Chinese Evaluative Information Analysis

Yiou Wang   Jun'ichi Kazama   Takuya Kawada   Kentaro Torisawa
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{wangyiou, kazama, tkawada, torisawa}@nict.go.jp

Abstract

Together with the ever-growing amount of Chinese web data, the number of opinions voiced by Chinese users is rapidly increasing, and analyzing them is an important task. This paper introduces a Chinese Evaluative Information Analyzer (CEIA) and proposes a method to improve its performance. We use *evaluative information* as a unifying term for the information about attitudes, opinions, sentiments and so on. This paper makes three contributions: (i) CEIA can identify and analyze a more diverse and richer set of evaluative information than previous studies for Chinese; (ii) to implement the system, we constructed an original annotated corpus for Chinese evaluative information and built a large sentiment dictionary; (iii) we introduce syntactic dependency, semantic class and distance features to improve the evaluative information extraction. The performance of the system and the effectiveness of the newly introduced features are evaluated in a series of experiments on our Chinese evaluative information corpus.

Title and Abstract in Chinese

中文评价信息分析

随着网络的不断普及，中文用户在网络平台上发表的评价性文本的数量也在迅速增长，因而分析这些日益膨胀的评价性信息则成为一项重要的课题。本论文介绍了一个中文评价信息分析系统（CEIA）并提出新的方法来提高系统的性能。在这里我们用评价信息来统一表达关于态度，观点和情感等的评价性相关信息。本文的贡献主要体现在以下三个方面：（i）与以往研究相比，CEIA可以识别和分析更加多种多样，更加丰富的评价信息；（ii）为了实现CEIA，我们标注了一个中文评价信息语料并且构筑了一个大规模的情感词典；（iii）我们引入了句法依存特征，语义聚类特征和距离特征来提高评价信息抽取的性能。通过在中文评价信息语料库上一系列实验，给出了系统的整体性能，并验证了新方法的有效性。

Keywords: Chinese Evaluative Information, Opinion Mining, Sentiment Analysis.

Keywords in Chinese: 中文评价信息，观点挖掘，情感分析.

Proceedings of COLING 2012: Technical Papers, pages 2773–2788,
COLING 2012, Mumbai, December 2012.

1 Introduction

To automatically find or track the attitudes, feelings and evaluations in texts, opinion mining and sentiment analysis have been extensively studied from different perspectives (Pang and Lee, 2008). With the ever-growing number of Chinese users (over half a billion users only in mainland China), the amount of web opinions in Chinese is rapidly increasing, and analyzing them is an important task. However, research and resources about the Chinese opinion analysis lag behind those for extensively studied languages, such as English. Therefore, opinion analyzers, which can deal with Chinese web data of a great variety of topics and styles, are especially in great need.

To meet this requirement, we introduce a Chinese Evaluative Information Analyzer (CEIA) that can mine a wide variety of evaluative information from Chinese web documents. We use *evaluative information* as a unifying term for the information concerning attitudes, opinions and sentiments, and so on, which is useful to provide a view of evaluation.

The system automatically analyzes Chinese evaluative information through the following processes: (1) *extracts evaluative expressions*; (2) *identifies evaluation holders*; (3) *extracts evaluation targets*; (4) *determinates evaluation types*; (5) *determinates the sentiment polarities of the evaluative expressions*.

CEIA has the following two characteristics:

Firstly, CEIA can analyze a more diverse and richer set of evaluative information than the previous studies for Chinese. The previous research on Chinese opinion analysis focuses on subjective expressions (*opinionated sentences*) (Liu, 2010), as in the Multilingual Opinion Analysis Task (MOAT) of NTCIR (Seki et al., 2010). However, some objective expressions that describe positive or negative facts are also informative in that they express some kinds of evaluations. Also, requests are some kinds of representations of opinions or attitudes. Consider the following sentences,

1. *Many people are using mobile phone A.*
2. *The users hope company A will offer them a security lock function.*

The sentence 1 suggests that "mobile phone A" is popular and has been chosen by many people. The sentence 2 claims that the company A does not offer a security lock function now and the user request the company to offer it. In some sense, this sentence also includes the evaluation or unsatisfied feelings of the users. We want to consider such cases as "implicit" evaluations for "mobile phone A" and "company A", in addition to subjective expressions such as "I love mobile phone A".

To the best of our knowledge, this is the first paper that treats the above implicit evaluations in Chinese evaluative information analysis. Implicit evaluations have been considered by Nakagawa et al. (2008) for Japanese. They presented the study about extracting subjective and objective Japanese evaluative expressions from the web and their work was used in WISDOM system (Akamine et al., 2010) [1] , and shown to be useful to support users' judgement of information credibility. Inspired by their work, we adopt the task definition and expand the research scope of Chinese evaluation information analysis.

Secondly, CEIA can deal with the data in diverse topics and writing styles. The existing studies about Chinese opinion analysis are domain-limited. For example, Chinese Opinion Analysis Evaluation (COAE) (Zhao et al., 2008) mainly deals with opinion analysis of reviews. MOAT (Seki et al.,

[1] http://wisdom-nict.jp/

2010) deals with the analysis of news articles, which are written in a formal writing style. To make our system more robust to the web data of a great variety of topics and styles, we constructed an original annotated Chinese evaluative information corpus whose sentences are extracted from web pages of wide range of topics and styles. CEIA consists of many machine learning modules such as CRFs and SVMs and the corpus was used to train these modules, resulting in a robust evaluative information analyzer.

