sentences (not those topic sentences) sometimes affect the sentiment analysis result, how to deal with them by Machine Learning methods would be our next work, too.

Acknowledgment. Research of this paper is funded by the Natural Science Foundation of China (NSFC, Grant No.61070119, 61370139), the Project of Construction of Innovation Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality (Grant No.IDHT20130519), the Beijing Municipal Education Commission Special Fund (Grant No.PXM2013_014224_000042, PXM2014_014224_000067) and the Fund of Beijing Information Science and Technology University (Grant No. 5221410935)

REFERENCES

[1] Hao Huili, Discussions shallowly on the sentiment of News and the objective reports, *Press Circles*, vol. 6, 12-13, 2003. [In Chinese]

[2] Gu Runde, Discussions on three levels for News sentiment, *Journal of Nanjing University(Philosophy, Humanities and Social Sciences)* , vol.3, 183-187, 1999. [In Chinese]

[3] Li Xiaojun, Dai lin, Shi Hanxiao and Huang Qi, Survey on sentiment orientation analysis of texts, *Journal of Zhejiang University (Engineering Science)*, vol.45, no.7, 1167-1174, 2011. [In Chinese]

[4] Wang Fei, Wu Yunfang, Qiu Likun, Exploiting Discourse Relations for Sentiment Analysis, *Proceeding of COLING 2012: Posters*, 1311-1320, 2012.

[5] Zuo Weisong, Studies on sentiment analysis of chapter based on rule and statistics, Zhengzhou University, Zhengzhou, 2010. [In Chinese]

[6] Shen Xiaoye, Feng Huamin, Wu fei, Opinion orientation analysis framework for Web News based on the Sematic, *Journal of Zhengzhou University(Natural Science Edition)*, vol. 41, no.11, 33-35, 2009. [In Chinese]

[7] Tao Fumin, Gao Jun, Wang Tengjiao, Zhou Kai, Topic oriented sentimental feature selection method for News comments, *Journal of Chinese Information Processing* , vol. 24, no.3, 37-43, 2010. [In Chinese]

[8] Lun Weiku, Liang Yuting, Chen Hsinhsi, Opinion extraction, summarization and tracking in news and Blog corpora, *Proceeding of the AAAI 2006 Spring Symp. on Computational Approaches to Analyzing Weblogs*, Menlo Park: AAAI Press, 2006.

[9] Zhao Yuehua, Research on key words identification and communication trends of hot event at the forum, Jiangsu University, Nanjing, 2011. [In Chinese]

[10] He Wei, Wang Yu, Extracting topic sentence from Web text based on sentence relationship map, *New Technology of Library and Information Service*, vol.176, no. 3, 57-61, 2009. [In Chinese]

[11] Wang Wei, Zhao Dongyan, Zhao Wei, Identification of topic sentences about key event in Chinese News, *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 47, no.5, 789-796, 2011. [In Chinese]

[12] Xu Linhong, Lin Hongfei, Panyu Wang et al. Constructing the affective lexicon Ontology, *Journal of the China society for scientific and technical information*, vol. 27, no.2, 180-185, 2008. [In Chinese]