Extraction of Comparative Sentences and their Components from BBS Messages

Keita Ozaki, Fuminori Kimura, and Akira Maeda

Abstract—**In this paper, we describe a method for extracting comparative sentences and their components from messages posted on bulletin board systems (BBSs) or discussion forums. A comparative sentence is a message representing the merits and demerits between two or more targets. We propose a method for extracting a set of object-attribute-evaluation triples, which are the elements that make up a comparative statement. We first extract comparative sentences that meet our definition of a comparative sentence by using the method devised from observations of actual sentences in a BBS. Then, by analyzing the messages by using the difference in characteristics of each component of the comparative sentence, we identify object-attribute-evaluation triples, which are the components of a comparative sentence. We conducted evaluation experiments with the proposed method. As a result, we raised the extraction of the comparative sentence and the extraction accuracy of a component.**

Index Terms—**Comparative Sentence, Forum, Expression of evaluation**

I. Introduction

On the Internet, the number of communities, in which many and unspecified people can exchange opinions, are increasing. One typical example is the bulletin board system (BBS). A BBS consists of many threads that have comments posted in relation to the topic of the thread. In each thread, users can communicate, talk, and argue with many people about the topic. In some of these threads, users talk about the differences between things related to the topic, for example, the superiority or inferiority of the one to the other. Messages in these threads contain information for judging which is superior in a certain aspect. It is helpful to judge which is superior if we can obtain these information.

In this paper, we call messages containing this information "comparative sentences." We aim to obtain this information from comparative sentences. In this paper, we use threads on the bulletin board system "2channel" [1] as the target of an experiment. We conducted experiments on comparative sentence extraction and its components in order to evaluate the proposed method.

Keita Ozaki is with the Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan (corresponding author to provide phone: 077-599-4365; fax: 077-599-4365; e-mail: is011084@ed.ritsumei.ac.jp).

Fuminori Kimura is with the Kinugasa Research Organization, Ritsumeikan University, Kyoto, Kyoto 603-8577, Japan (e-mail: fkimura@is.ritsumei.ac.jp).

Akira Maeda is with the College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan (e-mail: amaeda@is.ritsumei.ac.jp).

II. Related Work

In this section, we introduce related research on extracting and classifying comparative sentences and evaluation aspects.

For the extraction of comparative sentences, Kurashima et al. [2] focused on the superiority or inferiority between comparative objects, e.g., a sentence saying which restaurant is cheaper or more delicious between restaurants A and B, and proposed a method for extracting four kinds of elements that constitute a comparative sentence, i.e., criteria, object, attribute, and evaluation, from sentences, by using rules devised from observing actual comparison expressions. Jindal and Liu [3] collected comparative sentence candidates comprehensively using a manually created list of clue word clauses that express comparisons in English, and they proposed a comparative sentence classifier that uses class sequential rules created from the list as the feature. Regarding the extraction of evaluation aspects, Iida et al. [4] extracted groups of two or more evaluation candidates and one attribute, obtained the optimal combination of an attribute and its evaluation by supervised learning, and extracted pairs of attribute-evaluation. Suzuki et al. [5] extracted the candidates of object-attribute-evaluation triples by combining the naive Bayes classifier and EM algorithm.

In our proposed method, comparative sentences are extracted independent of specific expressions. In addition, in the extraction of evaluation aspects, our method focuses on colloquial expressions that are used mainly on BBSs in Japanese, which has not been dealt with in previous studies. To tackle the problem of colloquial expressions in Japanese that are often grammatically broken, we use a clause, instead of a word or a clause, as the unit of comparative elements, i.e., target, attribute, and evaluation.

III. The Definition of a Comparative Sentence

In this section, we define a comparative sentence and explain the comparison components of the such sentences.

We define comparative sentences as messages that fulfill any of the below three definitions.

Definition 1
A message contains two "objects" (candidate for comparisons).
Definition 2
A message contains "evaluation" for at least one "object."
Definition 3
A message mentions relationships between "objects," such as "superiority or inferiority," "equivalent," "the best," and "the feature."

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

We introduced an example of a comparative sentence on the bulletin board that fulfills the above three definitions.

