PKUTM Experiments in NTCIR-8 MOAT Task

Chenfeng Wang*, Tengfei Ma*, Liqiang Guo,
Xiaojun Wan and Jianwu Yang

Institute of Computer Science & Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistics (Peking University), MOE, China

{ wangchenfeng,matengfei,guoliqiang,wanxiaojun, yjw}@icst.pku.edu.cn

ABSTRACT

This paper describes our work in the Simplified Chinese opinion analysis tasks in NTCIR-8. In the task of detecting opinioned sentences, various sentiment lexicons are used, including opinion indicators, opinion operators, degree adverbs and opinion words. The linear SVM model is selected as the main classifier, and four groups of features are extracted according to punctuations, words and sentiment lexicons. We also try a two-step classification to improve the SVM result. For extracting the opinion holder and target, we use a synthesis of CRF and heuristic rules. The evaluation results on NTCIR-8 MOAT Simplified Chinese side show that our system achieves the best f-measure in two tasks. This demonstrates that the proposed framework is promising.

Keywords: *NTCIR, Sentiment Analysis, Subjectivity Classification, Opinion Holder/Target Extraction*

1. INTRODUCTION

Opinion mining, also called opinion analysis or sentiment analysis, has become a hot area of computational linguistic. It focuses on the analysis of subjectivity, sentiment and opinion extraction in text. With the web's fast development, opinion mining has more and more applications, such as identifying the web users' opinions toward products, people or events.

To date, a lot of work has been done for opinion mining. The task becomes more mature and several subtasks are derived. Subjectivity analysis, or opinionated identification, aims at automatically recognizing subjective content. Sentiment classification, or polarity classification, attempts to predict the orientation and strength of the sentiment towards the opinion target in the subjective content. Besides, opinion extraction tries to analyze further into sentence and extract opinion holders and opinion targets of the opinion expression.

Because many researchers perform their experiments on different datasets, a fair comparison are difficult to make. NTCIR Multilingual Opinion Analysis Tasks (MOAT) [8] proposes several tasks to evaluate and compare different methods for opinion mining.

In NTCIR-8, we participate in three subtasks for the Simplified Chinese side:

1. Opinionated subtask.
2. Opinion holder extraction.
3. Opinion target extraction.

We apply a machine learning method in opinionated sentences identification task. The corpora in NTCIR-6, NTCIR-7 and MPQA[1] are also used for training because the provided training data in NTCIR-8 is not plenty enough. In addition, we translate the Traditional Chinese corpus into Simplified Chinese and add them into the training set.

For opinion holder and target extraction, we use CRF [9] to combine nouns or pronouns into candidate phrases and then exploit a few heuristic syntactic rules to choose the best candidate.

Three runs are submitted to evaluate our classifier's effectiveness. The results show that our system achieves promising results on two subtasks: the highest precision and F-measure on the opinionated task and the highest F-measure on the holder extraction task among all submitted runs.

The rest of this paper is organized as follows. In Section 2 we briefly review the existing works on opinion mining. Our opinionated identification system is described in Section 3, and Section 4 describes the method in holder and target extraction. Section 5 gives the evaluation results and discussion. In the last section we give our conclusion.

2. Related Work

Researches on opinion mining range from word-level to sentence-level and document-level. In the early time, researches focus on words [1]. Then, sentences and documents' opinions are studied. Mainly two kinds of approaches are proposed to solve the problem. One is unsupervised[2][3], which uses linguistic knowledge on sentiment words and heuristic rules to predict the content's opinion. The other is supervised [4], which extracts features from labeled data and builds a classification or labeling model, and predicts the content's opinion. The most commonly used classification models include Naive Bayes[4], Maximum Entropy[10] and Support Vector Machine[11]. Nowadays, more and more methods combine the two approaches, by using the sentiment knowledge, heuristic rules as well as machine learning models [21].

Opinion holder/target identification is a more challenging problem in the field of opinion analysis. Bethard[14] uses the

*The authors contribute equally to this paper.

[1] Available at http://www.cs.pitt.edu/mpqa

technique of semantic parsing and syntactical features to extract propositions and holders. Kim and Hovy[13] design a system to automatically learn the syntactic features signaling opinion holders using a Maximum Entropy ranking algorithm trained on human annotated data. Other researchers[7][9] use CRF [16] with some extraction patterns.

