Three-Phase Opinion Analysis System at NTCIR-6

Hironori Mizuguchi   Masaaki Tsuchida   Dai Kusui
NEC Internet Systems Research Laboratory
8916-47 Takayamacho, Ikoma, Nara 630-0101 Japan
{hironori@ab, m-tsuchida@cq, kusui@ct}.jp.nec.com

Abstract

We developed an opinion analysis system at NTCIR-6. Our system can detect opinion sentences and extract opinion holders by executing three phases: (1) opinion sentence classification by SVM that distinguishes an opinionated sentence from others, (2) opinion-holder candidate extraction using named entity recognition, (3) opinion-holder detection by rules that find the correspondence between the sentence and the holder. Characteristics of the system are the following two points: (a) in phase 1, the end of a sentence expression is added to the feature in SVM vector, and (b) phase 3 is separated into the author detection and the others. As a result of the evaluation, in opinion sentence judgment both precision and the recall ratio improved based on point (a). In opinion holder extraction, precision has improved greatly based on point (b).

Keywords: *Opinion Analysis, Opinion-Holder Detection, Author Detection.*

1 Introduction

In recent years, opinions about certain products or events have been made widely known on Internet product review sites, weblogs and so on. If we are able to extract and analyze these opinions, we can research products' markets and investigate public opinion.

Our research group studied reputation information extraction [3] [2] mainly for information on product review sites. Reputation information is information that contains expression of the evaluation of a product or service and so on. For example, ”Let It Be はすごく良い (Let It Be is very nice)” includes the expression of the evaluation, ”良い (nice)”.

Based on such background, we participated in the two Japanese subtasks in the Opinion Analysis Pilot Task of NTCIR-6: (1) opinionated sentence judgment and (2) opinion holder extraction. An opinionated sentence contains not only reputation information but also suggestion information. For example, ”水口氏は大統領を信任すべきだといった (Mr. Mizuguchi said that we should trust in the president)” is an opinion sentence that is a suggestion but does not give reputation information. ”水口氏 (Mr. Mizuguchi)” is the opinion holder.

First, we analyzed two tasks using the sample data in section 2. From the results of the task analysis, we developed our system in section 3. Our system has the following two characteristics: (a) as for opinionated sentence judgment, we pay attention to the end of a sentence expression; and (b) as for opinion holder extraction, it distinguishes the author from the rest. Lastly, an evaluation result and a problem are described in sections 4 and 5.

2 Task analysis

In this section, to get a clue as to opinion sentence judgment and opinion holder extraction, we investigate the sample data set.

2.1 Task analysis for opinion sentence judgment

The sample data are composed of 585 opinion sentences and 2167 non-opinion sentences. The rate of opinion sentences is about 20 percent.

We compared opinion sentences and non-opinion sentences to get a clue about opinion sentence judgment and discerned the following three attributes.

1. The end expression of an opinion sentence has a certain characteristic.
2. An opinion sentence often continues.
3. A remark is often an opinion sentence.

As for 1, we examined the use frequency of three characters at the end of each of the opinion sentences and the non-opinion sentences in the sample data. The top 10 rankings are shown in Table 1. The characters with hatching are expressions that appeared in both.

There are many end expressions that appear in opinion sentences but don't appear in non-opinion sentences. Therefore, the end expression is a valid handhold in opinion sentence judgment. On the other hand,

In opinion sentence

Characters	Freq.
だろう	40
ている	35
はない	16
られる	14
いない	14
もある	14
そうだ	14
である	13
がある	10
れない	10

In non-opinion sentence

Characters	Freq.
ている	145
だった	39
にした	36
ていた	33
なった	29
てきた	26
になる	24
された	23
られる	22
がある	21

Table 1. Frequency of three characters at the end of each opinion sentence

because there are expressions that appear in both, it is difficult to judge an opinion sentence using only the end expression.

For attribute 2, we investigated the number of opinion sentences that appear continuously and found it to be 239 sentences. To decide that the opinion sentences appear continuously, we compute the number of continuation appearances when an opinion sentence is randomly written. We mark 585 sentences of 2752 sentences randomly and count the number of continued marks. After doing this computation 500 times, the average was about 121 sentences. Therefore, because the number of continuously appearing sentences (239) is bigger than the number of continuous sentences (121) that are marked randomly, the opinion sentences appear continuously.

For attribute 3, we investigated the number of opinion sentences and non-opinion sentences that are in the remark. In Japanese, the kagikakko symbol (「」, Japanese quotation marks) often contains remark contents. So, we defined that a remark is a part equal to or more than 10 characters in the Japanese quotation marks.