In the message shown in Figure 1, there are two candidates for comparison. The first is "ウイイレ" (the name of the "Winning Eleven" game series), and the second is "FIFA" (the name of a game series), a typical soccer game (definition 1). Arbitrary evaluations are given to both objects (definition 2), and the superiority or inferiority between the objects can be read as a relationship (definition 3). Therefore, this message is a comparative sentence. In actuality, a comparative sentence needs to fulfill not all three definitions but at least one.

Although the playability of ウイイレ is severe, the playability of FIFA is good.

Fig. 1. Message with relation of merits between objects

The components of a comparison sentence, "object," "attribute," and "evaluation," are extracted from the comparison sentence. A clause is extracted as these components. The reason for using a clause is that colloquialisms are used mainly for messages on bulletin boards. It is difficult for a morphological analyzer to separate colloquialisms into the correct words. For example, when we extract the phrase "kore dake na no ka" as word units, it is divided like "kore" "dake" "na" "no" "ka." We thought that using clauses is suitable for the colloquialisms on bulletin boards because we can extract such words as one block when extracting them as clause units.

When we extract the components of the comparison sentence from the above example message, the object becomes "FIFA," the attribute "playability" and the evaluation "good."

IV. Proposed Method

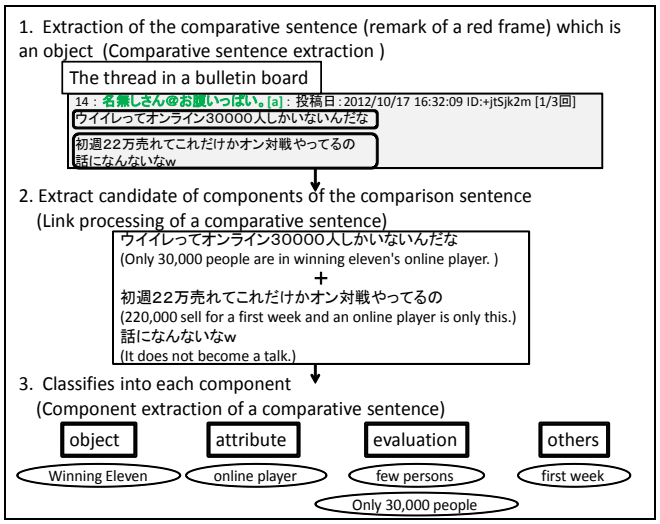

Fig. 2. Outline of proposed method

An outline of the proposed method is shown in Figure 2. First, the system extracts the comparative sentence. Second, it extracts the sentence that becomes a group of the components of that comparative sentence ("object," "attribute," "evaluation"). Third, it classifies clauses into these components.

4.1. Comparative Sentence Extraction

In this section, we explain the method for extracting a comparative sentence from a bulletin board. Considering definitions 1 and 2 of a comparative sentence, most comparative sentences have at least one component pair that consists of an "object" and "evaluation" (the underline part "it was good" or not "not quite satisfactory" shown in Figure 3).Also, when we collected the comparative sentences of 340 messages from the bulletin board and observed them, we found that the "evaluation" in these component pairs related to their "object".

裏は黒はよかった。白はいまいち。
(The black reverse side was goodWhite was not so good.)

Fig. 3. Comparison message on actual bulletin board

In the comparative sentence extraction process, we extract sentences that have component pairs including such a feature. Figure 4 shows the processing flow of this process.

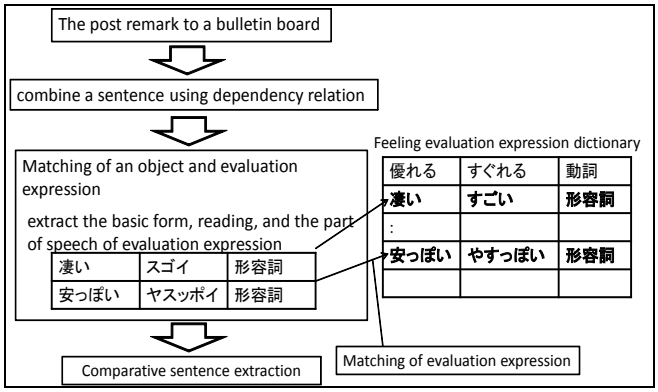

Fig. 4. Flow of comparative sentence extraction

First, the system conducts a morphological analysis for sentences in order to judge if "evaluations" relate to "objects." The system then extracts all clauses pairs that are in a dependency relation. In this system, we use Cabocha [6] in order to obtain a dependency relation. Cabocha is a tool that outputs the result of a morphological analysis, which divides each word, and the divided data of a clause from the sentence of an analysis object.