3. Detecting Opinionated Sentences

In our system, we regard the sentence level subjective detection as a classification problem. First, data is processed by a POS (Part-Of-Speech) Tagger and a NER (Named Entities Recognizing) tool. Each sentence is represented by a vector of features extracted from the dataset. After that, a Linear-SVM Classifier is applied to get the basic classification result and finally an iterative classifier is used to improve the prediction.

3.1 Dataset and Preprocessing

NTCIR-6 and NTCIR-7's corpora are used in our system. The MPQA and our in-house labeled corpora are also used in lexicon building and feature selection parts of our system. As we only experiment in Simplified Chinese, all the Traditional Chinese corpora are translated into Simplified Chinese using a translation tool, ConvertZ[2]. While, the MPQA corpus is first processed and extracted in English, then the words are translated by Google Translation Tool[3].

Similar to most existing opinion mining tools, we first preprocess the dataset before extracting the opinions of sentences. We use our own word segmentation tool to get each Chinese word as well as its Part-Of-Speech tag. After that, we use our in-house NER Tool to recognize the named entities. Besides, we build some sentiment-related lexicons to identify the opinion features, which are listed below.

Opinion Operators

An opinion operator lexicon is first collected from the sentiment dictionary of *HowNet*[4] and labeled operators in NTCIR-6, then expanded using the *Synonymy Thesaurus* [19], and finally filtered using labeled datasets. Only distinguishable verbs which appear mostly in opinionated sentences are selected as good operators, e.g.表示(express), 声称(claim), 相信(believe), 盛赞(praise).

Opinion Indicators

An opinion indicator is a word indicating the orientation of an opinion or the orientation trend of multiple opinions [20]. We collect 17 opinion indicators manually, such as 但是(but), 并且 (however), 尽管(although).

Degree Adverbs

Degree adverbs, which frequently co-occur with opinion words, can strengthen or weaken the degree of sentiment or even reverse the polarity of the sentiment. Moreover, sometimes they can even activate a normal word to be sentimental. Some examples of the degree adverbs include 非常(very), 缺乏(lack of), 不(not), 尤其 (especially), etc.

[2] http://alf-li.tripod.com

[3] http://translate.google.com

[4] http://www.keenage.com

Opinion Words

Opinion word lexicon is used in most of sentiment analysis systems, and it plays a key role in opinionated sentences identification. Our initial opinion word lexicon is collected from three linguistic resources:

1. Our In-house Opinion Word List, which consists of 5975 words. Each word is marked with a positive or negative label and its opinion degree. The Word List is first built based on *HowNet*, then expanded by using *Synonymy Thesaurus* and the SVM classifier, and manually check at last.
2. The opinion word lexicon provided by National Taiwan University (NTU), which consists of 2,812 positive words and 8,276 negative words [17].
3. The opinion word lexicon provided by Jun Li [18], which includes 4468 negative words and 5567 positive words.

The final lexicon consists of 28421 opinion words, such as 欢喜 (happy), 当头棒喝 (a severe warning), 怪罪 (blame), 难堪 (intolerable). We only use them as subject words without considering whether they are positive or negative.

Strong Opinion Words

Only 4680 words' probabilities of appearing in opinionated sentences are no less than 50% in all the corpora. Some words like 大气(lordly or air), 推进(boost or push forward) might have different meanings and sentiment in Chinese. So we decided to build a strong opinion word lexicon. We assume that all the idioms in the opinion word lexicon are more likely to be opinion words and added them to the strong opinion lexicon. Besides, we also add the word that appeared in opinionated sentences more often than in objective sentences. Finally, we obtain a strong opinion word lexicon with 6471 words.

3.2 Feature Selection

Based on experiments on NTCIR-7's corpus, we select four groups of features including punctuations features, words and entities features, sentiment features and collocation features. Some of these features are borrowed from Ruifeng Xu's work [5]. For each given sentence, we extract the following features and add them to the vector space model.

