Supervised Approaches and Ensemble Techniques for Chinese Opinion Analysis at NTCIR-7

Bin Lu, Benjamin K. Tsou and Oi Yee Kwong
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
lubin2@student.cityu.edu.hk {rlbtsou, rlolivia}@cityu.edu.hk

Abstract

For the opinion analysis task on traditional Chinese texts at NTCIR-7, supervised approaches and ensemble techniques have been used and compared in our participating system. Two kinds of supervised approaches were employed here: 1) the supervised lexicon-based approach, and 2) machine learning approaches. Ensemble techniques were also used to combine the results given by different approaches. By making use of these approaches and ensemble methods in various combinations, we submitted three runs for each of the two subtasks we participated in: opinionated sentence recognition and opinion polarity classification. The results show that our system achieved state-of-the-art performance on both subtasks: the highest F-measure on the opinionated sentence recognition task and the second highest F-measure on the opinion polarity classification task amongst all runs submitted by seven participants. Furthermore, the ensemble combination of different classifiers markedly outperformed individual classifiers on the opinion polarity classification task, without showing much improvement, however, on the opinionated sentence recognition task.

Keywords: Chinese *opinion analysis, machine learning, ensemble techniques, supervised approaches*

1 Introduction

The analysis of opinion information has been an active area of computational linguistics and NLP fields in recent years. With the large amount of opinionated texts available online, opinion analysis has many potential applications such as helping companies analyze customer opinions, or helping governments obtain feedbacks on policies and regulations.

Opinion analysis is also known as sentiment classification [10], polarity classification [18], opinion mining [4], subjectivity analysis [2], etc. Although these terms have subtle differences, for the purpose of this paper we do not differentiate amongst them. Many researchers have studied opinion analysis from various perspectives: initial analysis of polarity-bearing words or phrases [2], the analysis of subjective and objective attributes in terms of sentences and documents [21], the polarity classification of sentences and documents [10], and even more complicated opinion mining and sentiment summarization [4].

The NTCIR-7 Multilingual Opinion Analysis Task (MOAT) [13] provides an opportunity to evaluate the techniques used by different participants based on a common framework of evaluation in Chinese (simplified and traditional), Japanese and English, following the opinion analysis pilot task (OAPT) at NTCIR-6 in 2007. NTCIR-6 OAPT was introduced in the overview paper [12], as well as the reports of five participants in [5, 7, 8, 23, 24]. A newspaper corpus and 16 news topics were used for the evaluation at the traditional Chinese side of NTCIR-7 MOAT. The data was drawn from a corpus covering news from 1998 to 2001 from a variety of sources in Japanese, English, and Chinese. Participants need to tag opinions at a sub-sentence level and annotate the opinion features, including opinionatedness, relevance, polarity, opinion holder and opinion target. More descriptions are covered in the overview paper.

This paper describes the CityU (HK) system used in the traditional Chinese task at NTCIR-7 MOAT. We participated in two of the five subtasks: opinionated sentence recognition and opinion polarity classification. Three runs were submitted for both subtasks exploring two kinds of supervised approaches: the supervised lexicon-based approach and machine learning approaches, as well as the ensemble methods. The result shows that the system achieved state-of-the-art performance on both subtasks: the highest F-measure on the opinionated sentence recognition task and the second highest F-measure on the opinion polarity classification task amongst all runs submitted by seven participants. Furthermore, the ensemble combination of different classifiers shows marked improvement on the opinion polarity classification task when compared with individual classifiers, while it did not improve the opinionated sentence recognition task much, if not deteriorating it.

The rest of this paper is organized as follows. Section 2 presents the Chinese opinion analysis from the linguistic perspective. The supervised approaches to recognize opinionated sentences and to determine the polarity of opinionated sentences, as well as the ensemble methods, are described at Section 3. Section 4 gives the evaluation results and finally, Section 5 concludes this paper.

2 Linguistic analysis of opinions

According to *the 2008 Trial English Opinion Annotation Instructions* for NTCIR-7 MOAT, the annotation strategy follows the one in [21], which is centered on the notion of private state, a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. In [21], three types of private state expressions are annotated: 1) explicit mention of private states; 2) speech events expressing private states; 3) expressive subjective elements. The guidelines for each type are good indicators for what should be considered opinions. In the following, we will discuss what kinds of linguistic clues can help us with the subtasks.