Second, matching processing of an object and evaluation is performed. To judge whether the "evaluation" concerning an "object" is suitable, an evaluation expression dictionary [7] is used. When all three pieces of information, "the basic form," "reading," and the "part of speech" in the result of the dependency analysis for a word match the three elements of information in the dictionary, it is judged with the word being "evaluation." In Figure 4, the item that is in agreement with all three elements of information, "凄い" (superb), "スゴイ (sugoi), "形容詞" (adjective), is included in the dictionary. When this matching processing is performed on each sentence and two pairs of "object" and "evaluation" are matched with this processing, that sentence is extracted as a comparative sentence.

4.2. Object Scope Extension

In this process, we extend the scope of influence of objects to next sentences. Some sentences contain no "objects," although they contain "attributes" or "evaluations." The components "attributes" and "evaluations" in these sentences are also informative if they are related to "objects" in preceding comparable sentences. However, the comparative sentence extraction mentioned in section 4.1 cannot extract these components. To do so, the method conducts an object

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

scope-extension process. An example of this processing is shown in Figure 5.

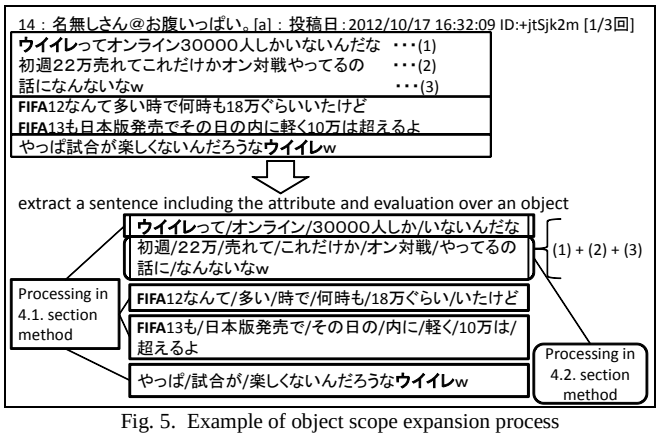

Fig. 5. Example of object scope expansion process

This process analyzes a message in order from the top sentence to bottom sentence. If a sentence contains the candidate for comparison, that sentence is analyzed as a starting point (upper side of Figure 5). Next, the system moves focus onto the next sentence of the starting point. If there are no "objects" in this next sentence, this sentence is merged into the starting point sentence in order to extend the scope of the "object" at the starting point. This step is repeated until new "objects" appear in the focused on sentence. If a new "object" appears, the system sets the focused on sentence as the new starting point and repeats these steps.

For example, when the candidate for comparison "ウイイレ" exists in sentence (1) in Figure 5 and the candidate for comparison does not exist in sentence (2), it is judged as being a sentence in which the evaluation of the candidate for comparison "ウイイレ" is included, and sentence (2) is combined with sentence (1). Let this be sentence (1+2). Since the candidate for comparison is not contained in sentence (3), sentence (3) is combined with sentence (1+2). An object, its attribute, and evaluation are extracted from the group of the sentences divided by this processing. At this time, the contributed messages on a bulletin board are colloquial expressions in many cases, and there are many patterns that are not grammatically correct Japanese, like literary expressions. Therefore, the component of a comparative sentence is not taken out per word but is taken out per clause.

4.3. Component Extraction of Comparative Sentence

After the object scope-extension process, the system classifies clauses divided by the dependency analysis into either of each component, "object," "attribute," "evaluation," and "others" (a clause that is not applied to a component), of a comparative sentence. The flow of this processing is shown in Figure 6. There are two kinds of classification processing. One is classification 1, which uses the appearance pattern of the part-of-speech sequence in a clause, and the other is classification 2, which uses the evaluation value computed by using the word importance and part-of-speech information in a clause. First, the system classifies clauses into components by classification 1. Second, clauses that are unclassified by classification 1 are classified by classification 2.