Table 1. Features used in the opinionated subtask

Punctuations Features
Presence of quotation marks like “,「,', 」and ”
Presence of colon followed by quotation marks
Percentage of punctuations in sentences
Words and Entities Features
The percentage of numeral words
The presence of pronoun
The presence of a named entity
The presence of a word which indicates a sequence
Lexical Subjective Clues
The presence of opinion operator
The presence of opinion indicator
The logarithm of percentage of opinion words
The logarithm of percentage of strong opinion words
The presence of degree verb

Collocation Features
The presence of collocations between named entities and opinion operators
The presence of collocations between pronouns or nouns and opinion operators
The presence of collocations between opinion operators and opinion words
The presence of collocations between pronouns and opinion words
The presence of collocations between nouns or pronouns and opinion words
The presence of collocations between degree adverbs and opinion operators
The presence of collocations between degree adverbs and opinion words
The presence of collocations between nouns or named entities and opinion words

With all these features our system achieves the F-measure value of 0.702 under lenient evaluation and 0.743 under strict evaluation on NTCIR-7's Simplified Chinese test corpus. Sentiment features and collocation features contribute the most, and the other two groups also make a slight improvement on some corpora. The features different from other works are explained below.

3.2.1 Punctuations Features

Punctuation features contain some special punctuation in sentences, which may be good indicators for opinionated or un-opinionated sentences. After browsing a lot of corpora and doing experiments we find the following three useful features in opinionated judgment.

Colon followed by Quotation Marks (CQM): Although there is already a quotation-related feature, we find that a sentence is usually someone's words or some decelerations when it has a quotation strictly following a colon.

Punctuation Content (PuC): The percent of the punctuations appearing in the sentence. The intuition behind this feature is that we find that a sentence which has a low percentage of punctuations tends to be objective.

3.2.2 Words and Entities Features

In the group of words and entities features, we add two features which might be good indicators for objective sentence.

Percentage of numeral words (PeNW): Ratio between the number of punctuations and the number of all words and punctuations. Numeral words often appear as data in objective sentences.

Presence of sequence indicator (PSI): The presence of a word which can indicate the sentence to be a part of a sequence or list, such as 首先(first of all), 第一(first), 其次(then), 一(1st, 1).

3.2.3 Lexical Subjective Clues

This group of features is commonly used in most machine learning based opinionated identification methods. In our system, the opinion operator feature (OOF) and opinion indicator feature (OIF) use the presence of the word; the opinion words feature (OWF) and strong opinion words feature (SOWF) use the logarithm of the percentage of the given words in sentence instead. Tests on NTCIR-7 show that the logarithm of the percentage performs better than term frequency or presence, which is also proved on our self-labeled dataset.

3.2.4 Collocation Features

Opinion is often expressed by several words together instead of one word. By referring Rui's paper, we make some changes on careful tests and finally determine to use the eight features listed in table 1:

Combining the pronouns and nouns: The original features are collocation between some words with pronouns or collocation with some word with nouns respectively**.** We try to combine the features and change them into collocation of some word with pronouns or nouns. Here the nouns include Named Entities. It can increase the feature's coverage with little lost in precision. After a series of test, we find that the combined feature, collocations between pronouns or nouns and opinion operators (PrPNOO) contribute more than only use collocations between pronouns and opinion operators and collocation between nouns and opinion operators.

The presence of collocations between opinion operators and opinion words (PrOpOw): This is an enhancement feature for operator, and it would be more accurate.

3.3 Classification Model*Basic Classifier*

We try several commonly used classifiers such as SVM, Naive Bayes, Max Entropy and Decision Tree. They are all trained on a combined dataset including NTCIR-6's Traditional Chinese corpus, NTCIR-7's Traditional Chinese corpus and the training set of NTCIR-7's Simplified Chinese corpus. We use three metrics for evaluation: Precision (P), Recall(R) and F-measure (F). The results are shown in Figures 1 and 2.

SVM-linear uses the linear kernel of SVM and SVM-rbf uses the Radial Basis Functions kernel of SVM. ME means Max Entropy model. J48 is the Decision Tree Model using C4.5, and NB is Naive Bayes. The linear SVM exceeds all the other models under both lenient evaluation and strict evaluation. Thus we choose the linear SVM as our basic classifier.

3.3.2 Iterative Classifier

Besides using the basic classifier, we also try some methods to improve the basic classifier's result. An iterative classifier is adopted according to Xu's paper[6], and we only use the opinionated result other than polarities and relevance.