2.1 Opinionated sentence recognition

The *opinionated sentence recognition* task is a binary sentence classification problem with only two labels, opinionated or not. The opinions in news text maybe explicitly mentioned, or they may be expressed indirectly by the types of words and the style of language that a speaker or writer uses [20]. Three kinds of lexical clues are exploited for this task:

1) **Reporting verbs**: verbs indicating speech events;
2) **Polar items**: items containing polarity, referring to positive, negative or neutral, may be words or phrases.
3) **Adverb clues**: adverbs frequently co-occurring with opinions.

Consider the following sentences[1]:

*a) 報導中引述 KGB 在德國的上司卡魯金的話**說**，普亭的間諜工作並 **不特別成功**。*(The report **said** the spying work of Putin was **not particularly successful**, citing the words of the supervisor of KGB in Germany.)

*b) 金融市場 **可能** **面臨** 不能提供所需資金的**風險**。*(The financial market was **perhaps** **facing the danger** of not being able to provide necessary funds.)

In sentence a) above, the reporting verb *說* (said) indicates a speech event expressing an opinion given by *KGB 在德國的上司卡魯金*. At the same time, the positive item *成功* (successful) plus the negation marker *不* (not) express the negative polarity of the opinion.

In sentence b) above, the adverb *可能* (perhaps) does not mean an opinion, but it co-occurs with an opinion ***面臨…風險*** (face the danger of) here. Some adverbs (e.g. *一定* (certainly), *一向* (all along), *也許* (maybe), *仍然* (still)) exhibit the same characteristics with *可能* (perhaps), i.e. frequently co-occurring with opinions.

2.2 Opinion polarity classification

The *opinion polarity classification* task is to classify opinion fragments into positive, negative, or neutral with respect to the topic. In addition to the *polar items* mentioned in section 2.1, two other kinds of lexical clues were used for polarity classification:

4) **Negation marker:** words used to reverse the polarity of a particular polar item.
5) **Discourse marker:** those discourse markers that may reverse the polarity of previous clause(s) [16].

Each negation marker is assumed to negate the opinion polarity of the closest polar item. In sentence a) above, the negation marker *不* (not) negates the polarity of the polar item *成功* (successful), and the contextual polarity for the phrase *不成功* (not successful) is negative.

Rhetorical structures expressed with discourse markers represent the logical relationships between discourses. Consider the following sentence:

*c) 普亭 **雖然** **支持**這項法案，**但** 俄國民意對這項法案的**反對聲浪高漲**。*(Although Putin **supports** this bill, the majority of Russian people is **highly against** it.)

In sentence c) above, the discourse markers *雖然* (although) and *但* (but) involve an *adversativity* relation between the first clause and the second one, which means that the polarity of the two clauses are opposite to each other. Under such circumstances, we could consider only the polarity of the second but main clause to get the correct polarity for the whole sentence, i.e. the polarity for sentence c) should be negative.

3 Supervised approaches for opinion analysis

The techniques used for sentiment analysis can be divided into machine learning approaches and knowledge-based approaches [1, 24]. The advantage of the machine learning method is the high accuracy given a high-quality training corpus, but it needs extensive manual work to annotate data. In contrast, the methods based on the linguistic knowledge and semantic resources can be easily used in many domains, but sometimes they cannot yield the comparable accuracy to machine learning methods.

3.1 Overview

The motivation of our system is to make full use of both the manually labeled lexicons and annotated corpora to improve opinion analysis. The collected lexicons are introduced in section 3.2.1, and the training corpus includes the Chinese sample and test data for NTCIR-6 OAPT, as well as the sample data in traditional Chinese for NTCIR-7 MOAT.

Different approaches were employed and compared in our system. First, the supervised lexicon-base approach was used to tune and filter the word lists of collected lexicons based on the training corpus, and then employed to the tasks. Then, two standard machine

[1] All the Chinese sentences in this paper are selected from the traditional Chinese data of NTCIR-7 MOAT.

learning approaches were also used based on the training corpus: Naive Bayes classification and support vector machines (SVM). Furthermore, the results of the above individual approaches were then combined to obtain more accurate results under the assumption that the individual approaches can complement each other.