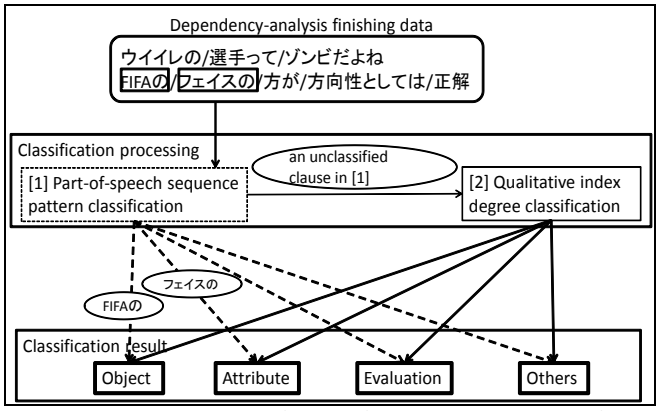

Fig. 6. The component classification of the comparative sentence from a clause

4.3.1. Part-of-Speech Sequence Pattern Classification

In this process, a clause is classified into each component by using the appearance pattern of the part of speech in it. Some patterns of the part of speech in clauses are obviously classified into specific components. The appearance patterns are considered from the data of the clause used as the correct answer collected manually.

An appearance pattern is a part of speech of one word or a sequence of the part of speech that appears in the clause frequently as a certain component. For example, the appearance patterns of "object" in the part of speech in a clause are a "noun (proper noun) – particle (linking particle)". (e.g. "FIFA/は(FIFA is)" (noun (proper noun) – particle (linking particle))).

A clause that is in agreement with this feature is classified into the component "object."

If there are unclassified clauses in a sentence after conducting classification 1, they are classified by classification 2.

4.3.2. Qualitative Index Degree Classification

Next, we use the part-of-speech information on a word. When extracting a comparison component, the part of speech in the clause, which is an object of analysis, is also used. We thought that the extraction accuracy of a comparison component could be raised by using the appearance probability, which shows the probability that each part of speech will appear as one of the components ("object," "attribute," "evaluation," and "others").

We use the part of speech of each word contained in a clause and the sequence of the part of speech of words in a clause. An example of the part-of-speech information to be used is shown below. In this example, each part of speech, called the "noun" and a "particle" of clause (1), is used as part-of-speech information. In addition, the sequence of "noun–particle," which is a sequence of the part of speech in clause (1) is used as part-of-speech information. The part-of-speech information included in the clause below clause (1) is used as well.

- 「ウイイレ/って (Winning Eleven is)」名詞 + 助詞 (noun-particle) - (1)
- 「オン/対戦 (online player)」名詞 + 名詞(noun-noun)
- 「これ/だけ/か (only this)」名詞 + 助詞 + 助詞 (noun-particle-particle)

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The appearance probability of the part of speech in the word unit in the clause of each component and the appearance probability of the sequence of the part of speech in all the words in a clause are used. Bayes's law is applied in order to compute appearance probability. In this calculation, the probability to solve for is the probability that each part of speech belongs to each element. To search for this appearance probability, we collected the data of each component manually.

The system uses these data in order to calculate the appearance probability that a clause is a certain sequence of a part of speech when the clause is supposed to be classified into a certain component. For example, the joint probability P (A ∩ B) (prior probability) in the case where a component is an "object" and a part of speech is a "noun" is calculated by the following formula (1).

$$e.g.\quad P(object \cap noun) = P(noun \mid object)P(object) \tag{1}$$

Generally, it is expressed with the following formula.

$$P(A \cap B) = P(A \mid B)P(B) \tag{2}$$

By using Bayes's law for formula (2), it can be denoted by the probability to which element each part of speech belongs. The following formula (3) calculates the probability value (posterior probability). The notations from A_0 to A_3 mean each of the components, "object," "attribute," "evaluation," and "others," respectively. *B* expresses each part of speech, such as a noun and an adjective.

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{i=0}^{n} P(B \mid A_i)P(A_i)} \tag{3}$$

In this process, we apply a tfidf value to the calculated probability value in formula (3). Since this value makes importance high in the order of the component "object," "evaluation," "evaluation," "others," it is necessary to carry out weighting in the order of "object" > "attribute" > "evaluation" as a value to apply. For weighting, it is necessary to take into consideration the character between each comparison component. If the appearance probability of an object or an attribute in a certain part of speech is high, it will get higher weighting, and it will get lower weighting if the appearance probability of evaluation or others is high.