4. Extracting opinion holders and targets

Some former researchers consider the holder/target extraction as a sequence tagging problem [7][9], and directly use CRF to address the task based on some useful features, such as position, POS, and dependency tree. However, limited by the size of data sets and the scope that features can represent, labeling the holder always results in a low recall. In this study, we use CRF to extract candidate phrases, and then choose the best candidate as the final holder/target with a few heuristic rules.

Figure 1.Comparison of different models under lenient evaluation

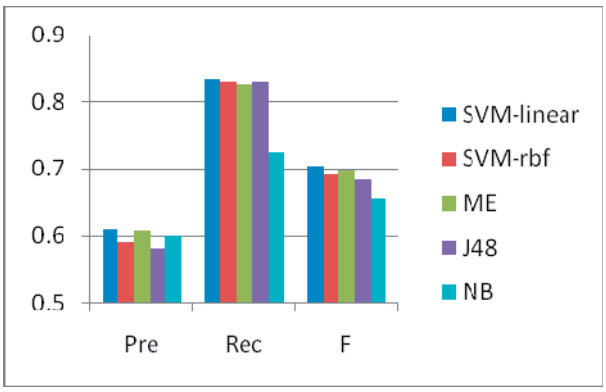
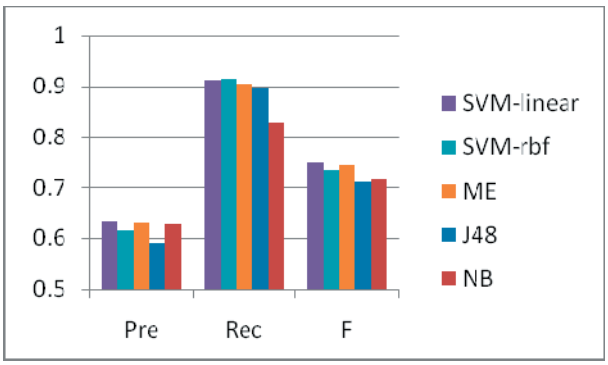


Figure 2.Comparison of different models under strict evaluation



4.1 Generating Semantic Phrases

Given a sequence $x_1x_2\ldots x_n$, we generate a corresponding sequence of tags $y_1y_2\ldots y_n$. The possible tag values are chosen from the set $\{B, I, N\}$, where B indicates the beginning of a phrase, I is the non-initial token of the phrase, and N represents that the word is not in any phrases.

We use the "Chinese Proposition Bank 2.0"[5] as the original training dataset, which is a corpus of Chinese text annotated with basic semantic propositions. Due to the differences between our task and the task of semantic proposition identification, the labeled information should be adapted. We only select the noun phrases and pronouns as the labeled phrases, because they are likely to be the holder or target. The features for CRF mainly include: word tokens, POS tokens, contextual features between words and between POSs.

After training on the modified Chinese proposition bank, we randomly choose a few samples from the file of "NTCIR-8MOATSample_SC" and test the result to find possible wrong patterns. Then the tested sentences are corrected and added into the training set.

5 http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T07

4.2 Finding Opinion Holders and Targets by Heuristic Rules

Opinion holders are usually the subjects of the sentences and they co-occurred with the opinion operators. Sometimes, however, the opinion holder may not exist in the operating sentence but is inherited from the neighboring sentences. Based on the observation, we design some heuristic rules for identifying opinion holders which are shown as follows.

1. The holder is before an opinion operator (include a colon) or following a quotes.
2. If a candidate phrase is governed by a preposition, it is not the opinion holder.
3. If a sentence is in a quotation or contains only part of a quotation, the opinion holder must be before or behind the quotation (in previous or posterior sentences).
4. If no candidate phrase conforms to the former conditions, try to use nouns or pronouns as candidates.
5. The opinion holder of subjective sentences, which have no opinion operators or extract no candidate as a holder, is regarded as "作者"(the author).

Opinion target identification is similar to the opinion holder recognition. An opinion target usually is the object of an operator or in the clause of opinion expression, but when the holder is in other sentences, it is the subject of the operating sentence. Due to the property of our training set, the extracted phrases are mainly subjects or objects of a verb. This further decreases the complexity of filtering rules.