By making use of these approaches and ensemble methods in various combinations, we submitted three runs for both the opinionated sentence recognition task and the opinion polarity classification task. For the first task, the first and second runs were based on ensemble methods, and the third run on the supervised lexicon-based approach. For the latter task, three runs were submitted by classifying the opinionated sentences in the three runs for the first task into different polarities, respectively: the first run was based on the ensemble method, the second run on Naive Bayes, and the third one on supervised lexicon-based approach.

Words were treated as basic features for opinion analysis in traditional Chinese texts in our system. Because there are no word boundaries in written Chinese, the sentences were segmented using a production segmentation system of our center to get words. This production segmentation system draws heavily on a large dictionary derived from processing a very large amount (over 200 million Chinese characters) of synchronous textual data[2] gathered systematically over 13 years from various Chinese speech communities, including Beijing, Hong Kong, Taipei and others [17].

The supervised lexicon-based approach and machine learning approaches, as well as the ensemble methods, are described in details in the next sections, respectively.

3.2 Supervised lexicon-based method

To our best knowledge, most lexicon-based methods just use the original sentiment lexicons without any supervised training [5, 7, 23, 24]. We applied a tuning algorithm introduced in section 3.2.3 to remove lexical items with low precision from the existing lexicons based on the training corpus.

3.2.1 Lexicon preparation

Over the past years, several traditional Chinese resources of polar items have been collected for opinion analysis in our center, including NTU Sentiment Dictionary (NTUSD)[3], *The Lexicon of Chinese Positive Words (*LCPW*)*[11], *The Lexicon of Chinese Negative Words* (LCNW) [26], and CityU's polar word and phrase list (CPWP), which were manually marked in the political news data by trained annotators from our center. Polar items marked with the *SENTIMENT_KW* tag (SKPI), including only positive and negative items but not neutral ones, were also automatically extracted from the Chinese sample data of NTCIR-6 OAPT.

All these polar item lexicons were combined, and the combined polar item lexicon[4] consists of 13437 positive items and 18365 negative items, a total of 31802 items. Apart from words, an item can also be a phrase, e.g. *不負責任的表現* (irresponsible performance) in CPWP, *努力解決* (work hard to solve) in NTUSD. We also did some manual cleanup on the polar items in the combined lexicon, especially those marked as both positive and negative, to make sure most items are appropriate. Table 1 shows the statistics of lexicons mentioned above.

The SKPI lexicon contains some adverb clues mentioned in section 2.1, and we, on purpose, did not remove them because we think they could help the opinionated sentence recognition task. For example, the adverbs mentioned in section 2.1 all appear in SKPI.

The reporting verbs were firstly collected from the Chinese sample data of NTCIR-6 OAPT in which the *OPINION_OPR* tag was used to mark them. The list was extended from 68 to 308 words through manual synonym search in HowNet[5], WordNet and Tongyici Cilin [9].

There were 48 negation markers manually collected by inspection of HowNet and Tongyici Cilin. We only choose a small set of discourse markers (i.e. 5 single markers and 7 pairs of markers), such as the *adversativity* and *concession* markers including *但 (but), 可是 (however), 雖然…但是… (Although…, but …), 儘管…還… (Although...still...)*, which can be used to reverse the polarity of previous discourses, although there are different kinds of Chinese discourse markers according to [16].

3.2.2 Recognizing opinionated sentences and classifying polarity

The above lexicons were exploited to detect opinionated sentences and classify their polarity based on the polarities of its constituent items. Given a segmented sentence, we checked whether a polar item (including adverbs) or a reporting verb occur in it. If a sentence contains at least one item in the combined polar item lexicon (including the adverb clues) or the reporting verb lexicon, it was reported as opinionated, otherwise not opinionated.

The polarity of an opinionated sentence was calculated by checking the polar item lexicon. A polar item gets a score of 1 if it is marked as positive in the combined lexicon, a score of -1 if negative, and a score of 0 if both positive and negative. The sentence was reported as positive if its sum score of all its component polar items is bigger than 0, negative if smaller than 0 and neutral equal to 0.