As shown in the following formula (4) considering such an idea, the importance of the part of speech in the comparison component of a comparative sentence is expressed with the one value *W*(B).

$$W(B) = \frac{3 \times P(A_0 \mid B) + 2 \times P(A_1 \mid B) + P(A_2 \mid B)}{P(A_1 \mid B) + 2 \times P(A_2 \mid B) + 3 \times P(A_3 \mid B)} \tag{4}$$

Tf-idf Scoring

In this section, we explain how to use a tf-idf value.

It is a high possibility that "objects" represent a specific proper noun. Therefore, it is thought that the difference in importance appears as "object" > "attribute" > "evaluation" between each component. To calculate the importance of a word in a clause, which is each component, we used the tfidf method. We consider one thread of a bulletin board as one document and calculate a tfidf value. Finally, when classifying a clause into the component of a comparative sentence by using the value calculated with formula (5), we combine a tfidf value with the part-of-speech information in a clause.

In this section, we explain the technique of extracting the comparison component of a comparative sentence by combining the tfidf value and value computed by using the part-of-speech information searched for with Section 4.3.2.

$$I = \frac{\sum_{x=1}^{t} tf \times \log \frac{N}{n} \times W_1(B) \times W_2(B)}{t} \tag{5}$$

Formula (5) calculates the importance in a clause by combining tfidf and part-of-speech information. In formula (5), *tf* is the frequency of the appearance of a certain word in a certain thread, and *idf* is the number that divides the total number of threads by the number of the threads containing a certain word. Here, *t* shows the number of words in a clause. $W_1(B)$ is the numerical value that uses the part-of-speech information on the word unit in a clause and is calculated in the course from formula (1) to formula (3). $W_2(B)$ is the numerical value that uses the row of the part of speech in all words in a clause and is calculated in the course from formula (1) to formula (3). *I* is a numerical value that shows the importance in a clause.

We use as significance the value that multiplies the tfidf value and part-of-speech information. Moreover, since the difference of the importance *I* of a clause becomes large with the number *t* of words contained in a clause, it is normalized with the number *t* of words in a clause.

In TABLE I, the values in cells from the 2nd to the 5th row are the posterior probability that use Bayes's law, and the value of the right side is W1 (B) calculated by using the numerical value of the left four posterior probabilities.

TABLE I
POSTERIOR PROBABILITY AND $W_1(B)$ (PART OF THE LIST)

Part of speech	Object	Attribute	Evaluation	Others	Part-of-speech importance
noun(not independent and adjective verb stem)	0	0	0.3283	0.6717	0.1229
noun(proper noun and region)	0.1228	0.7125	0	0.1647	1.486
symbol(general)	0.0125	0.0081	0.2082	0.7712	0.0956
noun(general)	0.4428	0.1508	0.0269	0.3795	1.2335
noun(not independent and adverb possible)	0	0	0.0112	0.9888	0.0038

Finally, since the difference of the importance *I* widens between each component, a threshold value is manually set up between "object" and "attribute," "attribute" and "evaluation," "evaluation," and "others." The system then extracts each component.

V. EXPERIMENTS

In this section, we describe the results of the evaluation experiments with the method proposed in Section 4.

We evaluate the proposed method in two stages: an experiment to extract a comparative sentence from one thread and another to extract components from comparative sentences.

5.1. Evaluation of Comparative Sentence Extraction

We used "2channel" as the target BBS for the experiments and conducted the experiments for three threads on that BBS. A total of 1215 messages from threads A, B, and C, shown in TABLE II, were used in the experiments. Comparative

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

sentences that exist in those threads were extracted by using the method proposed in Section 4.1. It is necessary to first input two objects to the system. The actual objects entered are shown in the "object" column of TABLE II.