As a result of the investigation, 182 opinion sentences and 189 non-opinion sentences were in the remarks. The ratio of opinion sentences to non-opinion sentences is 1 to 4. In the remarks, the ratio of opinion sentences to non-opinion sentences is 1 to 1. Therefore, a remark often becomes an opinion.

2.2 Task analysis for opinion holder extraction

In this section, we investigated the following two characteristics of the opinion holder.

Relative position	Author(%)	Non-Author(%)
Before 6 or more	19.9	15.8
Before 5 sentences	0.5	1.2
Before 4 sentences	0.4	1.2
Before 3 sentences	0.5	2.3
Before 2 sentences	0.5	3.9
Before 1 sentence	1.1	6.8
Same sentence	0.4	28.8
After 1 sentence	3.6	2.7
After 2 sentences	0.4	2.1
After 3 sentences	0.5	0.4
After 4 sentences	0.7	0.4
After 5 sentences	0.5	0
After 6 or more	5.1	0.3
TOTAL	**34.1**	**65.9**

Table 2. Ratio of author and non-author opinion holders appearing in relative positions

1. The characteristic of the words in the opinion holder.

2. The difference of the characteristic between author of opinion holder and non-author.

First, we investigated the characteristics of the words of opinion holders in the sample data set. There were the following two characteristics.

1. The opinion holder is almost always a person, location or organization.

2. The opinion holder is composed of the repeat of the noun, the symbol and the case particle ”の (in)”, coordinating particle.

The opinion holder includes not only the person but also the affiliation and the location of the person. For example, ”米国の研究者 (Researcher in the U.S.)” includes the person ”研究者 (Researcher)” and the location ”米国の (in the U.S.)”. Therefore, there is a second characteristic.

Next, we describe the difference of characteristics between the author of an opinion holder and a non-author.

34.1% of opinion holders are authors. Empirically, the opinion holder is almost always the author who wrote an article. So, more than 30% of opinion holders are authors.

We compare the relative position of the opinion sentence to the author holder and the non-author holder. Table 2 shows the ratio of opinion holders that are authors and non-authors, and the relative position of an

Depth level	Holder(%)
1	34.3
2	27.3
3	10.5
4	4.1
5	7.8
other	15.9

Table 3. Ratio of non-author holders and depth of parse tree

opinion sentence to the sentence that contains its opinion holder. Each value is the rate of the opinion holders that appear at the each position. There are the following differences between the author holder position and the non-author holder position. When the opinion holder is an author, the distance between the opinion sentence and the holder is far because the reporter (author) name appears at the head or the end of the article. When the opinion holder is a non-author, it appears in the place around the opinion sentence. About half of opinion holders (28.8%) of all non-author holders (65.8%) appear at same opinion sentence.

Next, we analyze the characteristic of the strings and the contexts of the neighborhood of the author and the non-author holder.

When the opinion holder is an author, the strings around the opinion holder have the same patterns. 79.1% of author holders exist in these patterns, ”（author）” or ”【author】”. The authors appear in the top or end of paragraphs.

When the opinion holder is a non-author, the strings around the opinion holder do not have the same patterns. Therefore, we investigated the context around the opinion holders and found the following three characteristics.

1. Most of all of the opinion holders are outside of the remark contents: 98.9% of non-author holders exist outside of the Japanese quotation marks that denote remarks.

2. The particles ”は (be)” or ”が (be)” often follow the opinion holder: 38.1% of opinion holders are followed by these particles, which become topic or subject markers.

3. The depth from the node of the opinion holder to the root node in the parse tree of the sentence that contains the opinion holder is low level. Table 3 shows the ratio of holder and depth level. There are 60% of holders at equal to or less than depth level 2.

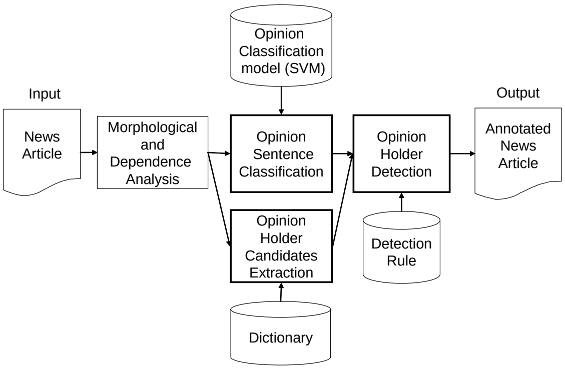

Opinion Classification model (SVM)
Input
News Article
Morphological and Dependence Analysis
Opinion Sentence Classification
Opinion Holder Detection
Output
Annotated News Article
Opinion Holder Candidates Extraction
Detection Rule
Dictionary

Figure 1. System architecture

3 Our System

In this section, we describe our system to judge opinion sentences and extract opinion holders.