Negation markers and discourse markers were also considered. Each negation marker would negate the polarity of its following polar items within a window of certain words. We tuned the window size to get the

[2] http://livac.org

[3] http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html

[4] An item could be positive and negative at the same time, and we treat it as two items. For example, *防止* is marked as positive in NTUSD, and negative in SKPI.

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

optimal performance by using the supervised technique based on the training corpus, and here it was set to 3 words. When a discourse marker in the lexicon occurs in the sentence, we would only consider the polarity of the second but main clause to get the correct polarity for the whole sentence.

Table 1. Statistics of lexicons

	NTUSD	LCPW	LCNW	CPWP	SKPI	**Combined**
# Positive items	2812	5046	0	5838	2426	13437
# Negative items	8276	0	3499	9002	1252	18365
Total	11088	5046	3499	14840	4234	31802

3.2.3 Supervised lexicon adjustment

The above combined polar item lexicon, as well as the reporting verb lexicon, were adjusted on the training data, respectively. There are two reasons for this adjustment: a) lexical items were collected from multiple sources, which have not been cleaned, and could contain errors or typos; especially those polar items in SKPI are not quite clean, such as *随着 (with)*, *可以 (be able to)*, etc.; b) these items were marked by annotators with their own subjective judgement, and they could be contextual or not suitable for the news domain. The training algorithm for the opinionated sentence recognition task is illustrated in Figure 1 and Figure 2. For the tuning procedure, the lenient gold standard of the training data was used, and the standard precision (P), recall (R) and F-measure (F) were used as an evaluation metric.

```
Input: the polar item lexicon, the reporting verb lexicon, and the training corpus;
Output: the filtered polar item lexicon, and the filtered reporting verb lexicon;
Algorithm 1: Supervised Lexicon Tuning
1. For each reporting verb, compute its precision for this task based on the training data.
2. For each polar item, compute its precision for this task based on the training data.
3. Compute the best threshold combination of the reporting threshold $\theta_{reporting}$ for reporting verbs
   and the polar threshold $\theta_{polar}$ for polar items: $\theta_{reporting}^{Max}$ and $\theta_{polar}^{Max}$.
   The concrete algorithm for this step is illustrated in Figure 2.
4. Remove reporting verbs with precision lower than $\theta_{reporting}^{Max}$ from the original lexicon
   to get the filtered reporting verb lexicon;
5. Remove polar items with precision lower than $\theta_{polar}^{Max}$ from the original lexicon
   to get the filtered polar item lexicon.
```

Figure 1. Algorithm for supervised lexicon tuning

```
Input: the polar item lexicon, the reporting verb lexicon, and the training corpus;
Output: the best threshold combination of $\theta_{reporting}^{Max}$ and $\theta_{polar}^{Max}$;
Algorithm 2: Computing Best Threshold Combination
1. For $\theta_{reporting}$ from 0%  to 100% (1% for each step)
2. Remove reporting verbs with precision lower than $\theta_{reporting}$
3. For $\theta_{polar}$ from 0%  to 100% (1% for each step)
    a) Remove polar items with precision lower  than $\theta_{polar}$
    b) Compute the F-measure of the opinionated sentence recognition task for the current thresholds
       $\theta_{reporting}$ and $\theta_{polar}$ based on the whole training set, with the method mentioned in section 3.2.2.
    c) Record the best threshold combination of $\theta_{reporting}^{Max}$ and $\theta_{polar}^{Max}$
       that give the highest F-measure.
```

Figure 2. Algorithm for computing best threshold combination

Two kinds of items could be filtered out during this process: 1) noisy terms which were actually not reporting verbs or polar items according to our judgment, e.g. 觀光 (sightseeing) in LCPW, 定下 (set) and 前往 (head for) in SKPI; 2) reporting verbs or polar items which may present facts and frequently occur in factual sentences, e.g. 暴雨 (downpour) in NTUSD and 襲擊 (attack) in NTUSD and CPWP. The lexicons of remaining reporting verbs and polar items obtained from algorithm 1 were used for the opinionated sentence recognition task with the method in section 3.2.2.