To achieve high system performance is also a primal goal of evaluative information analysis. In this work, we introduce new features to improve the performance. Specifically, syntactic dependency features, semantic class features and distance features are added to the baseline models. To demonstrate the performance of our system and the effectiveness of our new features, we conducted a series of experiments on the Chinese evaluative information corpus.

2 CEIA

In this section, we describe the entire picture of CEIA and the resources for the system. We first introduce the specifications of the evaluative information on which this study is focused, and then we explain how an evaluative information corpus is constructed. Finally, we explain each process of CEIA in detail.

2.1 Evaluative Information

There is a wide variety of evaluative information on the web, such as reviews of products and criticisms of policies. The information reflects various perspectives of individuals or organizations. Research on evaluative information analysis are conducted from different points of views and at different levels of granularity (Kobayashi et al., 2004; Kaji and Kitsuregawa, 2006; Liu, 2010; Pang and Lee, 2008; Akamine et al., 2010). In this section, we describe the specifications of evaluative information on which this study is focused.

We analyze the evaluative information at a fine-grained level. We use a 5-tuple that consists of (1) an evaluative expression, (2) an evaluation holder, (3) an evaluation target, (4) an evaluation type, and (5) sentiment polarity as the basic unit of evaluative information and call it an *evaluative information set*. Each item is defined as follows.

Evaluative expression is a span of text that describes the evaluation. It can be a single word, a multi-word expression, or a sentence.

Evaluation holder is a person, a group or an organization that expresses the evaluation.

Evaluation target is a thing, a matter, or an entity about which the evaluation was expressed.

Evaluation type is the category to which the evaluative expression belongs. It will be explained in detail in the following subsection.

Sentiment polarity indicates whether the evaluation expression for the evaluation target is positive or negative from the viewpoint of the evaluation holder. For some cases, it may differ from the polarity of the whole sentence. For examples, *Mike strongly objected to the war*. Although the entire sentence is not negative, the sentiment polarity of evaluative expression "*strongly objected to*"is negative. That is to say, the evaluation holder "*Mike*" has a negative opinion on the evaluation target "*war*". From this point of view, we consider the sentiment polarity in the connection with specific evaluation holders and evaluation targets at fine-grained levels.

2.1.1 Evaluation Type and Sentiment Polarity

There are various kinds of evaluative expressions such as approving or opposing attitudes, description of merits or desirable events, and so on. To clarify the scope of evaluations that we address in this study,we classify evaluative expressions into several categories. Such categorization is also helpful for further use of evaluative information.

Following the work of Akamine et al. (2010), we use the following evaluation types. Each type, except for "Request", has sentiment polarities: positive (+) or negative (−). We use underline to show evaluation targets, boldface for evaluative expressions, and italics for the evaluation holders.

- Emotion+/−: an expression that expresses human feelings or emotions.
  e.g., *XiaoLi* **is not interested in** product A. (Emotion−)
- Comment+/−: an expression that expresses approval/disapproval or praise/criticism.
  e.g., *Mike* said that movie A **is one of the best he has ever seen**. (Comment+)
- Merit+/−: an expression that cites good points/shortcomings or merits/demerits.
  e.g., Drug A **starves and kills cancer cells**. (Merit+)
- Event+/−: an expression that describes good/bad events, desirable/undesirable experience.
  e.g., Camera X **broke just three days after I bought it**. (Event−)
- Adoption+/−: an expression that shows adoption, promotion or rejection.
  e.g., **Nobody bought** Mike's ebook. (Adoption−)
- Request: an expression that expresses proposals, obligations, advices, hopes or requests.
  e.g., *The users* **hope Company A can offer them a security lock function**. (Request)

2.2 Chinese Evaluative Information Corpus

To train our system and analyze a wide variety of evaluative information, we constructed an evaluative information corpus which consists of Chinese sentences extracted from web pages of wide range of topics and styles. We chose 66 topics which relate to things we use in daily life, controversial policies, movie reviews and so on. The followings are the steps for the corpus construction:

(1) Use the topic as the keyword and search documents using a Web search engine.

(2) Collect HTML files of 900 web pages from the retrieval results for each topic. Specifically, the first 300 pages in the retrieval results from forum sites, the first 300 pages from blogs and the first 300 pages from general sites.[2] In this way, the corpus can cover different writing styles and reflect more diverse perspectives.

(3) Randomly choose candidate sentences that include topic keywords from the above files. For each topic, we randomly collected 200 sentences, and for each sentence, we provided context information (the previous two sentences and subsequent two sentences) for annotation reference.

(4) Trained annotators judged whether a sentence contained any evaluative expressions or not. If the sentence contained evaluative expressions, the annotator annotated the evaluation holders, the text spans of the evaluative expressions, the text spans of the evaluation targets, the evaluation types and the sentiment polarities. That is to say, an evaluative information set was annotated for each evaluation expression. For evaluation holder annotation, if the writer is the evaluation holder, [*author*] is annotated as the holder. If the holder is neither explicitly written in the sentence

[2] We suppose the URL including "forum", as the web pages from forum sites, the URL including "blog", as the web pages from blog sites, and the rest are general sites, although it may include some noise.