Figure 1 shows our system architecture for inputting news articles and outputting them annotated with opinion sentences and opinion holders. Our system contains four components. At first, a news article is analyzed by the Morphological and Dependence Analysis component in order to preprocess for following the phases. The system processes three phases to output annotated data.

1. Opinion Sentence Classification by Support Vector Machine (SVM) that classifies sentence as opinion sentence or not

2. Opinion Holder Candidates Extraction using named entity resolution

3. Opinion Holder Detection by rules that find the correspondence of the opinion sentence and the holder in opinion holder candidates.

3.1 Morphological and Dependence Analysis

The news article is analyzed and output with a part of speech tag (POS tag) and dependency tree. We use the product ”Ko-BaKo/J” [1] in this analysis.

3.2 Opinion Sentence Classification

This component judges whether a sentence is opinion or not. We think that it is difficult to develop rules that detect opinion sentences because the judgment of an opinion sentence depends on the person. The sample data set contains 2800 sentences. We use SVM that learns the Opinion Classification Model to distinguish opinion sentences from other sentences.

We describe the features used by SVM. We consider the results of section 2.1 and use the following features:

Feature 1 Words, original form and POS tag of all morphemes

Feature 2 Words, original form and POS tag of the end phrase

Feature 3 Flag that denotes whether previous sentence is opinion or not

Feature 4 Words, original form and POS tag that distinguish between the front, inside and back of Japanese quotation marks

Each feature is placed in different elements in SVM.

3.3 Opinion Holder Candidates Extraction

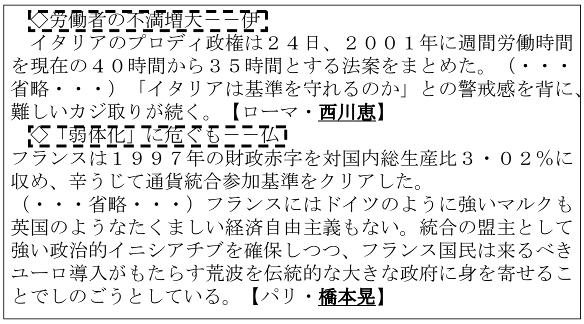
This component extracts the candidates of opinion holders. From the results in section 2.2, we must consider that an opinion holder is a person, location or organization and is composed of the repeat of the noun or some symbols or words. So, we achieve the following three steps:

1. Searching the noun phrase from the head of the sentence and adding this word to the word list
2. Adding to the word list on the condition that the following words are a noun phrase, case particle ”の (of)”, parallel particle or symbol
3. Extracting holder candidate if the end word of the word list is a person, organization or location.

These steps are executed on all noun phrases. In step 3, we use the predefined semantic category in ”Ko-BaKo/J”. The dictionary in Figure 1 includes the correspondence of the semantic category and the named entity category (person, organization, location).

3.4 Opinion Holder Detection

This component detects the correspondence of opinion sentence and the opinion holder in opinion holder candidates. From the results in section 2.2, we consider that there are many author holders. Author holders and non-author holders have different characteristics about the relative positions or the contexts. So, we separate author holder detection and non-author holder detection and process the following three steps:



◇労働者の不満増大ーー伊

イタリアのプロディ政権は２４日、２００１年に週間労働時間を現在の４０時間から３５時間とする法案をまとめた。（・・・省略・・・）「イタリアは基準を守れるのか」との警戒感を背に、難しいカジ取りが続く。【ローマ・**西川恵**】

◇「弱体化」に危ぐもーー仏

フランスは１９９７年の財政赤字を対国内総生産比３・０２％に収め、辛うじて通貨統合参加基準をクリアした。

（・・・省略・・・）フランスにはドイツのように強いマルクも英国のようなたくましい経済自由主義もない。統合の盟主として強い政治的イニシアチブを確保しつつ、フランス国民は来るべきユーロ導入がもたらす荒波を伝統的な大きな政府に身を寄せることでしのごうとしている。【パリ・**橋本晃**】

Figure 2. Example of Title Paragraph

1. Extracting author from each news articles.
2. Relating opinion holder candidates to opinion sentence.
3. Relating author to opinion sentence.