5. Experimental Results and Discussion

The NTCIR-8's MOAT Simplified Chinese test corpus consists of 19 topics, 385 documents, including 4492 sentences and 4512 sentiment sub-sentences. There are two annotators and only the sentences labeled as opinionated by both annotators are considered as opinionated. 18.99% of the sentences are opinionated, and the others are objective.

5.1 Opinionated Sentence Recognition

Three runs trained on different datasets were submitted. In run 1, we used all corpora as training set, including NTCIR-6's corpus, NTCIR-7's Simplified Chinese and Traditional Chinese corpus and NTCIR-8's Simplified Chinese training set. We applied both the basic classifier and the iterative classifier for this run. The NTCIR-7's Simplified Chinese corpus and NTCIR-8's Simplified Chinese training set were used in run 2. Only the basic classifier was applied. The NTCIR-7's Traditional Chinese corpus was added to run2's training set in run 3.

The results are measured by Precision(P), Recall(R) and F-measure(F). All evaluation results for identifying opinionated sentences are given in Table 2.

Table 2. The result for identifying opinionated sentences

	Precision	Recall	F-measure
Run1	0.3721	0.8370	0.5152
Run2	**0.4134**	0.8335	**0.5527**
Run3	0.3405	**0.9062**	0.4950

Compared with other participants, our system achieves the highest F-measure in run 2, which also achieves the best precision with a little lose in Recall.

5.2 Opinion Holder and Target Identification

In the opinion holder task, we achieves the best f-measure both for opinionated sentences and for all sentences among the four participating groups and eight submitted results. Our target identification also gains a good performance. This demonstrates the effectiveness of our system. The results of the two tasks are showed in Table 3 and Table 4 respectively, where the three runs are based on the opinionated results of the above three runs in Section 5.1, respectively.

Table 3. Evaluations Results for Opinion Holders

		Precision	Recall	F-measure
Only for opinionated sentences	Run1	0.892	0.736	0.806
	Run2	**0.896**	0.732	0.805
	Run3	0.877	**0.792**	**0.832**
For all sentences	Run1	0.339	0.736	0.464
	Run2	**0.385**	0.732	**0.504**
	Run3	0.307	**0.792**	0.442

Table 4. Evaluations Results for Opinion Targets

		Precision	Recall	F-measure
Only for opinionated sentences	Run1	0.550	0.434	0.485
	Run2	**0.554**	0.431	0.485
	Run3	0.548	**0.473**	**0.508**
For all sentences	Run1	0.204	0.434	0.277
	Run2	**0.232**	0.431	**0.301**
	Run3	0.186	**0.473**	0.267

5.3 Discussion and Data analysis

Compared with other teams, our system achieves the best in F-measure and precision though the recall is not high enough in the opinionated sentence identification task. Looking through the results of the three runs, we find the run 2 outperforms the others and ranked the top in the list. It can be simply explained by the fact that NTCIR-7's SC corpus is more similar with NTCIR-8's than other corpora. Furthermore, it might be the differences in style between SC and TC when expressed. It is not always true that the larger the training set is, the better the model would be. For example the basic classifier's result on Run 1's train set is worse than on Run 2's train set, thought the train set of Run 1 is larger than that of Run 2. We also perform some experiments about the iterative classifier. Although the iterative classifier improves run 1' result, it does not make help on run 2's result, and the result even decreases a little. Run 3 performs the worst, but it is better than run 1 if run 1 just uses the basic classifier. This proves that the run 3's dataset is better than run 1 when testing on NTCIR-8.

However, in contrast with NTCIR-7 Simplified Chinese's results, in which the values of precision and f-measure from the best run achieved 0.5862 and 0.6839 under lenient evaluation, NTCIR-8's results are relatively low in both precision and F-measure. We explain the phenomenon as follows.

First of all, the test corpus in NTCIR8 may be harder to predict, and the differences in structure or semanteme between training and test corpora often leads to worse results.

Secondly, there may be differences in annotations. In NTCIR-8, there are only two annotators and the lenient evaluation considers all that sentence that be labeled as opinionated by both annotator as opinionated. The percentage of opinionated sentences is only 18.99% while it is 38.32% in NTCIR-7.

At last, feature distributions are different. Table 5 shows the coverage and precision of some well-performing features in NTCIR-7, including CQM, OOF and PrPNOO. We can notice that their precision declined a lot, which may be a reason why our system cannot get a high precision value.