Dictionary Origin	Positive	Negative	Postive + Negative
JSD	6,270	19,394	25,664
Giga-word	1,977	770	2,747
Total	8,247	20,164	28,411

Table 1: The statistics of sentiment dictionary

nor is the writer, [*undefined*] is annotated. When annotating the current sentences, the annotator could refer to its context information. In some cases, one sentence may contain multiple evaluative targets or multiple evaluation expressions, and then multiple evaluative information sets must be annotated. For example, *Mike said that Movie A is great but it is not better than Movie B which is the best movie he has seen.* Two evaluative information sets should be annotated: (1) (is great, Mike, Movie A, Comment+) and (2) (is the best movie he has seen, Mike, Movie B, Comment+). Note that *it (Movie A) is not better than Movie B* is a comparative expression. we do not deal with the comparative sentences that do not show clear sentiment polarities at present.

The total number of sentences in the corpus was 6,680. There were 5,111 evaluative information sets in the corpus. It took 380 man-hours to construct the entire corpus.

2.3 Sentiment Dictionary

A sentiment dictionary is a set of words and their polarities (for example, [break a record, +], [break the law, −]). Such a dictionary is a fundamental resource for evaluative information analysis. We built a Chinese sentiment dictionary in the following way.

(1) Since it is time-consuming to built a dictionary without any reference, we semi-automatically translated an existing Japanese sentiment dictionary (JSD)[3] to Chinese. We mapped the entries of JSD with a Japanese-Chinese bilingual dictionary, and obtained Chinese translations and their polarities transferred from Japanese entries. Unmapped entries were translated by human. The resulting Chinese entries and polarities were finally manually checked. There were 36,981 entries in JSD (9,030 positives and 27,951 negatives) , and we obtained 25,664 entries for Chinese.

(2) So that the dictionary covers the frequently used polarity-bearing words, we also auto-segmented and tagged the XIN_CMN portion of Chinese Gigaword Version 2.0 (LDC2009T14), which has approximately 311 million words, and collected adjectives (with POS tags "VA" and "JJ") and idiom candidates with high frequency. We removed the overlap between the words collected from JSD, and manually checked the rest of the candidates, and tagged them with polarity.

Finally, we build a Chinese sentiment dictionary with 28,411 entries. Its detailed statistics are shown in Table 1. It is used in evaluative expression extraction and polarity classification models.

2.4 CEIA System

CEIA flow is shown in Figure 1. First, the user inputs raw sentences; second, the system (1) extracts the evaluative expression from the input sentences, (2) identifies the evaluation holder, (3) extracts the evaluation target, (4) categorizes the evaluation type and (5) determinates the sentiment polarity. Finally the results from these processes are summarized and displayed as output to user. The rest of this section describes the above processes in detail.

[3] The dictionary is distrubuted only to the member of the ALAGIN forum (http://alaginrc.nict.go.jp). It is for the freely available package of opinion extraction tool, which can be obtained from http://alaginrc.nict.go.jp/opinion/index_e.html

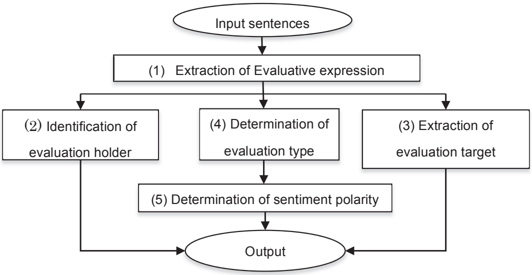
Input sentences
(1) Extraction of Evaluative expression
(2) Identification of evaluation holder
(4) Determination of evaluation type
(3) Extraction of evaluation target
(5) Determination of sentiment polarity
Output

Figure 1: CEIA flow

Type	Feature	Description
Word feature	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$, $w_{i-1}\&w_i, w_i\&w_{i+1}$	Word surfaces of the previous but one, previous, next, and next but one words; word surface bigram of the previous (next) word and the current word.
POS tag feature	$t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}$, $t_{i-1}\&t_i, t_i\&t_{i+1}$	POS tags of the previous but one, previous, next, and next but one words; POS tag bigram of the previous (next) word and the current word.
Polarity feature	$p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2}$, $p_{i-1}\&p_i, p_i\&p_{i+1}$	The word polarities of the previous but one, previous, next, and next but one words; word polarity bigram of the previous (next) word and the current word.

Table 2: Feature templates for evaluative expression extraction

2.4.1 Extraction of Evaluative Expressions

The goal of this process is to identify the words, phrases or sentences that express the evaluations in the text. We use the sequence tagging method with the BIO tag-set, which was initially used for opinion extraction by Breck et al. (2007). In the method, each word is tagged with one of three types of labels based on its position in the evaluative expressions: (B) beginning of an evaluation expression, (I) inside of an evaluation expression or (O) outside of an evaluation expression. For example, for the sentence, "*The chief editor* **really loves** book A." , the BIO tags are encoded in the following way:

The/O chief/O editor/O **really/B loves/I** book/O A/O ./O

We employ the linear chain CRFs (Lafferty et al., 2001) as our learning model for BIO tagging. Specifically, we use CRF++ (version 0.54) implementation by Taku Kudo. [4]

The features shown in Table 2 are used in the CRF for the i-th word in a sentence. Here, w_i, t_i, and p_i denote the current word surface, the part-of-speech tag and the polarity of the i-th word in the input sentence, respectively. A word's polarity is obtained from the sentiment dictionary. To search the word in the dictionary, we use forward maximum matching. We generate the above features with the unigram template of CRF++ (i.e., as the combination with the output tag at the current position, o_i). We also use the tag bigram feature (i.e., $o_{i-1}\&o_i$) .

