Opinion Target Extraction in Chinese News Comments

Tengfei Ma **Xiaojun Wan***

Institute of Compute Science and Technology
The MOE Key Laboratory of Computational Linguistics
Peking University
{matengfei, wanxiaojun}@icst.pku.edu.cn

Abstract

News Comments on the web express readers' attitudes or opinions about an event or object in the corresponding news article. And opinion target extraction from news comments is very important for many useful Web applications. However, many sentences in the comments are irregular and informal, and sometimes the opinion targets are implicit. Thus the task is very challenging and it has not been investigated yet. In this paper, we propose a new approach to uniformly extracting explicit and implicit opinion targets from news comments by using Centering Theory. The approach uses global information in news articles as well as contextual information in adjacent sentences of comments. Our experimental results verify the effectiveness of the proposed approach.

1 Introduction

With the dramatic development of web 2.0, there are more and more news web sites allowing users to comment on news events. These comments have become valuable resources for researchers to make advanced opinion analysis, such as tracking the attitudes to a focused event, person or corporation. In these advanced opinion analysis tasks, opinion target extraction is a necessary step. Unfortunately, former works did not focus on the domain of news comments. Though some researchers and workshops have investigated the task of opinion target extraction in product reviews and news articles, the methods cannot perform well on news comments. Actually, target extraction in news comments significantly differs from that in product reviews and news articles in the following ways.

1) Products usually have a set of definite attributes (e.g. size) and related opinion words (e.g. large), and thus researchers can use a small fixed set of keywords to recognize frequent feature words (Zhuang et al., 2006), or leverage the associated rules between feature words and opinion words to improve the performance (Hu and Liu, 2004; Su et al., 2008; Jin and Ho, 2009; Du and Tan, 2009). But news comments are more complicated. There are much more potential opinion targets in news comments. In other words, the candidate targets are in a much more open domain. On the other hand, the opinion targets in news comments are not strongly associated with the opinion words. We cannot judge a target by a special opinion word as easily as in product reviews.

2) The opinionated sentences in news articles mostly contain opinion operators (e.g. believe, realize), which can be used to find the positions of opinion expressions. However, news comments have already been considered to be declared by readers and they do not have many operators to indicate the positions of opinion targets.

3) Furthermore, many comment sentences are of free style. In many cases, there are even no manifest targets in the comment sentences. For example, a news article and its relational comment are as follows:

News: "迪拜将建超千米全世界最高摩天大楼"
(Dubai will build the highest skyscraper in the world)
Comment:
"真的很高，起到什么作用呢？"
(Really high, but what (is it) used for?)

* Contact author

Coling 2010: Poster Volume, pages 782–790,
Beijing, August 2010

The comment sentence obviously comments on "skyscraper" by human understanding, but in the sentence we cannot find the word or an alternative. Instead, the real target is included in the news article. Now we give two definitions of the phenomenon.

Implicit targets: The implicit targets are those opinion targets which do not occur in the current sentence. The sentence is called implicit sentence.

Explicit targets: The explicit targets are those opinion targets which occur in the current right sentence, and the sentence is called explicit sentence.

In Chinese comments, the phenomena of implicit targets are fairly common. In our dataset, the sentences with implicit targets make up nearly 30 percents of the total.

In this paper, we focus on opinion target extraction from news comments and propose a novel framework uniformly extracting explicit and implicit opinion targets. The method uses both information in news articles and information in comment contexts to improve the result. We extract focused concepts in news articles as candidate implicit targets, and exploit a new approach based on Centering Theory to taking advantage of comment contexts.

We evaluate our system on a test corpus containing different topics. The results show that it improves the baseline by 8.8%, and the accuracy is also 8.1% higher over the popular SVM-based method.

The rest of this paper is organized as follows: The next section gives an overview of the related work in opinion analysis. Section 3 introduces the background of Centering Theory and Section 4 describes our framework based on Centering Theory. In Section 5 we test the results and give a discussion on the errors. Finally Section 6 draws a conclusion.

2 Related Work

The early research of opinion mining only focused on the sentiment classification (Turney et al., 2002; Pang et al., 2002). However, for many applications only judging the sentiment orientation is not sufficient (eg. Hu and Liu, 2004). Fine-grained opinion analysis has attracted more and more attention these years. It mainly includes these types: opinion holder extraction (Kim and Hovy, 2005; Choi et al., 2005), opinion target extraction (Kim and Hovy, 2006; Ruppenhofer et al., 2008), and the identification of opinion proposals (Bethard et al., 2004) and some special opinion expressions (Bloom et al., 2007). Also, there are some other related tasks, such as detecting users' needs and wants (Kanayama and Nasukawa, 2008). However, these general systems are different from ours because they do not have or use any contextual information, and implicit opinion targets are not recognized and handled there.

