A System for Summarizing and Visualizing Arguments in Subjective Documents: Toward Supporting Decision Making

Atsushi Fujii
Graduate School of Library,
Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

Tetsuya Ishikawa
The Historiographical Institute
The University of Tokyo
3-1 Hongo 7-chome, Bunkyo-ku
Tokyo, 133-0033, Japan
ishikawa@hi.u-tokyo.ac.jp

Abstract

On the World Wide Web, the volume of subjective information, such as opinions and reviews, has been increasing rapidly. The trends and rules latent in a large set of subjective descriptions can potentially be useful for decision-making purposes. In this paper, we propose a method for summarizing subjective descriptions, specifically opinions in Japanese. We visualize the pro and con arguments for a target topic, such as "Should Japan introduce the summertime system?" Users can summarize the arguments about the topic in order to choose a more reasonable standpoint for decision making. We evaluate our system, called "OpinionReader", experimentally.

1 Introduction

On the World Wide Web, users can easily disseminate information irrespective of their own specialty. Thus, natural language information on the Web is not restricted to objective and authorized information, such as news stories and technical publications. The volume of subjective information, such as opinions and reviews, has also been increasing rapidly.

Although a single subjective description by an anonymous author is not always reliable, the trends and rules latent in a large set of subjective descriptions can potentially be useful for decision-making purposes.

In one scenario, a user may read customer reviews before choosing a product. In another scenario, a user may assess the pros and cons of a political issue before determining their own attitude on the issue.

The decision making in the above scenarios is performed according to the following processes:

(1) collecting documents related to a specific topic from the Web;
(2) extracting subjective descriptions from the documents;
(3) classifying the subjective descriptions according to their polarity, such as positive/negative or pro/con;
(4) organizing (e.g., summarizing and/or visualizing) the classified descriptions so that users can view important points selectively;
(5) making the decision.

Because it is expensive to perform all of the above processes manually, a number of automatic methods have been explored. Specifically, a large number of methods have been proposed to facilitate processes (2) and (3).

In this paper, we focus on process (4), and propose a method for summarizing subjective information, specifically opinions in Japanese. Our method visualizes the pro and con arguments for a target topic, such as "Should Japan introduce the summertime system?"

By process (4), users can summarize the arguments about the topic in order to choose a more reasonable standpoint on it. Consequently, our system supports decision making by users.

However, process (5) is beyond the scope of this paper, and remains an intellectual activity for human beings.

We describe and demonstrate our prototype system, called "OpinionReader". We also evaluate the components of our system experimentally.

Section 2 surveys previous research on the processing of subjective information. Section 3 provides an overview of OpinionReader, and Sec-

Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 15–22,
Sydney, July 2006.

tion 4 describes the methodologies of its components. Section 5 describes the experiments and discusses the results obtained.

2 Related Work

For process (1) in Section 1, existing search engines can be used to search the Web for documents related to a specific topic. However, not all retrieved documents include subjective descriptions for the topic.

A solution to this problem is to automatically identify diaries and blogs (Nanno et al., 2004), which usually include opinionated subjective descriptions.

For process (2), existing methods aim to distinguish between subjective and objective descriptions in texts (Kim and Hovy, 2004; Pang and Lee, 2004; Riloff and Wiebe, 2003).

For process (3), machine-learning methods are usually used to classify subjective descriptions into bipolar categories (Dave et al., 2003; Beineke et al., 2004; Hu and Liu, 2004; Pang and Lee, 2004) or multipoint scale categories (Kim and Hovy, 2004; Pang and Lee, 2005).

For process (4), which is the subject of this paper, Ku et al. (2005) selected documents that include a large number of positive or negative sentences about a target topic, and used their headlines as a summary of the topic. This is the application of an existing extraction-based summarization method to subjective descriptions.

Hu and Liu (2004) summarized customer reviews of a product such as a digital camera. Their summarization method extracts nouns and noun phrases as features of the target product, (e.g., "picture" for a digital camera), and lists positive and negative reviews on a feature-by-feature basis.

