NTCIR-6 at Maryland: Chinese Opinion Analysis Pilot Task

Yejun Wu and Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
{wuyj,oard}@umd.edu

Abstract

For the Chinese opinion analysis pilot task at NTCIR-6, we tested two techniques for each of the four subtasks—identifying opinionated sentences, making polarity decisions, identifying opinion holders, and retrieving topically relevant sentences. Our opinion detection technique is based on sentiment lexicons. We explored three main issues: the effect of the size of sentiment lexicons on the accuracy of opinionated sentence identification and of polarity decisions, the effect of a simple approximation to anaphora resolution on the accuracy of opinion holder identification, and the effect of sentence expansion on the effectiveness of relevant sentence retrieval.

Keywords: *NTCIR, Opinion analysis, sentiment detection, sentence retrieval.*

1 Introduction

Accurate and efficient techniques for opinion analysis could serve many purposes, ranging from current applications to marketing and public opinion research to scholarship on evolution of attitudes and opinions over time. The vast majority of the reported research on opinion classification has focused on written English. The NTCIR-6 Chinese opinion analysis pilot task is the first effort in the world to create a test collection for Chinese opinion detection at sentence level.

NTCIR provided a test collection of 843 topically relevant news documents with 11,907 sentences in Traditional Chinese for opinion analysis. There are 4 topics (i.e., 1, 2, 3, 26) for training and 28 topics (i.e., 4-25, 27-32) for evaluation. The training documents are marked with opinionated words, sentence relevance, and sentence polarity. Evaluations of the task were made by the NTCIR organizers. More descriptions of the test collection and the task design are covered in the overview paper [3].

The pilot task evaluates 4 subtasks—identification of opinionated sentences, identification of opinion holders, retrieving relevant sentences, and deciding opinion polarity. We tested relatively simple techniques with a one-month effort, and submitted 2 runs (a baseline run and an alternative run) for each subtask. This paper reports our methods for and the results of accomplishing the subtasks.

2 Methods

Supervised machine learning techniques have been applied to the identification of semantic polarity at the scale of words [6], sentences [9], and documents [8]. Regardless of scale, however, words typically have provided the base feature set that those classifiers exploited. We treat words as our basic features for Chinese opinion detection, although characters would also have been a reasonable choice.

In our NTCIR-6 Chinese opinion analysis experiments, we tested the following ideas: (1) detecting opinionated sentences using sentiment lexicons, (2) aggregating sentence polarity from word polarity, (3) using BBN Identifinder and some simple anaphora resolution technique to identify opinion holders, and (4) retrieving relevant sentences based on sentence expansion.

2.1 Segmentation

Since there are no word boundaries (other than those that are coincident with sentence boundaries) in written Chinese, we segmented the sentences using a one-best partition by the Stanford Segmenter [5] to get words. Although segmentation of Chinese sentences is a research problem itself, one-best segmentation techniques offer an obvious starting point for our purpose.

2.2 Detecting Opinionated Sentences and Deciding Polarity

We used sentiment lexicons to detect opinionated sentences and to aggregate the polarity of a sentence from the polarities of its constituent words. We first describe the source of our lexicons and then explain how they were used for each task.

2.2.1 Lexicon Acquisition and Preparation

We obtained the Chinese sentiment lexicon used by Ku et al.[1], which contains 2,812 positive words and 8,276 negative words in Traditional Chinese [2]. We also manually rekeyed two books, the Chinese Positive Dictionary and the Chinese Negative Dictionary [4, 7], both published in Simplified Chinese. We rekeyed them as traditional Chinese. Some of the words in each book were marked as being positive (or negative) in some contexts but neutral in other contexts; we segregated 325 of those both positive and neutral words into the Neutral lexicon. This yielded a total of 5,184 positive words, 3,116 negative words and 325 neutral words. Combining these with Ku et al's lexicons, removing duplicates, and cleaning up some punctuation and the characters preceding an ellipsis, yielded a positive lexicon with 6,743 Chinese words and a negative lexicon with 10,294 words.