2.4.2 Extraction of Evaluation Targets

The evaluation target is extracted from a sentence that contains the evaluative expression with a BIO tagging method using a CRF, as in the extraction of evaluative expressions. We use the same word feature and POS tag feature as in evaluative expression extraction and introduce position

[4] http://crfpp.sourceforge.net/

Type	Feature	Description
unigram p	$w_1, t_1, w_2, t_2, ... w_s, t_s,$	For the words previous to the evaluation expressions, the word and POS tag unigrams are added as type-p unigram features
unigram x	$w_{s+1}, t_{s+1}, ... w_{s+n}, t_{s+n},$	For the words in the evaluation expressions, the word and POS tag unigrams are added as type-x unigram features
unigram n	$w_{s+n+1}, t_{s+n+1}, w_{s+n+2}, t_{s+n+2}, ... w_l, t_l,$	For the words next to the evaluation expressions, the word and POS tag unigrams are added as type-n features
bigram	$w_{s+1}\&w_{s+2}, t_{s+1}\&t_{s+2}, ... w_{s+n-1}\&w_{s+n}, t_{s+n-1}\&t_{s+n}$	For the words in the evaluation expressions, the word bigram and POS tag bigram features are added
category	$c_i\&w_{i-1}, c_i\&w_i, c_i\&w_{i+1}, c_j\&w_{j-1}, c_j\&w_j, c_j\&w_{j+1}..$	For the words in the evaluative expressions, the category and word bigram feature are added: the category and the previous word bigram, the category and current word bigram and the category and the next word bigram

Table 3: Feature templates for evaluation type determination

features. The position feature setting is $\{e_{i-2}, e_{i-1}, e_i, e_{i+1}, e_{i+2}, e_{i-1}\&e_i, e_i\&e_{i+1}\}$. Here, e_i is a flag that expresses the position of w_i with respect to the evaluative expression. If w_i is previous to an evaluative expression, then e_i is "p"; if w_i is in an evaluative expression, then e_i is "x"; and if w_i is next to an evaluative expression , then e_i is "n". For example, for the sentence, " *The chief editor* **really loves** book A." , the e_i is encoded in the following way. If no holder was found by the CRF model, [*undefined*] was set as the evaluation target of the current evaluation expression.

The/p chief/p editor/p really/x loves/x book/n A/n ./n

2.4.3 Determination of Evaluation Types

We predicted the evaluation types using one-versus-rest multi-class linear kernel support vector machines (SVMs). We used the features shown in Table 3 for SVMs. Here w_i, t_i and c_i denote the word surface, the part-of-speech tag and the type category of the i-th word, respectively. l, n and s denote the number of words in the input sentence, the number of words in the evaluative expression and the number of words previous to the evaluative expression in the input sentence, respectively.

A word's type category is obtained from a type category dictionary. In the investigation of evaluation types, which has been described in Section 2.1.1, we found that each evaluation type has some characteristic words. Therefore we manually listed such characteristic words for each evaluation type and generated a type category dictionary, which includes 141 entries, for example, [希望(hope), Request], [憎恨(hate), Emotion], [称赞(praise), Comment]. For words in the evaluative expressions, the category feature can be generated only when the word is in the type category dictionary. The category feature can provide some improvement in performance according to our preliminary experiments.

2.4.4 Identification of Evaluation Holders

While the evaluation holders are sometimes stated explicitly in sentences where the evaluative expressions are contained, in more than half of the cases in Chinese, they are not clearly stated in the sentence. When evaluation holders are not expressed in the sentence, the evaluation holder is usually the information sender, i.e., the "author" in other words. Therefore, we consider that the opinion holder identification consists of a classification task and an information extraction task. That is, the evaluation holder is identified in two steps in CEIA: (1) use linear kernel support vector machines (SVMs) to determine whether the evaluation holder is *author* or *not-author*; (2) if the evaluation holder is not the author, then use a CRF tagging

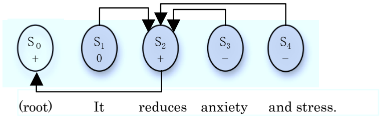

Figure 2: Example of CRFs with hidden variables

model to extract the evaluation holder for each evaluative expression. For the SVM model, in addition to the same features in types of unigram p, unigram x, unigram n and bigram as in Section 2.4.3, we also use the bigram p $\{w_1\&w_2, t_1\&t_2...w_{s-1}\&w_s, t_{s-1}\&t_s\}$ and bigram n $\{w_{s+1+1}\&w_{s+n+2}, t_{s+n+1}\&t_{s+n+2}...w_{l-1}\&w_l, t_{l-1}\&t_l\}$ features to add the bigram information for the words previous to the evaluation expressions and the words next to the evaluative expressions. For CRF model, we use the same features as in Section 2.4.2. If no holder was found by the CRF model, [*undefined*] was set as the evaluation holder of the current evaluation expression.