TABLE II
EXPERIMENTAL RESULTS OF COMPARATIVE MESSAGE EXTRACTION

Thread	Object	Number of messages	Number of extraction messages	Precision
A	白(white), 黒(black)	355	90	0.867
B	iPad, Nexus7	422	11	0.909
C	FIFA, ウイイレ	438	49	0.857

The resultant precision is shown in TABLE II. Recall is also an important measure for comparative sentence extraction. However, recall is not calculated in consideration that the data extracted as noise is important in component extraction from the comparative sentences, which is the purpose of this research, and it is time-consuming to collect all the correct answer sentences.

5.2. Evaluation of Comparative Sentence Component Extraction

In this experiment, the components of the comparative sentences extracted by the experiment described in Section 5.1 are extracted. Thread C in TABLE II was used as the object of this experiment. The correct answer data used for this experiment was collected manually by one of the authors. The details of the collected correct answer data are summarized in the table.

TABLE III
COLLECTED CORRECT ANSWER DATA

Object	Number of messages	Number of extracted messages	Sum total
FIFA, ウイイレ	438	42	1058

At this time, the number of threads (the number of documents) used for the calculation of the tfidf value of a word was 108. The number of messages was 104,544, and the number of words was 3,072,131. When calculating tfidf values, symbols, numbers, and white-spaces are excluded.

When the components of the comparative sentences are extracted from the thread, the number of extracted clauses that are correct is regarded as the number of correct answer extraction. The "sum total" of TABLE III is the total number of clauses of "object," "attribute," "evaluation," and "others." The rate of correct answer extraction in the 366 clauses except for "others" is evaluated as recall. The rate of correct answer extraction in the total number of extraction is evaluated as precision.

TABLE IV
EXPERIMENTAL RESULTS OF COMPARATIVE MESSAGE COMPONENT EXTRACTION

Method	Precision	Recall	F-measure
tfidf	0.385	0.634	0.479
tfidf + Part of speech	0.509	0.522	0.515
tfidf + Part of speech + Part-of-speech sequence pattern	0.556	0.549	0.552

TABLE IV shows the precision, recall, and F-measure for three kinds of methods. The first one is the method of only using tfidf. The second is the method of combining tfidf and part-of-speech information. The third is the method of combining the part-of-speech sequence pattern classification, tfidf, and part-of-speech information. As shown in the table, F-measure improved about 3.6% by adding the part-of-speech information. Moreover, it improved about 3.7% by adding the part-of-speech sequence pattern classification.

The actual sentences extracted by the method of using "tfidf + part-of-speech + part-of-speech sequence pattern" are shown in Figure 7. The extracted clauses and the components estimated by the proposed method are shown in the square column at the bottom half of the figure. At the right of each component, the estimated category is shown, and "○" is marked if the estimated category is correct, and "×" is marked if it is not.

Since it is difficult to regard a message as a contributor's opinion clearly if it is an interrogative sentence, which contains "?" at the end of a sentence, these sentences are ignored. The "object" and "attribute" components contained in this comparative sentence were mostly extracted correctly, and "others" were correctly judged. Since the combinations of the part of speech such as "noun + particle," which appears frequently as an "object" and "attribute," are actually used for the "object" and the "attribute" in the message, they were successfully extracted.

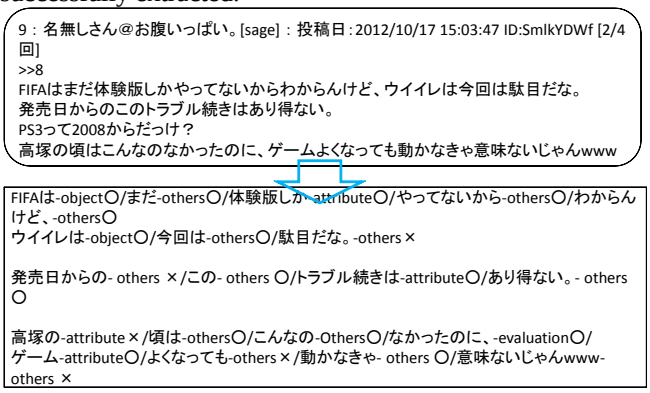

Fig. 7. Example of component extraction from comparative sentences (1)

The comparative sentences that were not correctly extracted with the "tfidf + part-of-speech + part-of-speech sequence pattern" of TABLE IV are shown in Figure 8. In the incorrectly extracted clauses, one of the noticeable cases is a clause like "楽しくないんだろうなウイイレw" (ウイイレ will not be fun). There are some cases that cannot be divided by the dependency analysis, i.e., between "な" and "ウ" in this case. It becomes a correct answer if "楽しくないんだろうな" (will not be fun) is judged as an "evaluation" and "ウイイレ" is judged as an "object." Also, although the clause "やってるの (doing)" should be extracted as an "attribute" that expresses the meaning of a "user," it was extracted as "others."