The 4 sets of training documents (for the 4 training topics) were marked with PSV (opinion indication words), POS (positive) words, NEG (negative) words, and NEU (neutral) words. We automatically extracted these annotated words and put them into 4 lexicons. Since the training documents were created by 3 annotators, we created 2 versions of the POS, NEG, and NEU lexicons - the bigger lexicons were created by combining all words annotated by the 3 annotators (i.e., by set union) whereas the smaller lexicons were composed of those words agreed by at least 2 annotators. A single PSV lexicon was created, by set union. Then we manually edited the lexicons by moving some words among the lexicons (see Figure 1 for full details). For opinion detection, the PSV lexicon was added into the NEU lexicon after removing the word "Bao Dao" (i.e., report) which appears in almost every news story. However, for identifying opinion holders, the full PSV lexicon was used.

Since negation words may reverse the polarities of opinionated words, we also manually created a lexicon of negation words by inspection from the acquired and extracted lexicons and by brainstorming. Some example negation words are listed in Figure 2.

By joining the acquired lexicons with a version of the extracted lexicons, we created two combined lexicons. The number of words in each lexicon is shown in Table 1.

2.2.2 Detecting Opinionated Sentences

We created a very lenient classifier for detecting opinionated sentences using the sentiment lexicons. If a sentence has at least one word appearing in the POS, NEG, or NEU lexicons, it was reported as opinionated. If no word in the sentence was found in those three lexicons, it was reported as "not opinionated." The baseline system used the smaller lexicons, whereas the alternative system used the bigger lexicons.

[1] http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html

```
Removed from PSV:
    報道 (report)

Removed from POS:
    提供 (provide)
    創始 (initiate)
    鼎立 (tripartite confrontation)

Added to NEG:
    隻是 (only)
    而已 (nothing more)

Removed from NEU:
    報道 (report)
    髮行 (issue/distribute)
    提供 (provide)
```

Figure 1. Manual changes to the extracted lexicons.

```
不 (no)
沒 (no)
無 (no)
未 (not)
不介入 (not involved in)
不受 (not subject to)
不相信 (not believe)
不以為然 (object to)
不再 (not any more)
不復 (not any more)
沒有 (no)
沒那麼 (not that)
從來未 (never ever)
從沒有 (never ever)
無意 (not intend to)
全無 (totally not)
```

Figure 2. Example negation words.

2.2.3 Aggregating Sentence Polarity

The polarity of a sentence was aggregated from its composing words by checking the POS, NEG, and NEU lexicons. If a word is found in the POS or NEG lexicon, it gets a score of 1 for positive or a score of -1 for negative, otherwise a score of 0.

Because negation characters (and bigrams) can be segmented as separate words, we added an additional rule to all the 3 sentence classifiers that flipped the polarity of any word that immediately followed such a negation word. The sentence was reported as positive if its aggregated opinion score is $>= 1$, negative if the score is $<= -1$, neutral if the score $= 0$ and there is at least one word appears in the POS, NEU, or NEU

Table 1. Lexicon size

	Bigger lexicon	Smaller lexicon
POS	9150	8195
NEG	13038	12425
NEU	1357	423
PSV	82	82
Negation	116	116

Table 2. Features of the two systems

Subtask	Baseline	Alternative
Sentence Retrieval	single sentence	expanded sentence
Opinion Holder	first identified entity	some anaphora resolution
Opinion Identification	smaller lexicon	bigger lexicon

lexicon. The baseline system used smaller lexicons whereas the alternative system used bigger lexicons.

According to the evaluation plan [3], "the polarity is to be determined with respect to the set topic description if the sentence is relevant to the topic, and based on the attitude of the opinion if the sentence is not relevant to the topic." Here we simplified the decision of polarity without taking the topic description into consideration due to time constraints.