3.4.1 Extracting author

This step extracts the author from a news article. There is a case of more than one author sharing an article. So, this step can relate the author to the paragraphs using ”Title Paragraph.”

Title Paragraph denotes a title of the following paragraphs in article contents. Figure 2 shows an example ”Title Paragraph” that is ”労働者の不満大ーー伊 ” and ”「弱体化」に危ぐもーー仏 ”. The role of these paragraphs is title. Also, when more than one author exists, the author often appears among these Title Paragraphs.

System detects an author by the following procedure.

- Finding a Title Paragraph whose length is less than about one third of the length of the previous paragraph. This ratio is set empirically.
- Extracting words that appear in specific patterns at the beginning or end of paragraphs. The specific patterns are ”（*） ” or ”【*】 ”. This time, this system used the pattern obtained from the analysis in section 2.2, but this process may achieve by a conventional pattern extraction technique and so on.
- Filtering words that are the person or organization using the ”Ko-BaKo/J” semantic category, the same as Opinion Holder Candidates Extraction in section 3.3.
- Relating the author to the sentences of paragraphs between Title Paragraphs if there are two or more authors. If there is one author, the author is related to all sentences.

3.4.2 Relating opinion holder candidate to opinion sentence

This step calculates the score in the combination of the following four rules, and detects the correspondence between opinion holder candidates and opinion sentence:

- Opinion holder candidate appears in opinion sentence.
- Opinion holder candidate is outside the Japanese quotation marks.
- The particle ”は(be)” or ”が(be)” follows the opinion holder candidates.
- The depth level of the node of the opinion holder candidate in the parse tree.

The opinion holder candidates that conform to these rules received the scores. The opinion holder that has the highest score is related to the opinion sentence.

3.4.3 Relating author to opinion sentence

This step relates the author to an opinion sentence that is not related to the opinion holder candidate.

Each sentence is already related to an author at step 1 (Extracting author). If the opinion sentence is not related to the opinion holder candidates until the previous step, the opinion sentence is related to the author as opinion holder.

4 Evaluations and discussions

In this section, we describe the two submitted systems (EHBN-1, EHBN-2) and the result of the formal run. After that, for a detailed evaluation of the system, we evaluate the opinion sentence judgment and opinion holder extraction.

We submitted two systems named EHBN-1 and EHBN-2 to apply to the Japanese Opinion Analysis Task. The two systems differ in the method of opinion holder detection. EHBN-1 relates an opinion sentence to the nearest opinion holder candidate as opinion holder in the previous sentences. EHBN-2 implements the method of section 3.4.

4.1 Evaluation of opinion sentence judgment

We evaluate the proposal method in section 3.2 of opinion sentence judgment by experiment. To evaluate, we conduct the following two experiments. In experiment 1, to evaluate the general effects of the proposal method, we compare a simple method and the proposal method by precision, recall and F-value. In experiment 2, to evaluate the effect of the proposed features, we compare some combinations of features.

Method	Standard	Precision	Recall	F−measure
Proposal method (1 and 2 and 3 and 4)	Lenient	**58.91%**	**45.36%**	**51.25%**
	Strict	**45.85%**	**47.92%**	**46.86%**
Baseline (1)	Lenient	52.85%	45.26%	44.32%
	Strict	41.22%	47.92%	44.32%

Table 4. Results of Experiment 1

Method	Standard	Precision	Recall	F−measure
Method1 (2)	Lenient	**58.80%**	38.77%	46.73%
	Strict	**47.68%**	42.67%	45.04%
Method2 (1 and 3 and 4)	Lenient	53.20%	45.86%	49.26%
	Strict	40.83%	47.79%	44.04%
Method3 (2 and 3 and 4)	Lenient	57.53%	**49.09%**	**52.98%**
	Strict	45.47%	**52.67%**	**48.81%**

Table 5. Results of Experiment 2

We use sample data for the training and formal run data for the test.

4.1.1 Experiment 1

The proposal method in section 3.2 uses four kinds of features. General methods of the classification of documents or sentences use all morpheme information. Therefore, the baseline method merely uses morpheme information (feature 1 in section 3.2).

Table 4 shows the results of baseline and proposal methods. The precision ratio is improved by 4% in strict results, 6% in lenient results. The recall ratio approximately didn’t change. The effect of the proposal method could be confirmed because the recall ratio was equal and the precision ratio was improved.