For the polarity classification task, the separate and similar training algorithm was used to filter the polar items, but the reporting verbs were not used for this task. This separate training process was undertaken due to the following two reasons: 1) the majority of the reporting verbs are not related to polarity, thus the threshold combination computing becomes unnecessary, and the polar threshold needs to be recalculated; 2) some adverbs in the polar item lexicon are not related to polarity (e.g. *可能* (perhaps), *一定* (certainly), *一向* (all along), *也許* (maybe)), and hence their precision for the polarity task would be different from that of the opinionated task. When the separate tuning process was finished, the filtered polar item lexicon for the polarity task was ready. With the polarity computing method noted in section 3.2.2, the polarity was computed for each opinionated sentence.

3.3 Machine learning approaches

The *opinionated sentence recognition* task is a sentence classification problem with only two labels, opinionated or not. However, the *opinion polarity classification* is a three-class sentence classification problem with three labels, *positive*, *negative* or *neutral*. The unigram of Chinese words was used as the linguistic feature for machine learning.

Three machine learning approaches were explored: Naive Bayes classification (Bayes), maximum entropy classification, and support vector machines (SVM). Each machine learning method was trained on the training data, and then the learned models were applied to classify new test data. Note these models were respectively trained for each of the two subtasks, i.e. there were two respective models for two subtasks. For example, there are two SVM classifiers: the SVM opinionated classifier for the opinionated sentence recognition task and the SVM polarity classifier for the polarity classification task.

The maximum entropy model was not used in subsequent testing because of its unsatisfactory performance on the training data when compared with the other two models (i.e. we only used Bayes and SVM for the submitted runs). For SVM, we constructed one feature vector for each sentence with each unigram term of the sentence as one dimension and the frequency of the term as the weight of this dimension; the feature vector was normalized and then fed into the SVM algorithm for learning. Joachim's SVM^{light} package[6] was used for training and testing.

3.4 Ensemble combination

After obtaining the set of classifiers (e.g. the supervised lexicon-based classifier, SVM classifier and Bayes classifier), we could exploit different ensemble methods to combine the results of individual classifiers. The commonly used ensemble strategies include majority voting, sum, product, max, min, etc [11, 19]. We just used the intuitive majority voting strategy for the opinionated sentence recognition task and the sum strategy for the polarity classification task to derive the new result from the results of the component classifiers.

For the *opinionated sentence recognition* task, we exploited two ensemble schemes: the lenient voting scheme and the strict voting scheme. The underlying ideas behind these two schemes are the lenient and strict standard for evaluation, respectively. For the lenient voting scheme, if two of the three component classifiers mark a sentence as opinionated, the sentence would be marked as opinionated, just like the lenient standard for evaluation; while the strict voting scheme means that if and only if all three classifiers mark a sentence as opinionated, the sentence would be marked as opinionated, otherwise not opinionated.

For the *polarity classification* task, given an opinionated sentence and the polarity tags from three component polarity classifiers, we just get its polarity score by adding the scores of polarity tags (i.e. 1 for positive, 0 for neutral, and -1 for negative). The polarity tag would be *positive* if the sum of the scores is bigger than 0; *negative* if the sum is smaller than 0; otherwise neutral.

4 Results and discussion

The above approaches were applied to the traditional Chinese task at NTCIR-7 MOAT. Standard precision (P), recall (R) and F-measure (F) were used to evaluate the performance on each subtask, and a Set Precision (S-P) was also used [13]. More descriptions about the gold standards and metrics are covered in the overview paper. All the results presented in this section were released by the organizers. There were two types of evaluation: a lenient standard and a strict one, since each sentence was annotated by three annotators.

4.1 Opinionated sentence recognition

Three runs were submitted for this task. The first run used the lenient voting scheme based on three classifiers: Naive Bayes classifier, SVM classifier, and the supervised lexicon-based classifier; the second run applied the strict voting scheme based on the same three classifiers; and the third run only used the supervised lexicon-based classifier. Table 2 reports the performance of these three runs for this task.