2.4.5 Determination of Sentiment Polarity

A typical approach for sentiment classification is to use supervised machine learning algorithms with bag-of-words as features (Pang et al., 2002). However, this method cannot consider syntactic structures that seem essential to infer the polarity of a whole sentence. We follow the work of Nakagawa et al. (2010) and use a dependency tree-based method, which was demonstrated to perform better than other methods based on bag-of-words in both English and Japanese sentiment classification tasks. The sentiment polarity is classified using conditional random fields (CRFs) with hidden variables. In the method, the sentiment polarity of each dependency subtree, which is not observable in training data, is represented by a hidden variable. The polarity of the whole sentence is calculated by considering the interactions between the hidden variables. For example in Figure 2, each phrase (indicated by a circle) in the polarity-bearing sentence/expression has a random variables. The random variable represents the polarity of the dependency subtree whose root node is the corresponding phrase. Two random variables are dependent if their corresponding subtrees have head-dependent relations (indicated by an arc). Usually the polarity is labeled in expression/sentence level in the annotated corpus, and subtrees are not labeled, so all the random variables except for the root node are hidden variables (indicated by gray circles). In the model, if a head word tend to reverse the polarity of the dependent word, reversal polarity feature can be used. That is to say, it can deal with the reversal of sentiment polarities caused by polarity shifting words. For example, the "reduce" in the example is polarity shifting word. "Reduce anxiety" is positive, while "anxiety" is negative. In order to deal with the polarity shifting, 179 Chinese polarity shifting words were collected and used in the CEIA. As for the features, we used the same features as those in Nakagawa et al. (2010).

3 New Features

In this section, we describe our approach that effectively employs the dependency information, semantic class and distance information into the above evaluative information extraction (specifically evaluative expression extraction and evaluation target extraction).

3.1 Dependency Features

The use of syntactic or deep linguistic features has been tried in opinion analysis in the literature. Johansson and Moschitti (2010) demonstrated that the features derived from grammatical and

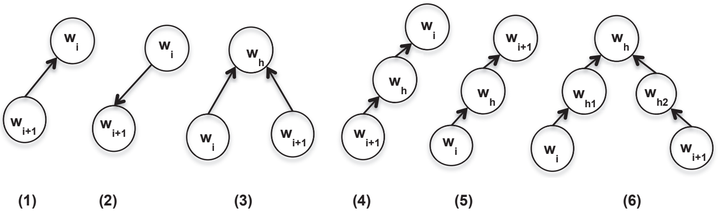

Figure 3: Different dependencies between w_i and w_{i+1} that can be linked by one or two arcs

semantic role structure can be used to improve the detection of opinionated expressions in subjectivity analysis. However, based on their evaluation, the precision decreases while the F-measure is increased. In addition, they claimed that a sequence tagging model cannot be used when using syntactic features, and they used reranking method, which will slowdown the processing. We introduce a simple dependency features for our tagging model that can be generated with the help of a Chinese dependency parser for evaluative information extraction.

Using a dependency parser, two kinds of dependency information can be obtained:

(i) *head* : the head of the current word, which is either a value of word ID, or zero ('0') if the word is the root node of the sentence.
(ii) *dependency relation*: the dependency relation of the current word to the head. The dependency relation is presented by the dependency labels: SBJ, OBJ, PRD, NMOD, VMOD, etc. The labels show function categories, such as the subject, object, predicate and so on.

We introduce the following two kinds of dependency features:

(i) *dependency head feature*: this feature is generated from the head information. The head-dependent relations between neighboring words w_i and w_{i+1} that can be linked by one or two arcs or can be linked to the same head by the same number (one or two) of arcs are summarized in Figure 3. We encoded the head-dependent relation into a new type of feature. We tried several feature representations and found that the features derived from the following method were most effective. We categorized the head-dependent relation between w_i and w_{i+1} into four groups:

- Near head-dependent relation (NH): the cases of (1), (2) and (3) in Figure 3.
- Medium head-dependent relation (MH): the cases of (4), (5) and (6) in Figure 3.
- Last word (LW): if w_i is the last word of the sentence/expression.
- Far head-dependent relation (FH): all the possible dependencies except for the above three groups.

The new features of $deph_i$ and $w_i\&deph_i$ are added for the i-th word in a sentence. Here $deph_i$ is head-dependent relation group of w_i and w_{i+1} , labeled with NH, MH, LW or FH. We suppose that such labels encode the syntactic distance information. For example, although w_i and w_{i+1} is the neighborhood in a sentence, they are distant syntactically, if the head-dependent relation group is labeled with FH.

(ii) *dependency relation feature*: this feature is generated with the information of the dependency relation. The dependency relation feature setting for evaluative expression extraction is

$\{depr_i, t_i\&depr_i\}$. Here, $depr_i$ is the dependency label of the relation between w_i and w_i's head in a sentence. Since the grammatical information is very important for evaluation target extraction, new features of $\{ depr_{i-2}$, $depr_{i-1}$, $depr_i$, $depr_{i+1}$, $depr_{i+2}$, $depr_{i-1}\&depr_i$, $depr_i\&depr_{i+1}\}$ are added for evaluation target extraction. With these features, the grammatical function information can be encoded in both the evaluative expression and evaluation target extraction tasks.

3.2 Semantic Class Features

The idea of combining semantic classes of words with discriminative learning has been previously reported in the context of named entity recognition (Miller et al., 2004; Kazama and Torisawa, 2008), dependency parsing (Koo et al., 2008) and Chinese word segmentation and POS tagging (Wang et al., 2011). We adopt and extend these techniques to evaluative information analysis and demonstrate their effectiveness in this task.

We produced the semantic classes of various levels of granularity, by using the Brown cluster hierarchy (Brown et al., 1992) at various lengths. Note that a semantic class is represented by a bit string that reflects the branching of the semantic class hierarchy.

We designed two kinds of semantic class features:
(i) full string feature: full string of the semantic class for w_i;
(ii) 6-bit prefix feature: 6-bit prefix of the semantic class for w_i.