A more special domain of feature extraction is product and movie reviews. Hu and Liu (2004) design a system to mine product features and generate opinion summaries of customer reviews. Frequent features are extracted by a statistical approach, and infrequent features are generated by the associated opinion words. The product features are limited in amount and they are strongly associated with specific opinion words, so researchers can use a fixed set of keywords or templates to extract frequent features (Zhuang et al., 2006; Popescu and Etzioni, 2005) or try various methods to augment the database of product features and improve the extraction accuracy by using the relations between attributes and opinions (Ghani et al., 2006; Su et al., 2008; Jin and Ho, 2009; Du and Tan, 2009). However, in news comments, the opinion targets are not strongly associated with specific opinion words and these techniques cannot be used.

There are also some works focusing on the target extraction in news articles, such as NTCIR7-MOAT (Seki et al., 2008). Different from the news comments, there are opinion indicators in the subjective sentences. However, in our task of this paper, the opinion holders are pre-assigned as the reviewers, so few opinion indicators and holders can be found.

To our best knowledge, this paper is the first work of extracting opinion targets in news comments. We analyze the complex phenomena in news comments and propose a framework to solve the problems of implicit targets. Our method synthesizes the information from related articles and contexts of comments, and it can effectively improve the extracting results.

3 Background of Centering Theory

Centering Theory (Grosz, Joshi and Weinstein, 1995) was developed for an original purpose of indicating the coherence of a discourse and choosing a referring expression. In the theory, the term "***centers***" of an utterance is used to refer to the entities serving to link this utterance to another utterance in a discourse. But this is not the only function of centers, and there are some other useful characteristics of centers to be recognized. Our observation shows that a center always represents the focus of attention, and the salience of a center indicates the significance of the component as a commented target. In news comments, we consider a comment as a discourse and a sentence as an utterance. If an utterance has a "center", then the center can be regarded as the target of the sentence.

Before introducing the common process of choosing the centers in utterances, several definitions are elaborated as follows:

Forward-looking center: Given an utterance U, there is a set of forward-looking centers $C_f(U)$ assigned. The set is a collection of all potential centers that may be realized by the next utterance.

Backward-looking center: Each utterance is assigned exactly one (in fact at most one) backward-looking center C_b. The backward-looking center of utterance U_{n+1} connects with one of the forward-looking centers of U_n. The C_b is the real focus of the utterance.

Rank: The rank is the salience of an element of C_f. Ranking of elements in $C_f(U_n)$ guides determination of $C_b(U_{n+1})$. The more highly ranked an element of $C_f(U_n)$, the more likely it is to be $C_b(U_{n+1})$. The most highly ranked element of $C_f(U_n)$ that is realized in U_{n+1} is the $C_b(U_{n+1})$. The rank is affected by several factors, the most important of which depends on the grammatical role, with SUBJECT > OBJECT(S) > OTHER.

Preferred center: In the set of $C_f(U_n)$, the element with the highest rank is a preferred center $C_p(U_n)$. This means that it has the highest probability to be $C_b(U_{n+1})$.

Table 1 is an example of the centers. In the example, the target of the first sentence is "Jack", which is exactly the preferred center; while in the second sentence, it is easy to see that "him" gets more attention than "the company" in this environment and thus the backward-looking center is more likely to be the target. So we assume that if $C_b(U_n)$ exists, it can be regarded as the opinion target of U_n, otherwise the $C_p(U_n)$ is the target.

Utterance	Center
U_1:杰克是把公司看作他的生命来做的。 (Jack regards the company as his life.)	C_f: 杰克(Jack)/ 公司(the company)/ 生命(life) C_b: null C_p: 杰克(Jack)
U_2: 公司能有今天的成果都是因为他。 (It attributes to him that the company can obtain today's achievement.)	C_f: 公司(the company)/ 成果(achievement)/ 他(杰克) (him(Jack)) C_b:他(杰克) (him(Jack)) C_p:公司(the company)

Table 1 Example of different centers.