Table 2. Results of opinionated sentence recognition

Run	Lenient			Strict		
	P	R	F	P	R	F
1	0.660	0.782	0.716	0.837	0.852	**0.844**
2	0.743	0.605	0.667	0.900	0.691	0.782
3	0.652	0.805	**0.721**	0.819	0.869	0.843

For the lenient standard, the supervised lexicon-based classifier (i.e. run 3) achieves the best performance based on F-measure, slightly better than that of the lenient

[6] http://svmlight.joachims.org/

voting scheme (i.e. run 1) and much better than the strict voting scheme (i.e. run 2), which improves the precision by 9-11%, but lowers the recall by about 18-20%. For the strict standard, the lenient voting scheme achieves the best performance based on F-measure, slightly better than that of the supervised lexicon-based classifier and much better than the strict voting scheme, which improves the precision by 7-9%, but deteriorates the recall by about 6-11%.

Overall, our system achieved the best performance in this task, with the highest F-measures for the lenient and strict standard reported amongst all submitted runs by the 7 participating teams, 0.7206 and 0.8444, respectively. This result demonstrates the effectiveness of the supervised lexicon-based classifier and the ensemble methods on this opinionated sentence recognition task.

4.2 Opinion polarity classification

Three runs were submitted for this subtask. The first run used the above ensemble method based on three component classifiers to determine the polarity of each opinionated sentence; the second run just used the Naive Bayes polarity classifier; and the third run only used the supervised lexicon-based polarity classifier. Table 3 shows the results of the three runs of polarity classification for opinionated sentences in the three runs for the opinionated sentence recognition task, respectively.

For both the lenient and the strict standard, the ensemble method (i.e. run 1) outperformed the other two runs based on Set-precision and F-measure, and achieved the second highest F-measures amongst all submitted runs by the seven participants for the lenient and strict standard, 0.3854 and 0.4642, respectively.

The results show the effectiveness of the classifier combination approaches on this polarity classification task. Another point is that the Bayes polarity classifier (run 2) achieves higher Set precision, but lower F-measure than the supervised lexicon-based polarity classifier (run 3).

Table 3. Results of polarity classification

Run	Lenient				Strict			
	S-P	P	R	F	S-P	P	R	F
1	**0.536**	0.354	0.453	**0.397**	**0.546**	0.457	0.494	**0.475**
2	0.520	0.386	0.339	0.361	0.529	0.476	0.388	0.427
3	0.505	0.329	0.440	0.377	0.526	0.430	0.485	**0.456**

4.3 Discussion

The above results show that our system achieved state-of-the-art performance on the two subtasks we participated in. The supervised lexicon approach performed well on the two subtasks, which shows that it made good use of both human-annotated lexicons and annotated corpus to improve the performance.

Since three classifiers were employed here, we further computed the overlap of answers given by them for both subtasks, which is shown in Table 4[7]. It is noted that the overlap ratio between Bayes and SVM is much higher that between the supervised lexicon-based classifier and the other two kinds of classifiers for both subtasks. This seems reasonable since SVM and Bayes are both machine learning methods which are based on prior knowledge of the training set without using manually constructed lexical resources.

Table 4. The overlap statistics of answers

	Opinionated	Polarity
Lexicon vs Bayes	0.724	0.510
Lexicon vs SVM	0.747	0.477
Bayes vs SVM	0.809	0.703

For the opinionated sentence recognition task, the ensemble method did not improve, if not deteriorate, the performance much on both the lenient and the strict standard. There are two possible reasons for this phenomenon: 1) the machine learning approaches did not perform as well as the supervised lexicon-based such that the ensemble combination lowers the overall performance, 2) the ensemble method lowered the overall performance although the machine learning approaches did not perform worse than the supervised lexicon-based.

For the polarity classification task, the ensemble method performed much better than the supervised lexicon-based approach on both the lenient and the strict standard. The marked improvement may be due to two reasons: 1) machine learning approaches can utilize word features and improve the performance when compared with the supervised lexicon-based approach; 2) the individual classifiers can complement each other, and thus improve the overall performance of this task.