2.3 Opinion Holder Identification

We applied the BBN Identifinder [1] named entity tagger (2001 version) to automatically annotate PERSON, ORGANIZATION, and LOCATION entities. We then used two strategies to identify the focal named entity—the simple reported the first identified PERSON, ORGANIZATION, or LOCATION (in that order), the complex strategy used PSV opinion indication words (usually verbs such as "said," "announced") to anchor the position of named entities, and reported the first identified PERSON, ORGANIZATION, or LOCATION (in this order) before the opinion indicating verbs. If named entities were not identified before any opinion indicating word, the named entity of the previous sentence was reported (which can sometimes function as a simple approach to anaphora resolution). If no opinion indicating words were identified, the simple strategy was applied. The baseline system used the simple strategy, whereas the alternative system used the complex strategy. Although more than one opinion holder might actually be mentioned in a sentence, we only reported one single opinion holder (if any) in order to simplify our efforts, and we did not use the 4 training topics to further train the BBN Identifinder.

2.4 Relevant Sentence Retrieval

In information retrieval, the unit of retrieval is often referred to informally as a "document," even though it might be just part of a document; we adopt that convention here. We created two collections, one for each system. For the baseline system we used sentences as the retrieval unit, while for the alternative system we expanded each retrieval unit to also include a second copy of that sentence, plus the previous and next sentences (if any). This was intended to partially mitigate the effect of vocabulary mismatch when indexing only sentences.

The documents were tokenized with the Stanford Chinese Segmenter, converted from BIG5 to UTF-8, and then indexed with Indri.[2] The queries were created by manually concatenating the title and description fields of the topic files, automatically tokenized with the Stanford Segmenter, and then automatically rendered in the Indri query format.

The Indri search engine reported a ranked list of retrieved documents for each query. We wrote a Perl script to remove duplicates from the ranked list. Lacking any principled way of determining how many top-ranked documents in the ranked list are relevant, we manually checked the ranked lists for 3 randomly selected topics to see at which point the ranked list should be cut off. For the baseline system, we ultimately chose to keep the top 65% documents in the ranked list, while for the alternative system, we chose to keep the top 99%. Since we performed this inspection using three evaluation topics rather than the training topics, our relevant sentence retrieval results should be considered a manual run.

3 Results and Discussion

Since all sentences were annotated by three assessors, there are two types of evaluation - a strict standard (all three assessors must have the same annotation) and a lenient standard (at least two assessors have the same annotation). Both were automatically computed for relevant sentence retrieval, detecting opinionated sentences, and deciding sentence polarity. The opinion holder evaluation required some manual judgment, and was only performed once for each participating group [3].

A brief description of the features of the two systems is presented in Table 2. Since NTCIR evaluated each category (opinionated sentence, holder, relevance, and polarity) separately, we report the performance of the two systems for the 4 subtasks.

[2] http://www.lemurproject.org/indri/

Table 3. Results for detecting opinionated sentences

Runs	Baseline		Alternative	
	lenient	strict	lenient	strict
P	0.6301	0.2388	0.6447	0.2452
R	0.9837	0.9927	0.9738	0.9863
F	0.7682	0.3850	0.7757	0.3928

Note: P=Precision, R=Recall, F=F-measure.

Table 4. Polarity decision results

Runs	Baseline		Alternative	
	lenient	strict	lenient	strict
P	0.2855	0.0812	0.2920	0.0854
R	0.4458	0.6035	0.4412	0.6148
F	0.3481	0.1431	0.3514	0.1500

Note: P=Precision, R=Recall, F=F-measure.

3.1 Detecting Opinionated Sentences

Table 3 reports the accuracy of the two systems for judging whether a sentence is opinionated or not. Compared with the baseline, the alternative system improves the precision by 2–3% relative, but hurts the recall by about 1% relative. Overall, the alternative system did 1–2% better than the baseline by the F-measure. We have not tested these results for statistical significance, but it seems reasonable to conclude that increasing the size of our sentiment lexicon was not generally harmful. However, the precision scores are relatively low for every condition (which means that our systems reported many non-opinionated sentences as opinionated), indicating that substantial room for improvement remains. We should note, however, that 0.7757 was the highest F-measure reported for any of the seven submitted runs by the 5 participating teams.

3.2 Deciding Sentence Polarity

Table 4 reports the accuracy of the two systems for polarity decision (positive, negative, neutral) if a sentence is computed as opinionated. Our systems did not take the relevance of topics to the sentence into consideration when making the polarity decision. The alternative system did a little better than the baseline by improving the F score by 1% relative for the lenient standard, or 5% relative for the strict standard. However, the precision scores are all very low, indicating that our way of aggregating sentence polarity from word polarity is likely flawed.