3.3 Distance Feature

The target extraction task is to extract a target for a given evaluative expression. In most cases, the evaluation target and the evaluative expression are near to each other. Therefore, we add the distance label between w_i and the evaluative expression as a new feature for evaluation target extraction. The distance labels are defined in the following way: we first compute the distance d between w_i and the evaluative expression in word count; then when d is larger than 10, the distance label is "L"; otherwise if w_i is on the lefthand side of the evaluative expression, the distance label is d; and if w_i is on the righthand side of the evaluative expression, the distance label is $-d$. The feature setting of distance feature is $\{dis_{i-2}, dis_{i-1}, dis_i, dis_{i+1}, dis_{i+2}, dis_{i-1}\&dis_i, dis_i\&dis_{i+1}\}$. Here, dis_i is the distance label of w_i . With these feature, the position information with regard to the evaluative expression can be encoded.

4 Experiments

We evaluated the performance of the CEIA system and the effect of the new features.

4.1 Experimental Setting

We used the Chinese evaluative information corpus described in Section 2.2 as the training and test sets and performed 10-fold cross validation experiments on the corpus.

To conduct the experiments, we used the Chinese morphological analyzer described in Wang et al. (2011) and a Chinese dependency parser (CNP) [5] to obtain the Chinese word segmentation, part-of-speech tags and dependency information.

To generate the semantic classes of words, we used the XIN_CMN portion of Chinese Gigaword Version 2.0 (LDC2009T14), which has approximately 311 million words, as a large raw data and set the number of classes to 1000.

[5] http://alaginrc.nict.go.jp/cnp/index.html

We use the following measures to evaluate the performance of the system:

Recall (R) : ratio of correctly extracted evaluative expressions/targets/holders to the number of expressions/targets/holders in the gold standard corpus.
Precision (P): ratio of correctly extracted evaluative expressions/targets/holders to the number of expressions/targets/holders in system's output.
F-measure (F): harmonic mean of recall and precision.
Accuracy (Acc): ratio of the number of correct system output to the number in the gold standard. The accuracy of each tasks is defined as follows:

Accuracy of evaluation type determination: ratio of correctly identified evaluation types to the number of evaluative expressions in the gold standard corpus.
Accuracy of evaluation polarity determination: ratio of correctly classified sentiment polarities to the number of evaluative expressions of polarity-bearing evaluation types in the gold standard corpus.

To calculate the recalls, precisions and F-measures of the evaluative expressions and the evaluation targets, we use the following three criteria:

Exact match: extracted expression/target/holder is regarded as correct if it exactly matches the gold standard.

Partial match: extracted expression/target/holder is regarded as correct if it overlaps the gold standard's one. Our partial match is different from the overlap-based precision and recall measures in Breck et al. (2007). A potential issue with their overlap-based precision and recall is that the measures may drastically overestimate the system's performance as follows: a system predicting the whole sentence as an extracted expression would achieve 100% overlap-based recall and precision, if the gold standard contains any evaluative expression. In order to avoid this problem, we deal with the duplicate matches as follows: an extracted expression is only counted as overlapping with the first gold standard one, even if it can overlap with more than one gold standard's ones. From this point of view, our metric is stricter than in Breck et al. (2007).

Span partial match: this evaluation metric takes the span coverage of extracted expression/target/holder with respect to the span of the gold standard's one into consideration. We define this metric by refining the soft precision and recall described in Johansson and Moschitti (2010). First the span coverage c of a span s with respect to another span s', which measures how well s' is covered by s , was defined: $c(s,s') = \dfrac{|s \cap s'|}{|s'|}$. In this formula, the operator $| * |$ counts tokens (Chinese characters), and the intersection $\cap$ represents the overlap of the two spans. Then, if two spans overlapped, instead of adding "1" to the number of correctly extracted expression as in partial match, we add the span coverage to the number of correctly extracted expression. For example, if the gold standard evaluative expression had 8 tokens and 6 tokens of extracted expression overlapped with the gold standard, then we consider $3/4$ of the expression is correctly extracted. We deal with the duplicated matches in the same way as in partial match to avoid the overestimation. Although Johansson and Moschitti (2010) tried to alleviate the overestimation problem with their soft precision and recall, their measure still tend to reward long spans in recall [6] and overestimate the precision in some cases. Our metrics solved both the overestimation in recall and precision. Our metric is bounded below the exact match and above the partial match.

[6] a system predicting the whole sentence as an extracted expression would achieve 100% soft recall in Johansson and Moschitti (2010)

Task	Exact match	Partial match	Span partial match

Evaluative expression extraction	R=0.1730
	P=0.2933
	F=0.2176
		R=0.4560
		P=0.7728
		F=0.5734
			R=0.3934
			P=0.6264
			F=0.4832

Evaluation target extraction	R=0.4171
	P=0.6530
	F=0.5089
		R=0.5442
		P=0.8521
		F=0.6640
			R=0.5226
			P=0.7934
			F=0.6300

Evaluation holder identification	R=0.7455
	P=0.9630
	F=0.8401
		R=0.7518
		P=0.9714
		F=0.8474
			R=0.7509
			P=0.8672
			F=0.8047

Evaluation type determination	Acc = 0.5787	-	-
Evaluation polarity determination	Acc = 0.8146	-	-

Table 4: The performance of CEIA

4.2 Performance of CEIA

The performance of the entire CEIA system is shown in Table 4. The figures are for the best combination of the features, which will be described later. The performance of each task was evaluated independently. For example, for sentiment polarity determination task that determine the polarity of the evaluative expressions, the input evaluative expressions are the gold standard ones rather than the system output of the evaluative expression extraction task. For the evaluative expression, the performance of the exact match seems to be low. This is because it is difficult to detect the exact span of an evaluative expression. The evaluative expression detection in English also came to such situation and most work use partial match measures (Johansson and Moschitti, 2010). The performance of our system for partial match is reasonably good. Although the recall was not high, to extract information from a large amount of raw data, such as billions of web documents, we believe that the precision is a very important metric. The precision of the evaluative information extraction is 0.77. With such a relatively high precision, we suppose the evaluative expression extraction can play an active role in the actual application.