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

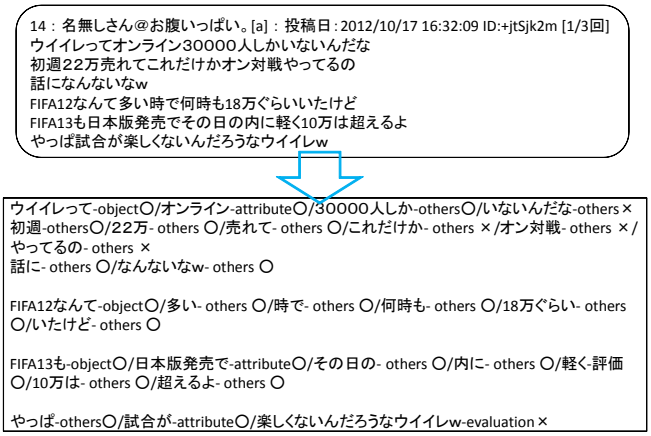

Fig. 8. Example of component extraction from comparative sentences (2)

VI. Discussion

In this section, we discuss the results of the experiment described in Section 5. One of the main causes of low F-measure in TABLE IV is that the precision of "evaluation" is low. The cause of this cause is that there is no difference in the importance of "evaluation" and "others." To solve this problem, it is necessary to consider the judgment of "evaluation" in an "evaluation" dictionary and the judgment with the message and surface character string that are contained in "others."

The case where a dependency analysis does not function is mentioned as a factor to which precision became low. A sentence like "楽しくないんだろうなウイイレ" was judged as one clause and suitable an "evaluation" and "object" were not able to be extracted. It turned out that a dependency analysis was not correctly made on the message currently written without the interval, such as the word "だろうな," which was applied to the ending, and the noun "ウイイレ."

"やってるの" will become a "やっ", "てる", and "の", if a morphological analysis divides into a word.

It turned out that that influence is the cause by which the importance using the tfidf value and part-of-speech information on a word unit becomes low.

The extraction method of the message in a message unit is also a problem, and there was a problem which also extracts an unnecessary message.

VII. Conclusion

In this paper, we proposed a method for extracting comparative sentences, which is based on syntactic analysis and dependency parsing, and the matching of the targets for comparison and "evaluations." In addition, we proposed a method for extracting the components of comparative sentences by using pattern classification by part of speech and the tfidf value and part-of-speech information. We conducted experiments with our proposed method by using actual BBS threads and obtained promising results.

For future work, we are planning to use features that are not used in the current method, such as the positions of components in messages and the information of topic categories, e.g., sports, entertainment, politics, etc., in order to improve the accuracy of the proposed method.

References

[1] 2channel http://menu.2ch.net/bbstable.html
[2] K. Kurashima, K. Bessyo, T. Utiyama, R. Kataoka, " Ranking Method using Comparative Relations ", DEWS2007.
[3] N. Jindal and B. Liu. Identifying comparative sentences in text documents. In Proc. of SIGIR 2006, 2006.
[4] R. Iida , N. Kobayashi , K. Inui , Y. Matumoto , K. Tateishi , S. Hukushima, "A Machine Learning-Based Method to Extract Attribute-Value Pairs for Opinion Mining " , The Special Interest Group Notes of IPSJ. IPSJ-SIGNL 2005(1), 21-28, 2005.
[5] Y. Suzuki，H. Takamura，M. Okumura，"Evaluation list expression extraction for Weblog"， The Japanese Society for Artificial Intelligence，SIG-SW&ONT-A401-02, 2004.
[6] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer http://code.google.com/p/cabocha/
[7] H. Takamura, T. Inui, M. Okumura, "Extracting Semantic Orientations using Spin Model, Transactions of Information Processing Society of Japan 47(2), 627-637, 2006.
[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)