Although the proposed approaches work well in general, there is still much room to improve, and several kinds of errors are analyzed in the following section. First, since sentences were automatically segmented into Chinese words using a production segmentation system and these words were used as the basic features for all three kinds of approaches, inevitably some errors from

[7] The overlap statistics on the polarity classification task is computed on the opinionated sentences in the first run of the opinionated sentence recognition task.

the word segmentation step might have propagated to the opinion analysis step.

Second, some polar items or reporting verbs are context-dependent. Consider the two sentences below: a) *友訊廣達與美股連動大近期可望有* ***表現*** (YouXun and GuangDa, closely connected with US stock market, are expected to have good performance in the near future); b) *受戰事拖累，印尼成為上周* ***表現*** *最差的市場* (Encumbered with wars, Indonesia market showed the worst **performance**). The word ***表現*** is positive in sentence a), neutral in sentence b), and also marked as positive in the Chinese sample data of NTCIR-6 OAPT. Some polar items or reporting verbs are also topic-related. For the example, the word ***暴雨*** (heavy downpour) is marked as negative in NTUSD, but in the second topic of traditional Chinese sample data at NTCIR-7 MOAT (*MOAT7-CT-N0: 中國大陸洪水:),* it just refers to the fact without any opinion, e.g. *持續不斷的****暴雨****, 使得福建、江西等地相繼傳出重大災情* (Due to the continuous **heavy downpour**, severe disasters happened in several provinces, such as Fujian, Jiangxi, etc.). The same word could carry different polarity when associated with different topics.

Third, if multiple polar items occur in one sentence, the relationship between them could not be simply treated as addition, and the polarity of the whole sentence may be wrong if we just add the polarity scores of the polar items. For example, consider the following sentence: *包括應用材料、科磊、Novellous 等企業未來兩季****獲利****將再度****下調****。* (Companies, such as Applied Material, KeLei, Novellous, etc., had again **decreased** their **profit** expectancy in the next two seasons.) It contains two polar items which exist in the combined polar item lexicon: ***獲利*** (profit) and ***下調*** (lower). If we just add the polarity scores of them, we would get a score of 0 (neutral) for the whole sentence; while this sentence is negative, in fact. Therefore, the accurate relationship between polar items has to be identified, possibly with some shallow parsing techniques.

These kinds of errors would adversely affect the performance of both the supervised lexicon-based approach and machine learning approaches, and therefore propagating to the classifier combination approach as well.

5 Conclusion and Future Work

This paper presents the supervised approaches and ensemble methods used in our system participating in the opinion analysis task on traditional Chinese texts at NTCIR-7. The supervised lexicon-based approach, machine learning approaches and ensemble combination methods were explored for two subtasks: opinionated sentence recognition and opinion polarity classification.

The Chinese training and test data for NTCIR-6 OAPT, as well as the traditional Chinese training data for NTCIR-7 MOAT, were exploited as the training set for the test of NTCIR-7 MOAT. The system achieved state-of-the-art performance on both subtasks. The supervised lexicon-based approach achieved good performance by making good use of both human-annotated lexicons and the annotated corpus. Furthermore, for the opinionated sentence recognition task, the supervised lexicon-based approach performed well, and the ensemble methods did not improve the performance much; while for the polarity classification task, the ensemble method performed much better than the supervised lexicon-based approach on both the lenient and the strict standard.

There is much room for the system to improve on opinion polarity classification. Hence part of our future work would be investigating the unique features of the opinion polarity classification. The potential contribution of contextual information, topic-related features, and shallow parsing techniques would be further investigated. At the same time, we would try our system on other subtasks at NTCIR-7 MOAT: relevant sentence detection and opinion holders/targets extraction.

Acknowledgement

Research presented here is supported in part by a Research Grant from the HKSAR Research Grants Council under grant No. CityU149607. We also greatly appreciate the help of our teammates (Jin Wang, Yan Nga Cheung, and Pui Ling Chong), and the help of Prof. Zhu Jingbo at the Northeastern University, China for providing the machine learning tools.

References

[1] Chaovalit P., Zhou L. 2005. Movie review mining: a comparison between supervised and unsupervised classification approaches, In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05) - Track 4*. p. 112c.

[2] Hatzivassiloglou V. and McKeown R.K. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the ACL 1997*. pp.174-181.

[3] Hatzivassiloglou V. and Wiebe J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of ACL 2000*, New Brunswick, NJ.