We also compared our system with the other works or systems reported in the literature, which are in the close task definition, although it is not fair to compare directly, because we deal with different languages and use different test sets. We just use their work as a reference to show that our system provided a reasonable result, when dealing with the same task in different language contexts.

As for sentiment polarity determination, we follow the work of Nakagawa et al. (2010). Their method was shown to perform better than other methods based on bag-of-words and provided accuracies ranging from 0.861 to 0.773 for a series of Japanese and English test sets. Because our test sets include various topics, this complicates the polarity classification task. Since our classification accuracy was 0.8146, we can say that Nakagawa et al. (2010)'s model also works well for Chinese.

As for the evaluative expression extraction, as we mentioned in the Section 1, Nakagawa et al. (2008) extracted subjective and objective Japanese evaluative expressions from the web. The result of their system with exact match is shown in Table5 . The performance scores are directly taken from their paper. The result indicates the difficulty of this task. The performance of our system is better than their work.

As for evaluation target extraction and evaluation holder identification, Multilingual Opinion Analysis Task (MOAT) of NTCIR-8 (Seki et al., 2010) included these tasks. Table 5 shows the best results with lenient match in opinion holder and opinion target identification tasks of simplified

work	Nakagawa et al. (2008)	MOAT of NTCIR-8	
Task	Evaluative Expression Extraction	Target Identification	Holder Identification
Recall	0.12	0.564	0.792
Precision	0.22	0.735	0.877
F	0.15	0.638	0.832

Table 5: Performance of previous works

Method	Exact match			Partial match			Span partial match		
Measure	R	P	F	R	P	F	R	P	F
Baseline	0.1628	0.3005	0.2110	0.4155	0.7678	0.5388	0.3599	0.6229	0.4557
Baseline+class (6-bit prefix)	0.1715	0.3018	0.2186	0.4386	0.7718	0.5593	0.3786	**0.6289**	0.4724
Baseline+class (full string)	**0.1746**	0.3005	**0.2208**	0.4474	0.7696	0.5655	0.3874	0.6257	0.4784
Baseline+dependency head	0.1688	0.3010	0.2162	0.4319	0.7695	0.5531	0.3735	0.6278	0.4681
Baseline+dependency relation	0.1686	**0.3036**	0.2167	0.4276	0.7699	0.5496	0.3709	0.6281	0.4660
Baseline+all features	0.1730	0.2933	0.2176	**0.4560**	**0.7728**	**0.5734**	**0.3934**	0.6264	**0.4832**

Table 6: Performance of new features in evaluative expression extraction

Chinese in MOAT. The results are directly taken from Seki et al. (2010). Although we use different test set and cannot compare the results directly, we can conclude that our system's F-measure is competitive with the systems that deal with a similar task.

4.3 Effect of New Features

We added the new features described in Section 3 to the evaluative expression extraction and target extraction models and performed 10-fold cross validation experiments to evaluate their effectiveness. We also tested the new features for evaluation holder extraction. However we omit the results here because the improvement by the new features was slight.

Table 6 shows the performance of the new feature in the evaluative expression extraction. Here, "all features" is the result of the combination of all the features. As mentioned in Section 4.2, to exactly identify the span of the evaluation is very difficult. Thus, we mainly refer to the results measured by partial match and span partial match here. Dependency features achieved an improvement in both recall and precision. The dependency features that introduced by Johansson and Moschitti (2010) only showed positive effect on recall with their soft partial match measure and partial match. Our span partial match and partial match are stricter measures than theirs. Note that we also evaluated our dependency head and dependency relation features use their soft precision and recall. There was no decrease in both soft precision and recall. In this point, our dependency features was comparably effective. Furthermore, our method uses the dependency features in sequence tagging model and is simpler than their method. The results also show that the full string semantic class features were the most effective ones and that a combination of four types of features achieves the best performance in F-measure. This suggests that these features are relatively independent in feature characteristics.

Table 7 shows the performance of the new features in the evaluation target extraction. The results show that the semantic class feature shows less effect in target extraction task than in evaluative expression extraction task and distance feature were the most effective one. 6-bit prefix features achieved an improvement in partial match. While the dependency head feature did not show a positive effect on the recall , it achieved best results on precision. Dependency relation features and distance features had positive effect for both recall and precision. The combination of all

Method	Exact match			Partial match			Span partial match		
Measure	R	P	F	R	P	F	R	P	F
Baseline	0.4040	0.6643	0.5021	0.5143	0.8454	0.6391	0.4942	0.7960	0.6094
Baseline+class (6-bit prefix)	0.4026	0.6595	0.5000	0.5186	0.8501	0.6440	0.4977	0.7983	0.6129
Baseline+class (full string)	0.4058	0.6546	0.5008	0.5267	0.8494	0.6495	0.5051	0.7933	0.6169
Baseline+dependency head	0.4039	**0.6753**	0.5052	0.5135	0.8590	0.6425	0.4930	**0.8083**	0.6122
Baseline+dependency relation	0.4074	0.6710	0.5069	0.5219	0.8594	0.6491	0.5006	0.8072	0.6177
Baseline+distance	0.4135	0.6717	**0.5117**	0.5290	**0.8597**	0.6548	0.5073	0.8081	0.6232
Baseline+all features	**0.4171**	0.6530	0.5089	**0.5442**	0.8521	**0.6640**	**0.5226**	0.7934	**0.6300**

Table 7: Performance of new features in evaluation target extraction

features can provide best result in recall and F-measure in partial match.