4 Proposed Approach

Due to the problems we introduced in Section 1, the techniques of target extraction in other domains are not appropriate in news comments, and general approaches encounter the problems of free style sentences and implicit targets. Fortunately, news comments have their own characteristics, which can be used to improve the target extraction performance.

One important characteristic is that though potential opinion targets may be in large quantities, most comments focus on several central concepts in the corresponding news article, especially in the title. So we can extract the focused concepts in the news and use them as potential implicit targets for the comments.

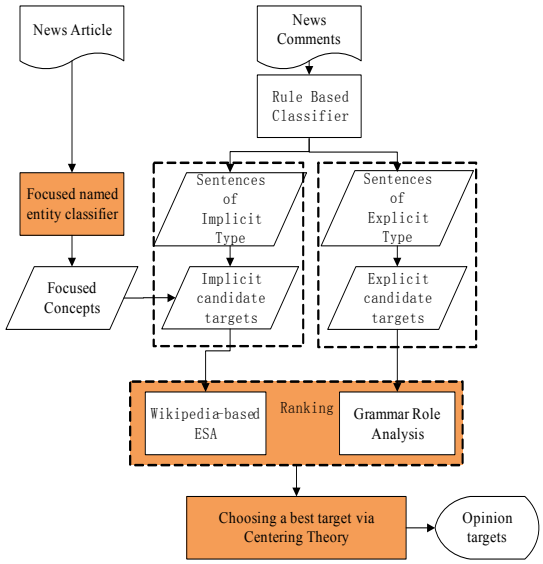

Figure 1: Framework of opinion target extraction in news comments

The other useful information comes from the fact that the sentences in one comment are usually coherent. As the comments may be long and each comment contains several sentences, the sentences within one comment are relevant and coherent. So the opinion targets in previous sentences have some influence on that in subsequent sentences. Using this kind of contextual information, we can eliminate noisy candidates and relax the dependence on an unreliable syntactic parser.

Considering the above characteristics, we propose a framework of target extraction based on focused concepts recognition and Centering Theory, as shown in Figure 1.

Given a news article and its relevant comments, we first adopt some syntactic rules to classify the comment sentences into implicit or explicit type. Whether a sentence includes an explicit target is mainly decided by whether it owns a subject. A few heuristic rules, such as the appearance of the subject, the combination of the POS, and the position of the predicate, are used based on the parse result by using a Chinese NLP toolkit[1], and the rule-based classification can attain an accuracy of 77.33%.

Then we exploit two different approaches for dealing with the two types of sentences, respectively. For the implicit type, we extract the focused concepts in the news article as candidate implicit targets, and rank them by calculating the semantic relatedness between the targets and the sentence. For the explicit type, all nouns and pronouns in the sentence are extracted as candidate targets and ranked mainly by their grammatical roles. At last, Centering Theory is used to choose the best candidate using the ranks and contextual information.

The details of the main parts are explained in the following sections.

4.1 Focused Concepts (FC) Recognition

As the comments usually point to the news article, it is highly probable that the implicit targets appear in the news article. Generally, the focused concepts of the news article are more likely to be the commented targets. Thus, if we extract the focused concepts of the news article, we will get the candidate implicit targets.

In general, the focused concepts are named entities (Zhang et al. 2004) or specific noun phrases. Taking the news

> "迪拜将建超千米全世界最高摩天大楼(*Dubai will build the highest skyscraper in the world*)" ----NEWS1

as an example, "迪拜(*Dubai*)" and "摩天大楼(*skyscraper*)" are the potential opinion targets. "*Dubai*" is a named entity, and "*skyscraper*" is a specific noun phrase. In addition, the focused concepts may also appear in the content of the news article, if they attract enough attention or have strong relations with the focused named entities in the title.

As the number of noun phrases is usually large, if we extract the two types of concepts together, there must be much noise to impact the final result. To be simple and accurate, we first extract focused named entities (FNE), and then expand them with other focused noun phrases, for the reason that the focused noun phrases usually have a strong relation with the focused named entities.

Entity Type	Person, Location, Organization, Time
Title	In title or not
Frequency	The number of occurrence
Relative Frequency	Frequency/the number of total words

[1] LTP, http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm
LTP is an integrated NLP toolkit which contains segmentation, parsing, semantic role labeling, and etc.