[4] Hu M.Q. and Liu B. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM-KDD 2004*, pp.168-177.

[5] Huang R.H., Pan L.X., Sun L. 2007. ISCAS in Opinion Analysis Pilot Task: Experiments with sentimental dictionary based classifier and CRF model. *Proceedings of the Sixth NTCIR Workshop*. pp. 296-300.

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

[7] Ku L.W., Wu T.H., Lee L.Y., and Chen H.H. 2007. Using Polarity Scores of Words for Sentence level Opinion Extraction. *Proceedings of the Sixth NTCIR Workshop*. pp. 316-322.

[8] Li Y.Y., Bontcheva K. and Cunningham H. 2007. Experiments of Opinion Analysis on the Corpora MPQA and NTCIR-6. *Proceedings of the Sixth NTCIR Workshop*. pp. 323-329.

[9] Mei J.J., Zhu Y.M., Gao Y.Q., Yin H.X. 1996. Tongyici Cilin (同義詞詞林) (The 2nd Version). Shanghai Lexicon Press.

[10] Pang B., Lee L., and Vaithyanathan S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, pp.79–86.

[11] Polikar R. 2006. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, vol.6, no. 3, pp. 21-45.

[12] Seki Y., Evans D.K., Ku L.W., Chen H.H., Kando N., and Lin C.-Y. 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proc. of the Sixth NTCIR Workshop*. May 2007, Japan.

[13] Seki Y., Evans D.K., Ku L.W., Sun L., Chen H.H., and Kando N. 2008. Overview of multilingual opinion analysis task at NTCIR-7. *Proc. of the Seventh NTCIR Workshop*. Dec. 2008, Japan.

[14] Shi J.L., Zhu Y.G.. 2006. The Lexicon of Chinese Positive Words (褒義詞詞典). Sichuan Lexicon Press.

[15] Turney P.D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL 2002*, pp.417–424.

[16] Tsou, B K, H L Lin, H C Ho & T B Y Lai. (1996). From Argumentative Discourse to Inference Trees: Using Syntactic Markers as Cues in Chinese Text Abstraction. *Readings in Chinese Natural Language Processing*, In C.-R. Huang, K.-J. Chen & B.K. Tsou eds. *Journal of Chinese Linguistics*. Monograph

[17] Tsou B.K.Y., Tsoi W.F., Lai T.B.Y., Hu J., and Chan S.W.K. 2000. LIVAC, A Chinese Synchronous Corpus, and Some Applications. *Proceedings of the ICCLC International Conference on Chinese Language Computing*, Chicago. pp. 233–238.

[18] Tsou B.K.Y., Yuen W.M.R., Kwong O.Y., Lai T.B.Y., Wong W.L. 2005. Polarity classification of celebrity coverage in the Chinese press. In *Proceeding of the 2005 International Conference on Intelligence Analysis*. Virginia, USA.

[19] Wan X.J. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In *Proceeding of EMNLP2008*. pp. 553-561.

[20] Wilson T. and Wiebe J. 2003. Annotating opinions in the world press. *Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*.

[21] Wiebe J., Wilson T., Bruce R., Bell M., and Martin M. 2004. Learning Subjective Language. *Computational Linguistics*, 30(3): pp.277-308.

[22] Wiebe J., Wilson T., Cardie C. 2005. Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.

[23] Wu Y.J. and Oard D. 2007. NTCIR-6 at Maryland: Chinese Opinion Analysis Pilot Task. *Proceedings of the Sixth NTCIR Workshop*. pp. 344-349.

[24] Xu R.F., Wong K.F. and Xia Y.Q. 2007. Opinmine – Opinion Analysis System by CUHK for NTCIR-6 Pilot Task. *Proceedings of the Sixth NTCIR Workshop*. pp. 350-357.

[25] Ye Q., Liu B., Li Y.J. 2005. Sentiment Classification for Chinese Reviews: A Comparison between SVM and Semantic Approaches. In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, pp.2341-2346.

[26] Zhu L., Zhu Y.G. 2006. The Lexicon of Chinese Negative Words (貶義詞詞典). Sichuan Lexicon Press.