5 Related Work

Some previous research extracted evaluative or polarity-bearing expressions from web documents with pre-defined linguistic patterns (Kobayashi et al., 2004; Kaji and Kitsuregawa, 2006). However, it is difficult to prepare a small number of fixed syntactic patterns to extract a wide range of evaluative expressions. Nakagawa et al. (2008) presented the study about extracting Japanese evaluative expressions from the web. Our task definition is based on their work. We applied these tasks to Chinese, made a Chinese corpus and presented our new features to improve the performance of evaluative information extraction.

In recent years, there have been several opinion-related evaluation workshops concerning Chinese opinion mining, such as Chinese Opinion Analysis Evaluation (COAE) (Zhao et al., 2008) and the Multilingual Opinion Analysis Task (MOAT) of NTCIR (Seki et al., 2010). Several subtasks are conducted in both COAE and MOAT, including the opinion-bearing sentence detection, opinion target extraction and polarity determination. The opinion target extraction task in COAE identifies the product features, which are defined as product components or attributes. Compared with COAE, the evaluation targets extracted by our system can cover a wider scope; they can be nouns, multi-word expressions or nouns modified by clauses. At the same time, we considered evaluation holders in this research. Since opinion expressers influence the credibility, identifying the evaluative holders is very important for analyzing the evaluations. MOAT also includes the opinion target and opinion holder extraction tasks. Compared with MOAT, we introduce evaluation types and extend the coverage of the opinion mining targets. Explicit and implicit opinions, and subjective and objective evaluations are considered in our research, while MOAT only considers the opinionated sentences, not including the general facts, such as positive or negative facts. Furthermore, COAE mainly deals with opinion analysis in reviews, and MOAT deals with the opinion analysis in news, which are written in a more formal writing styles. Since our system was trained with a corpus, which is written in more diverse writing styles and covers wide domains, we believe it is more robust to the web data of a great variety of topics and styles.

Conclusion

In this paper, we presented a Chinese evaluative information analysis system and proposed new simple yet effective features to improve its performance. Through a series of experiments, we demonstrated that our system can achieve reasonably good performance and that our new features provides substantial improvement in evaluative expression extraction and evaluation target extraction tasks.

References

Akamine, S., Kawahara, D., Kato, Y., Nakagawa, T., Leon-Suematsu, Y. I., Kawada, T., Inui, K., Kurohashi, S., and Kidawara, Y. (2010). Organizing information on the web to support user judgments on information credibility. In *Proceedings of the 4th International Universal Communication Symposium (IUCS2010)*, pages 122–129.

Breck, E., Choi, Y., and Cardie, C. (2007). Identifying expressions of opinion in context. In *Proceedings-IJCAI-2007*, pages 2683–2688.

Brown, P., Pietra, V. D., de Souza, P., Lai, J., and R.L.Mercer (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Johansson, R. and Moschitti, A. (2010). Syntactic and semantic structure for opinion expression detection. In *In Proceedings of ACL-2010*, pages 67–76.

Kaji, N. and Kitsuregawa, M. (2006). Automatic construction of polarity-tagged corpus from html document. In *Proceedings-COLING/ACL-2006*, pages 452–459.

Kazama, J. and Torisawa, K. (2008). Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-2008*, pages 665–673.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. In *Proceedings-IJCNLP-2004*, pages 584–589.

Koo, T., Carreras, X., and Collins, M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-2008*, pages 595–603.

Lafferty, J., McCallum, A., and Pereira, J. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML01*, pages 282–289.

Liu, B. (2010). *Sentiment Analysis and Subjectivity*. CRC Press, Taylor and Francis Group.

Miller, S., Guinness, J., and Zamanian, A. (2004). Name tagging with word clusters and discriminative training. In *Proceedings of HLT-2004*, pages 337–342.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Proceedings-HLT-NAACL-2010*, pages 786–794.

Nakagawa, T., Kawada, T., Inui, K., and Kurohashi, S. (2008). Extracting subjective and objective evaluative expressions from the web. In *Proceedings of the 2nd International Symposium on Universal Communication*, pages 251–258.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings-EMNLP-2002*, pages 79–86.

Seki, Y., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (2010). Overview of multilingual opinion analysis task at ntcir-8: A step toward cross lingual opinion analysis. In *Proceedings of NTCIR-8 Workshop Meeting-2010*, pages 209–220.

Wang, Y., Kazama, J., Tsuruoka, Y., Chen, W., Zhang, Y., and Torisawa, K. (2011). Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings-IJCNLP-2011*, pages 309–317.

Zhao, J., Xu, H., Huang, X., Tan, S., Liu, K., and Zhang, Q. (2008). Overview of chinese opinion analysis evaluation. In *Proceedings of the First Chinese Opinion Analysis Evaluation*, pages 1–20.