Distribution Entropy (Here we take N=5 according to the length of articles)	$Entropy = \sum_{i=1}^{N} p_i \log p_i$, where $p_i = \frac{Occurrence\ in\ the\ i^{th}\ Section}{Occurrence\ in\ Total}$

Table 2 Features of FNE classification

Extracting FNEs can be seen as a classification problem. In this work, we choose the features in Table 2.

Given a news document, we first recognize all named entities with our own named entity recognizer (NER).Then all named entities are classified based on the above mentioned features. The noun phrases in the title are also extracted and filtered by their frequency in the news article and co-occurrences with FNEs. The filtering threshold is set to a relatively high value to guarantee that not much noise is brought in. Thus we can get a small set of focused concepts in the news article.

4.2 Ranking Implicit Targets

We use the semantic relatedness to decide which potential target is most likely to be the right implicit target. There are many methods to calculate the semantic relatedness. We choose the Wikipedia-base explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007), for its adaptability and effectiveness for Chinese language. The method converts a word or a sentence to a series of wiki concepts, and then calculates the similarity between words or sentences.

Input: a Focused Concept t_0 in the news

Output: a vector C with a length of N. C= $<(c_j,w_j)>$, where c_j is a Wikipedia concept, and w_j is the weight of c_j

1. Find all nouns, adjectives and verbs co-occurring with t_0 in the same sentence, and put them into the set S= $\{t_i\}$.
2. Compute MI (Mutual information) of each t_i with t_0.
3. Choose 10 words in S with the highest MI (according to the total number of words, 10 is a proper value). Combine them with t_0 into a word vector and assign each word t_i a weight of its frequency v_i in the news article. The vector V= $<(t_i,v_i)>$, $|V|\leq 11$.
4. Let $<k_{ij}>$ be an inverted index entry for t_i, where k_{ij} quantifies the strength of association of t_i with Wikipedia concept c_j. Then the vector V can be interpreted as a vector constructed by All Wikipedia concepts. Each concept c_j has a weight $w_j = \sum_{t_i \in V} v_i k_{ij}$.
5. Select N concepts with the highest weights.

Table 3: Algorithm that converts a focused concept to a vector of Wikipedia concepts

Chinese Wikipedia is not as large as English Wikipedia. When some words are not included in the database, the original ESA algorithm will fail. To solve the problem, we first expand the input FC with a few words extracted from the news article. The words represent the semantic information related to the article, so they are more informative than a single concept while easily recognized by the Wikipedia database. The details of the algorithm are shown in Table 3.

On the other hand, when given a comment sentence, we segment it to words and remove the stop words (e.g. "的 (of)"). Then the serial of words are also converted by ESA into a vector of Wikipedia concepts.

After getting the vectors of wiki concepts for focused concepts and the comment sentence, we use the cosine metric to obtain their relatedness scores. In this way, the focused concepts are ranked by their relatedness scores with the sentence.

4.3 Ranking Explicit Targets

A comment sentence with explicit targets usually has a complete syntactic structure. According to Centering Theory, the ranks of explicit targets are decided mainly by their grammatical roles. Generally, a subject is most likely to be the opinion target, and the rank can be heuristically assigned by SUBJECT > OBJECT(S) > OTHER.

4.4 Choosing Best Candidate target via Centering Theory (CT)

After getting the candidate targets and their ranks, we start the matching step to make use of contextual information. The algorithm originates from the process of choosing preferred centers and backward-looking centers. A subtle adaption is that we add some global information in the news article as the context when dealing with the first sentence in a comment. The details of the algorithm are represented in Table 4.

Now we give an example to show the whole process of the framework. The following comment is associated with NEWS1 in Section 4.1.

U1:迪拜现在大力发展旅游和自由贸易。

(Dubai is developing travel and trades.)

U2:是一个很有活力的城市。

((It) is an active city.)

U3:在迪拜你可以感受到很多惊奇。

(In Dubai you can encounter many miracles.)

First, U1, U2 and U3 are classified as explicit, implicit and explicit, respectively. Then for U1 and U3 we choose noun phrases and pronouns in the sentence as candidate targets and rank them according to their grammatical roles. U2 chooses FC as candidates, and "Dubai" is more related than "skyscraper". At last, the final target is chosen by the algorithm in Table 4 and the whole process is illustrated in Table5.

Input: A comment with M sentences $S=\{s_i\}$, each sentence has a candidate target set $C_f(s_i)=\{c_i\}$;
The Focused Concepts set FC in the news article.

Output: A target set $\{t_i\}$, where each t_i is the opinion target of sentence s_i.