4.1.2 Experiment 2

Using the combination in the following features in section 3.2, we evaluate the effect of the features.

Method 1 Only feature 2

Method 2 Using feature 1, 3 and 4

Method 3 Using feature 2, 3 and 4

Feature 3 and 4 are only combined with feature 1 or 2 because feature 3 and 4 are supplementary features. We calculate each precision, recall and F-value and compare these combinations and the Proposal Method and Baseline of experiment 1.

The effect of feature 2 is evaluated by the comparison between method 2 and method 3. Features 3 and

国連貿易開発会議ＵＮＣＴＡＤ
渡辺賢一郎・国際金融情報センター統括審議役
アジア諸国など
堀内昭義・東大教授
Ｇ７
外交筋
地元紙

Table 6. Examples of opinion holders that are not in opinion holder candidates

4 with feature 1 are evaluated by the comparison between the Baseline and method 2, and feature 3 and 4 with feature 2 evaluated by method 1 and method 3.

Table 5 shows the results. When comparing tables 4 and 5, feature 3 and 4 contribute to the precision and feature 2 contributes to the recall. When not using all morphemes (feature 1), the recall became higher than the Proposal method and the F-value became high in the results.

As a result, the effective feature set is feature 1 or the combination of feature 2, 3 and 4.

4.2 Evaluation of opinion holder extraction

4.2.1 Evaluation of opinion holder candidates

We calculate the cover ratio of opinion holders out of opinion holder candidates, and describe the opinion holders that were not in opinion holder candidates.

We evaluate the cover ratio with existing opinion holders in the opinion holder candidates. The high cover ratio is a condition for the high recall of the opinion holder extraction. When the string of the opinion holder candidate partially equals the opinion holder, the opinion holder exists in opinion holder candidates. The number of opinion holders is counted in each article.

The cover ratio was 81.1%.

Next, the opinion holders that are not opinion holder candidates are shown in Table 6. Most were a named entity.

To raise the cover ratio, the named entity resolution method must be changed. Our system uses the output of ”Ko-BaKo/J” but it may be achieved by a different existing method [4].

4.2.2 Evaluation of opinion holder detection

We compare the Proposal method (EHBN-2) with the Baseline (EHBN-1), which relates the opinion sentence to the nearest opinion holder candidates. Table 7 is the result of the lenient of the formal run. Precision was 18% improved and recall was 1% improved.

	Precision	Recall	F-measure
EHBN-1 (Base)	13.83%	8.51%	10.53%
EHBN-2 (Rule + Author)	31.38%	9.74%	14.86%

Table 7. Results of opinion holder extraction of the lenient at formal run

Because precision was improved two-fold, according to the rule at the relating phase in section 3.4.2, a lot of wrong answers could be removed.

To evaluate the effect of author detection, we calculate the ratio of the author of the opinion holder that is the right answer in return by our system. As a result, about 50% of extracted holders were the authors. Therefore, we confirmed that there was a big effect in author detection.

Because half of the opinion holders are authors and the recall is low, the rules of relating opinion holder narrow down too much. In the future, it is necessary to create a rule that takes into account recall.

5 Conclusions

We developed a system that performs opinion sentence judgment and opinion holder extraction. As a result of analyzing sample data, our system has the following two characteristics: (1) opinion sentence judgment by SVM, which added the feature of the end expression of the sentence; (2) opinion holder extraction by different process of author holder detection and the others.

As a result of the evaluation, in opinion sentence judgment, both precision and recall were improved. In opinion holder extraction, precision was improved.

In the future, we will investigate the problem of the improvement of recall by improving the rule of relating the opinion holder to the opinion sentence.

References

[1] Japan System Applications Co., Ltd, Ko-BaKo/J, http://www.jsa.co.jp/LANG/ (in Japanese).

[2] K. Tateishi, Y. Ishiguro, and T. Fukushima. A reputation search engine that collects people’s opinions by information extraction technology. *IPSJ Transactions on Databases*, 22:115–123, 2004.

[3] M. Tsuchida, H. Mizuguhi, and D. Kusui. Opinion extraction by identifying object-attribute-evaluate relations (in japanese). *In Proc of 13th Annual Meeting of the Association for Natural Language Processing*, pages 412–415, March 2007.

[4] H. Yamada, T. Kudo, and Y. Matsumoto. Japanese named entity extraction using support vector machine (in japanese). *IPSJ Journal*, 43(1):44–53, 2002.