Table 5. Opinion holder identification results, sentence-based

Runs	Baseline		Alternative	
	lenient	strict	lenient	strict
Correct	917	441	1000	471
ParCorr	213	95	232	103
Incorrect	1051	442	964	405
Miss	257	97	243	96
FA	1976	631	1955	628
P	0.2206	0.2741	0.2409	0.2931
R	0.3761	0.4102	0.4100	0.4381
F	0.2781	0.3286	0.3035	0.3512

Note: P=Precision, R=Recall, F=F-measure, ParCorr=Partial Correct, FA=False Alarm

3.3 Opinion Holder Identification

Table 5 reports the accuracy of our two systems at identifying opinion holders on a sentence-by-sentence basis (referred to in the pilot task as "sentence-based" evaluation results). The definitions of precision, recall, and F-measure for this case are:

$$P = \frac{Corr}{Corr + ParCorr + Incorr + FA} \quad (1)$$

$$R = \frac{Corr}{Corr + ParCorr + Incorr + Miss} \quad (2)$$

$$F = \frac{2PR}{P + R} \quad (3)$$

where *Corr* is the number of sentences in which holders were correctly identified, *ParCorr* is the number of sentences in which partial holders were correctly identified (i.e., not all holders in the sentence were identified), *Incorr* is the number of sentences in which wrong holders were proposed, *FA* is the number of false alarms (i.e., the number of sentences in which holders should not have been proposed), and *Miss* is the number of missed identifications (i.e., the number of sentences in which holders were missed).

Table 5 shows that the alternative system seems to work better than the baseline in terms of precision, recall, and F-measure (although, again, we have not tested these results for statistical significance). Compared with the baseline, the alternative system improves the F-measure by 7–9% relative, indicating our very simple approximation to anaphora resolution is helpful. However, our systems also yields substantial numbers of incorrectly extracted entities and high false alarm scores, indicating that the BBN Identifinder might benefit from further training for this task,

Table 6. Opinion holder identification results, holder-based

Runs	Baseline		Alternative	
	lenient	strict	lenient	strict
Correct	1130	536	1232	574
PropH	4157	1609	4151	1607
ActH	2874	1266	2875	1266
P	0.2718	0.3331	0.2968	0.3572
R	0.3932	0.4234	0.4285	0.4534
F	0.3214	0.3729	0.3507	0.3996

Note: P=Precision, R=Recall, F=F-measure,
Correct=Correct holders identified,
PropH=Proposed holders,
ActH=Actual number of holders

and/or that always selecting the first identified person/organization/location entity could be a suboptimal approach.

Table 6 reports the accuracy of the same two systems computed another way, on a holder basis. In this case, precision and recall are defined as:

$$P = \frac{Correct}{PropH} \tag{4}$$

$$R = \frac{Correct}{ActH} \tag{5}$$

where *PropH* is the number of opinion holders in all opinion sentences proposed by the system, and *ActH* is the number of opinion holders in actual opinion sentences, and *Corr* is the number of correct identifications (i.e., the total number of holders in the correct opinion sentences proposed by the systems).

Table 6 shows that the alternative system seems to work better than the baseline in terms of precision, recall, and F-measure. Compared with the baseline, the alternative system improves the F-measure by 7–9% relative, again indicating that our simple approximation to anaphora resolution is helpful.

3.4 Relevant Sentence Retrieval

Table 7 reports the effectiveness of the two systems for retrieving topically relevant sentences. It shows that the alternative system performed worse than the baseline by decreasing the F-measure by 23% related for the lenient standard, or 9% relative for the strict standard. A precipitous drop in recall is responsible for that large difference, which is just the opposite of what we would have expected from an expansion-based technique. We therefore suspect a bug in our coding for the expansion process.

Table 7. Sentence retrieval results

Runs	Baseline		Alternative	
	lenient	strict	lenient	strict
P	0.6438	0.3537	0.6829	0.4036
R	0.9364	0.9530	0.5163	0.5653
F	0.7630	0.5160	0.5880	0.4709

Note: P=Precision, R=Recall, F=F-measure.