1. **For** each s_i in S
2. **If** $i=1$ (s_i is the first sentence)
3. **For** each c_i in $C_f(s_i)$
4. **If** c_i is contained in FC
5. Add c_i into the set $C_b(s_i)$
6. **If** $C_b(s_i)$ is not void
7. Choose the highest ranked element in $C_b(s_i)$ as t_i
8. **Else**
9. Choose the highest ranked element in $C_f(s_i)$ as t_i
10. **Else**
11. **For** each c_i in $C_f(s_i)$
12. **If** c_i realizes (equals or refers to) an element c'_i in $C_f(s_{i-1})$
13. Add c'_i into the set $C_b(s_i)$
14. **If** $C_b(s_i)$ is not void
15. Choose the highest ranked element in $C_b(s_i)$ as t_i
16. **Else**
17. Choose the highest ranked element in $C_f(s_i)$ as t_i

Table 4 Algorithm of choosing the best candidate target via CT

	type	ranks of candidates	target
U1	Explicit	迪拜>旅游>自由贸易 (Dubai >travel >trade)	迪拜 (Dubai)
U2	Implicit	迪拜>摩天大楼 (Dubai>skyscraper)	迪拜 (Dubai)
U3	Explicit	你>惊奇>迪拜 (you>miracles>Dubai)	迪拜 (Dubai)

Table 5 Example of the extraction process

5 Experiments

5.1 Evaluation Setup

To evaluate the whole system, we evaluate not only the result of the final target extraction but also some key steps. This makes the analysis of the bottleneck possible.

We first build a FNE dataset to evaluate the FNE classification result. As our target extraction task focuses on news comments, we collect 1000 news articles and the associated user comments from *http://comment.news.sohu.com,* which is a famous website offering a platform for users to comment on the news. Every news articles are annotated with its focused named entities, which are also the most possible commented targets.

Then we build the target dataset to evaluate the final target extraction. 9 articles and associated comments are randomly chosen from the FNE dataset, and each of their comment sentences is annotated with the opinion target. The target dataset focuses on 3 different topics: economics, technology and sports. Each document contains a news article and about 100 relevant comments, and there are 1597 comment sentences in total.

We assume that each comment sentence has one opinion target, but 108 sentences have more than one focused objects. In that case, we annotate all targets for evaluation and the result is regarded as true if we extract only one of the annotated targets.

In the target dataset, there are 444 sentences with implicit targets. This demonstrates that the implicit target extraction problem is prevalent and worth solving.

For the final target extraction, we use the accuracy metric to evaluate the result. It is defined as follows:

$$\text{Accuracy}=\frac{\text{Number of sentences with right extraction}}{\text{Number of total sentences}}$$

We do not use the precision and recall metric because every comment sentence in our dataset must have a target after extracting. The precision and the recall are both equal to the accuracy.

5.2 Evaluation Results

5.2.1 FNE Results

We perform a 4:1 cross-validation on the FNE dataset using a commonly used classifier SVM-light[2] and gain a mean f-measure of 80.43%.

Then, to assess the improvement by the FNE step and the classification of implicit and explicit sentences, we estimate the theoretic upper limit of the following three target extractions on the target dataset. **Test 1** assumes every noun phrases or nouns in the sentence can be possible to be extracted as the target. So if there is one candidate matching the target, we can recognize the sentence as extractable. **Test 2** adopts the annotation results of the classification of explicit and implicit sentences. For the manually annotated implicit targets, we adapt the candidate to be FC. Then, as same as Test 1, all candidates are determined whether to be the target. In **Test 3**, we follow the ruled-based classification of implicit and explicit sentences in our system and then judge the sentences whether extractable or not.

	Proportion of extractable sentences
Test 1	55.0%
Test 2	69.6%
Test 3	61.7%

Table 6 Improvement of the proportion of extractable sentences by FNE classification and explicit/implicit sentence classification

Table 6 shows the proportions of extractable sentences in the three tests. It is easy to see that the proportion of extractable sentences means the theoretic optimization of target extraction. So, by Test 2 we can see the extracted FC set is an effective complement of the candidate targets, while Test 3 demonstrates that the system still has much potential to improve the baseline after the rule-based classification of explicit and implicit sentences.

5.2.2 Target Extraction Results

To demonstrate the effectiveness of our approach, we design two baselines.

Baseline 1 treats all sentences as explicit type. In the method, we extract all noun phrases and pronouns in a sentence as candidates and obtain their ranks according to their grammatical roles.