The extracted features are sorted according to the frequency with which each feature appears in the reviews. This method allows users to browse the reviews in terms of important features of the target product.

Liu et al. (2005) enhanced the above method to allow users to compare different products within a specific category, on a feature-by-feature basis.

3 Overview of OpinionReader

Figure 1 depicts the process flow in OpinionReader. The input is a set of subjective descriptions for a specific topic, classified according to their polarity. We assume that processes (1)–(3) in Section 1 are completed, either manually or automatically, prior to the use of our system. It is often the case that users post their opinions and state their standpoints, as exemplified by the websites used in our experiments (see Section 5).

While our primarily target is a set of opinions for a debatable issue classified into pros and cons, a set of customer reviews for a product, classified as positive or negative, can also be submitted.

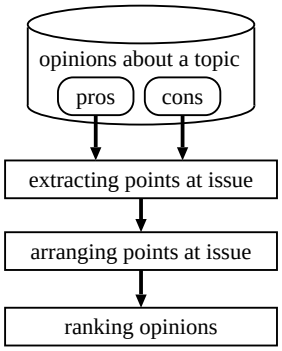

Figure 1: Process flow in OpinionReader.

Our purpose is to visualize the pro and con arguments about a target topic, so that a user can determine which standpoint is the more reasonable.

We extract "points at issue" from the opinions and arrange them in a two-dimensional space. We also rank the opinions that include each point at issue according to their importance, so that a user can selectively read representative opinions on a point-by-point basis.

The output is presented via a graphical interface as shown in Figure 2, which is an example output for the topic "privatization of hospitals by joint-stock companies". The opinions used for this example are extracted from the website for "BS debate"[1]. This interface is accessible via existing Web browsers.

In Figure 2, the x and y axes correspond to the polarity and importance respectively, and each oval denotes an extracted point at issue, such as "information disclosure", "health insurance", or "medical corporation".

Users can easily see which points at issue are most important from each standpoint. Points at issue that are important and closely related to one particular standpoint are usually the most useful in users' decision making.

By clicking on an oval in Figure 2, users can read representative opinions corresponding to that

[1] http://www.nhk.or.jp/bsdebate/

point at issue. In Figure 3, two opinions that include "information disclosure" are presented. The opinions on the right and left sides are selected from the pros and cons, respectively. While the pros support information disclosure, the cons insist that they have not recognized its necessity.

As a result, users can browse the pro and con arguments about the topic in detail. However, for some points at issue, only opinions from a single standpoint are presented, because the other side has no argument about that point.

Given the above functions, users can easily summarize the main points and how they are used in arguing about the topic in support of one standpoint or the other.

If subjective descriptions are classified into more than two categories with a single axis, we can incorporate these descriptions into our system by reclassifying them into just two categories. Figure 4 is an example of summarizing reviews with a multipoint scale rating. We used reviews with five-point star rating for the movie "Star Wars: Episode III"[2]. We reclassified reviews with 1–3 stars as cons, and reviews with 4–5 stars as pros.

In Figure 4, the points at issue are typical words used in the movie reviews (e.g. "story"), the names of characters (e.g. "Anakin", "Obi-Wan", and "Palpatine"), concepts related to Star Wars (e.g. "battle scene" and "Dark Side"), and comparisons with other movies (e.g., "War of the Worlds").

Existing methods for summarizing opinions (Hu and Liu, 2004; Liu et al., 2005). extract the features of a product, which corresponds to the points at issue in our system, and arrange them along a single dimension representing the importance of features. The reviews corresponding to each feature are not ranked.

However, in our system, features are arranged to show how the feature relates to each polarity. The opinions addressing a feature are ranked according to their importance. We target both opinions and reviews, as shown in Figures 2 and 4, respectively.