4 Conclusions and Future Work

We tested simple techniques with a one-month effort for accomplishing the four subtasks—identifying opinionated sentences, making polarity decisions, identifying opinion holders, and retrieving topically relevant sentences. For each subtask, we tried two techniques. Here we address what we have learned, and what we might do in the future with this test collection.

For identifying opinionated sentences and making polarity decisions, our alternative system, which relies on larger sentiment lexicons, seems to do about as well as, or perhaps a bit better than, our baseline system. However, both our systems classified a large number of non-opinionated sentences as opinionated, and the precision scores for polarity decisions are thus rather low. Now that we have a test collection to work with, we can try other ways for detecting opinionated sentences. Most obviously, we will want to optimize the thresholds beyond which a sentence is classified as opinionated. Once the precision of detecting opinionated sentences is improved, the precision of polarity decision can be addressed. For example, the training documents might serve as a basis for detecting patterns in the ways that word polarity contributes to sentence polarity. It seems reasonable to expect, for example, that systematic variations by topic might be identified if enough data is available.

For identifying opinion holders, both of our systems rely on the BBN Identifinder. The alternative system, which approximates anaphora resolution using a simple heuristic, did much better than the baseline (which always reports the first identified named entity). This suggests that a more principled approach to anaphora resolution might yield greater gains. We can also likely get some gains by focusing on linguistic cues for expression of opinion rather than simply taking the first entity that we encounter in some fixed search order. Such an approach would also be the first step in crafting a principled basis to identifying cases in which reporting more than one opinion holder would be appropriate. Finally, the fact that we observed large numbers of incorrectly identified named entities and high false alarm rates, so in the future we will want to try some additional task- and genre-specific training for the BBN Identifinder.

For retrieving topically relevant sentences, the obvious first step will be to conduct a failure analysis in order to understand why adjacent-sentence expansion proved to be so harmful. Once we understand that, we can begin to explore alternative expansion techniques and to explore more principled ways of selecting a classification threshold.

Perhaps our most important conclusion is that the NTCIR Opinion Analysis Pilot Task has definitely attained its objectives. We now have an evaluation framework, a test collection, and a community with mutual interest in these challenges. We look forward to discussing our results at the workshop, and to continuing our work on this important problem with the benefit of these new resources.

Acknowledgments

This work was supported in part by the DARPA GALE program.

References

[1] Bikel, Daniel M., Richard Schwartz and Weischedel, Ralph M., 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34(1-3), 211-213

[2] Ku, Lun-Wei, Liang, Yu-Ting, and Chen, Hsin-Hsi, 2006. Opinion extraction, summarization and tracking in news and blog corpora. *AAAI Spring Symposium Technical Report SS-06-03*, Palo Alto, California, 2006.

[3] Seki, Yohei, Evans, David Kirk, Ku, Lun-Wei, Chen, Hsin-Hsi, Kando, Noriko, and Lin, Chin-Yew. Overview of opinion analysis pilot task at NTCIR-6. *Proc. of the Sixth NTCIR Workshop*. May 2007, Japan.

[4] Shi, Jilin and Zhu, Yinggui (Ed.). *Bao Yi Ci Ci Dian (Positive Dictionary)*. Sichuan Dictionary Press, Chengdu, Sichuan, China. 2005.

[5] Tseng, Huihsin, Chang, Pichuan, Andrew, Galen, Jurafsky, Daniel and Manning, Christopher. A Conditional Random Field Word Segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*. 2005.

[6] Turney, Peter and Littman, Michael, 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS*, 21. 315-346.

[7] Yang, Ling and Zhu, Yinggui (Ed.). *Bian Yi Ci Ci Dian (Negative Dictionary)*. Sichuan Dictionary Press, Chengdu, Sichuan, China. 2005.

[8] Wilson, Theresa, Wiebe, Janyce, and Hoffmann, Paul, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT-EMNLP 2005*.

[9] Yu, Hong and Hatzivassiloglou, Vasileios, 2003. Toward answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *EMNLP-2003*, 129-136