Baseline 2, a SVM-based approach, is offered to compare with the popular target extraction methods. In this method we regard the target extraction as a classification problem. We extract the candidate noun phrases in a sentence first, and then use the semantic features to classify them as targets or not. The features mainly include: POS, whether or not a Named Entity, the positions in the sentence, the syntactic relations with the verb, and etc. As it is a supervised approach, the result is tested by a 2:1 cross validation.

Then we use a method called **FC-only (using only Focused Concepts)** to improve Baseline 1 by using the global information in news articles. For sentences of explicit type, we use the method in Baseline1. For sentences of implicit type, we take focused concepts in news articles as potential targets, and choose the highest ranked element as the final target.

Finally, our proposed approach **CT (using Centering Theory)** uses both Focused Concepts and Centering Theory. When the size of Wikipedia concept vector is set to be 800, the comparison results of the four approaches are shown in Table 7:

	Accuracy
Baseline1	34.38%
Baseline2(SVM-based)	35.13%
FC-only	37.25%
CT	**43.20%**

Table 7 Comparison results

FC-only is better than Baseline1, which demonstrates that the focused concepts are useful to provide information to implicit targets extraction. 444 implicit sentences are a large proportion of the total corpus. And the focused concepts do represent the global information and have influence on the target extraction.

Centering Theory is naturally another improvement. It mainly takes advantage of the information of contexts within a comment, using a rule of coherence to decide the center of attention. And the result indicates that it is very helpful.

Compared with the SVM-based approach, our approach is also much better. The SVM-based approach is only a little higher than Baseline 1. It seems that the manually annotated information is not very useful in target extraction in news comments. The reason may be that the target rules are complicated and exist not only in the current sentence. Using global and contextual

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

information is a more economic and effective way to improve the result.

In the Wikipedia-based ESA algorithm,there is a parameter of *N*, which is the vector size of the expanded vector. It is important to choose a proper parameter value to achieve a high accuracy and meanwhile keep a low computational complexity. The accuracy curves for FC and CT with different values of *N* are represented in Figure 2. Apparently, when *N* exceeds 600, the extraction performance almost does not change any more. So we finally take 800 as the value of *N* .

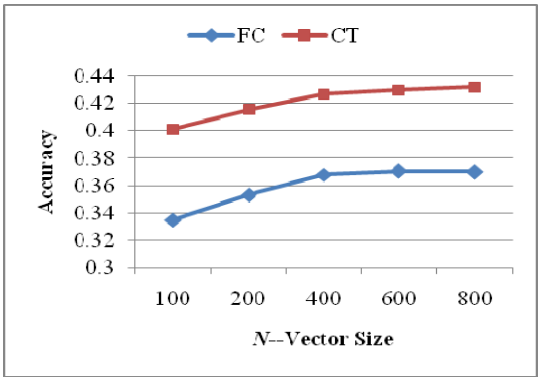

Figure 2: Accuracy vs. vector size *N*

5.3 Error Analysis

Generally there are two major types of errors in the extraction results. One common error is that the target is not in our extracted candidate nouns or noun phrases. For example:

> "可口买汇源，可真是中国饮料的灾难了." (It is a disaster of Chinese beverage that Coca Cola buys HuiYuan.)

The sentence comments on the event of "Coca Cola buys HuiYuan" but not a single concept "Coca Cola" or "HuiYuan". But our system cannot recognize this type of targets properly. Also there are some cases that the noun phrases missed to be extracted by the LTP toolkit. It causes that the target is not matched by the candidates.

Another error originates from the wrong classification of explicit and implicit sentences. For example,

> "还利于民才能化解中小企业生存危机." (Returning profits to civilians can get through the crisis of little companies.)

In this sentence, "还利于民(Returning profits to civilians)" is the opinion target and the sentence has a explicit target. But the rules based on the Chinese parser failed to recognize the phrase as a subject and thus the sentence is considered as implicit type by our approach. And lastly the target is extracted incorrectly.

In 5.2.1, we test the theoretic upper limit of the target extraction and prove the potential effectiveness of two steps. The tests also can be used to estimate the proportion of the types of errors and analyze the bottleneck. In Test 2, there are 298 un-extractable sentences among the annotated explicit sentences. It shows that there is at least 18.6% loss in accuracy caused by the candidate recognition, which accounts for the first error type. As for the second error type, its proportion can be computed by the reduction from Test 2 to Test 3, which is 7.9%.