Table 5. Features comparison between NTCIR-7 and NTCIR-8

	NTCIR-7		NTCIR-8	
	Coverage	Precision	Coverage	Precision
CQM	0.086	0.652	0.112	0.360
OOF	0.315	0.681	0.391	0.407
PrPNOO	0.303	0.694	0.383	0.414

6. Conclusion

In this paper, we propose a series of feature selection methods based on text information and opinion lexicon to solve the problem of opinionated identification. In opinion holder and target extraction, CRF is used to extract candidate phrases of nouns and pronouns. A few heuristic syntactic rules are adopted to choose the best candidate. Our proposed system achieves a best performance in the opinionated identification and holder extraction tasks. However, the results still have large space to be improved in further experiments by refining the model.

7. Acknowledgements

This work was supported by NSFC (60873155, 60875033), Beijing Nova Program (2008B03), NCET (NCET-08-0006), National High-tech R&D Program (2008AA01Z421) and National Development and Reform Commission High-tech Program of China (2008-2441).

REFERENCES

[1] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, 1997. ACL.

[2] V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrucken, Germany, 2000.

[3] Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 417–424.

[4] B. Pang, L. Lee, and S. Vaithyanathan, 2002.Thumbs up? Sentiment classification using Machine Learning Techniques. In *Proceedings of the EMNLP conference.*

[5] R.F.Xu,K.F. Wong and Y.Q.Xia, Coarse-Fine Opinion Mining – WIA in NTCIR-7 MOAT Task. In *NTCIR-7*,2008.

[6] R.F, Xu, K.F. Wong, Q. Lu and Y.Q. Xia, Learning Knowledge from Relevant Webpage for Opinion Analysis, In *Proc. IEEE/WIC/ACM WI-IAT*, 2008.

[7] Kang Liu, Jun zhao. NLPR at Multilingual Opinion Analysis Task in NTCIR. In *NTCIR-7, 2008.*

[8] Y. Seki, D.K.Evans, L.W. Ku, Overview of Multilingual Opinion Analysis Task at NTCIR-7, In *NTCIR-7*,2008

[9] Yejin Choi et al. Identifying Sources of Opinions with conditional Random Fields and Extraction Patterns. In *Proc of EMNLP 2005*.

[10] B. Pang and L.L. Lee, A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization based on Minimum Cuts. In *ACL04*, pp. 271-278, Spain, 2004

[11] Riloff, J. Wiebe and T. Wilson, Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *CoNLL03*, pp.25-32, 2003.

[12] Y. Seki, L.W. Ku, L.Sun,H.-H. Chen, N. Kando, Overview of Multilingual Opinion Analysis Task at NTCIR-8, In *NTCIR-8*,2010.

[13] Soo-Min Kim and Eduard Hovy. Identifying Opinion Holders for Question Answering in Opinion Texts. In *Proceedings AAAI-05 workshop on Question Answering in restricted domains*. 2005.

[14] S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou & D. Jurafsky. Automatic extraction of opinion propositions and their holders. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text.*2004.

[15] Yejin Choi et al.. Identifying Sources of Opinions with conditional Random Fields and Extraction Patterns. In *Proc of EMNLP 2005*.

[16] J. Lafferty, A. K. McCallum & F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of 18th International Conference on Machine Learning*. 2001.

[17] L.W. Ku and H.H. Chen, Mining Opinions from the Web: Beyond Relevance Retrieval, *Journal of American Society for Information Science and Technology,* pp. 1838-1850, 2007.

[18] Jun Li and Maosong Sun, Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques, in *Proceeding of IEEE NLPKE 2007*.

[19] J.J. Mei, Y.M. Zhu, Y.Q. Gao, H.X. Yin. Synonymy Thesaurus (Second Edition). *Shanghai Lexicographical Publishing House.*

[20] L. W. Ku, T. H. Wu, L. Y. Lee, and H. H. Chen. Construction of an evaluation corpus for opinion extraction. in Proc. of NTCIR-5 Workshop. pp. 513–520, Tokyo, Japan, 2005.

[21] Yunqing Xia,Linlin Wang,Kam-Fai Wong and Mingxing Xu.Sentiment Vector Space Model for Lyric-based Song Sentiment Classification. *ACL 2008.*