6 Conclusion and Future Work

In this paper, we propose a novel approach to extracting opinion targets in Chinese news comments. In order to solve the problem of implicit target extraction, we extract focused concepts and rank their importance by computing the semantic relatedness with sentences via Wikipedia. In addition, we apply Centering Theory to the target extraction system, for utilizing contextual information. The experiment results demonstrate that our approach is effective.

Currently, the result does not reach an absolutely high accuracy. One bottleneck is that Chinese parsing results are far from satisfactory. Actually this bottleneck has impacted the general target extraction long, such as the low performances of all participants in the target extraction task of NTCIR7-MOAT-CS. We hope to improve our results by avoid this disadvantage. Moreover, the phenomenon of implicit opinion targets exists not only in Chinese but also in English and other languages, while sometimes it is similar to zero anaphora. So the approach in this paper can be extended to news comments in other languages.

Acknowledgement

This work was supported by NSFC (60873155), Beijing Nova Program (2008B03), NCET (NCET-08-0006) and National High-tech R&D Program (2008AA01Z421). We thank the anonymous reviewers for their useful comments.

References

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. *Automatic Extraction of Opinion Propositions and their Holders*. In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.

Choi, Yejin, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. *Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns*. In Proceeding of HLT/EMNLP' 05.

Ding Xiaowen, Bing Liu, Philip S. Yu. 2008. *A Holistic Lexicon based Approach to Opinion Mining*. Proceeding of the international conference on Web Search and Web Data Mining (WSDM'08), 231-239.

Du, Weifu. and Songbo Tan. 2009. *An Iterative Reinforcement Approach for Fine-Grained Opinion Mining*. The 2009 Annual Conference of the North American Chapter of the ACL

Gabrilovich, Evgeniy. and Shaul Markovitch. 2007. *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI).

Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. 2006. *Text Mining for Product Attribute Extraction*. The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Grosz, Barbara J., Scott Winstein, and Aravind K. Joshi. (1995). *Centering: A Framework for Modeling the Local Coherence of Discourse*. In Computational Linguistics, 21(2).

Hu, Minqing and Bing Liu. 2004. *Mining Opinion Features in Customer Reviews*. In Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI-2004)

Jin, Wei and Hung Hay Ho. 2009. *A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining*. In Proceedings of the 26th International Conference on Machine Learning (ICML 2009).

Jin, Wei and Hung Hay Ho, Rohini K. Srihari. 2009. *OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction*. In The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kim, Soo-Min. and Eduard Hovy. 2006. *Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text*. In ACL Workshop on Sentiment and Subjectivity in Text.

Kim, Soo-Min. and Eduard Hovy. 2005. *Identifying Opinion Holders for Question Answering in Opinion Texts*. In Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains.

Pang, Bo and Lillian Lee, and Vaithyanathan, S. 2002. *Thumbs up?* Sentiment *classification using machine learning techniques.* In EMNLP 2002.

Popescu, Ana-Maria. and Oren Etzioni. 2005. *Extracting Product Features and Opinions from Reviews*. In Proceeding of 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05), 339-346.

Riloff, Ellen and Janyce Wiebe. 2003. *Learning Extraction Patterns for Subjective Expressions*. Proceedings of the 2003 Conference on EMNLP.

Ruppenhofer, Josef, Swapna Somasundaran, and Janyce Wiebe. 2008. *Finding the Sources and Targets of Subjective Expressions*. In LREC08.

Seki, Yohei, David K. Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2008. *Overview of Multilingual Opinion Analysis Task at NTCIR-7*. The 7th NTCIR workshop (2007/2008).

Su Qi, Xinying Xu, Honglei Guo, Zhili Guo, XianWu, Xiaoxun Zhang, Bin Swen and Zhong Su. 2008. *Hidden Sentiment Association in Chinese WebOpinion Mining*. In The 17th International World Wide Web Conference (WWW).

Turney, Peter D. 2002. *Thumbs up or thumbs down? Semantic* orientation *applied to unsupervised classification of reviews.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

Zhang, Li, Yue Pan, and Tong Zhang. 2004. *Focused Named Entity Recognition using Machine Learning*. The 27th Annual International ACM SIGIR Conference.

Zhuang, Li, Feng Jing. and Xiao-yan Zhu. 2006. *Movie Review Mining and Summarization*. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06), 43-50.