4 Methodology

4.1 Extracting Points at Issue

In a preliminary investigation of political opinions on the Web, we identified that points at issue can be different language units: words, phrases, sentences, and combinations of sentences. We currently target nouns, noun phrases, and verb phrases, whereas existing summarization methods (Hu and Liu, 2004; Liu et al., 2005) extract only nouns and noun phrases.

Because Japanese sentences lack lexical segmentation, we first use ChaSen[3] to perform a morphological analysis of each input sentence. As a result, we can identify the words in the input and their parts of speech.

To extract nouns and noun phrases, we use handcrafted rules that rely on the word and part-of-speech information. We extract words and word sequences that match these rules. To standardize among the different noun phrases that describe the same content, we paraphrase specific types of noun phrases.

To extract verb phrases, we analyze the syntactic dependency structure of each input sentence, by using CaboCha[4]. We then use handcrafted rules to extract verb phrases comprising a noun and a verb from the dependency structure.

It is desirable that the case of a noun (i.e., postpositional particles) and the modality of a verb (i.e., auxiliaries) are maintained. However, if we were to allow variations of case and modality, verb phrases related to almost the same meaning would be regarded as different points at issue and thus the output of our system would contain redundancy. Therefore, for the sake of conciseness, we currently discard postpositional particles and auxiliaries in verb phrases.

4.2 Arranging Points at Issue

In our system, the points at issue extracted as described in Section 4.1 are arranged in a two-dimensional space, as shown in Figure 2. The x-axis corresponds to the polarity of the points at issue, that is the degree to which a point is related to each standpoint. The y-axis corresponds to the importance of the points at issue.

For a point at issue A, which can be a noun, noun phrase, or verb phrase, the x-coordinate, x_A, is calculated by Equation (1):

$$x_A = P(pro|A) - P(con|A) \qquad (1)$$

$P(S|A)$, in which S denotes either the pro or con standpoint, is the probability that an opinion randomly selected from a set of opinions addressing

[2] http://moviessearch.yahoo.co.jp/detail?ty=mv&id=321602

[3] http://chasen.naist.jp/hiki/ChaSen/

[4] http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

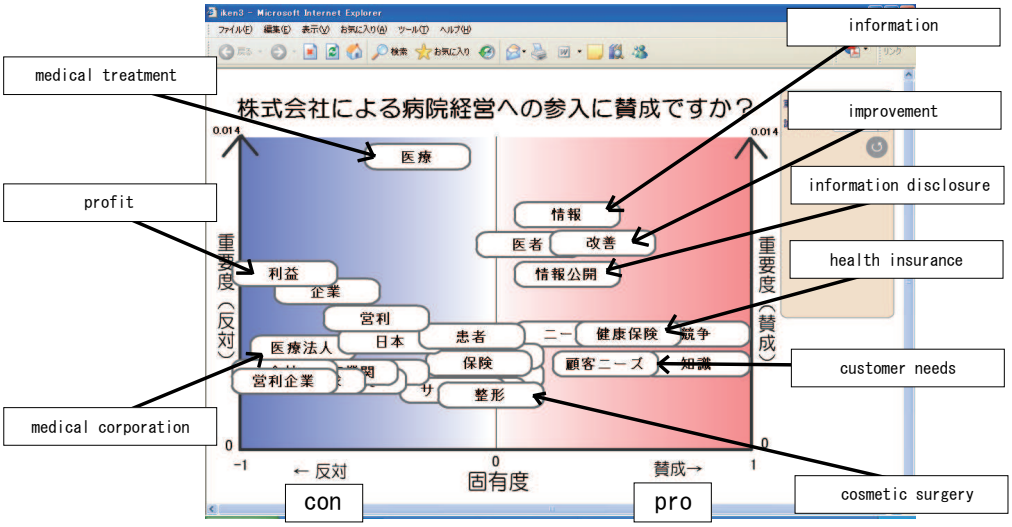

Figure 2: Example of visualizing points at issue for “privatization of hospitals by joint-stock companies”.

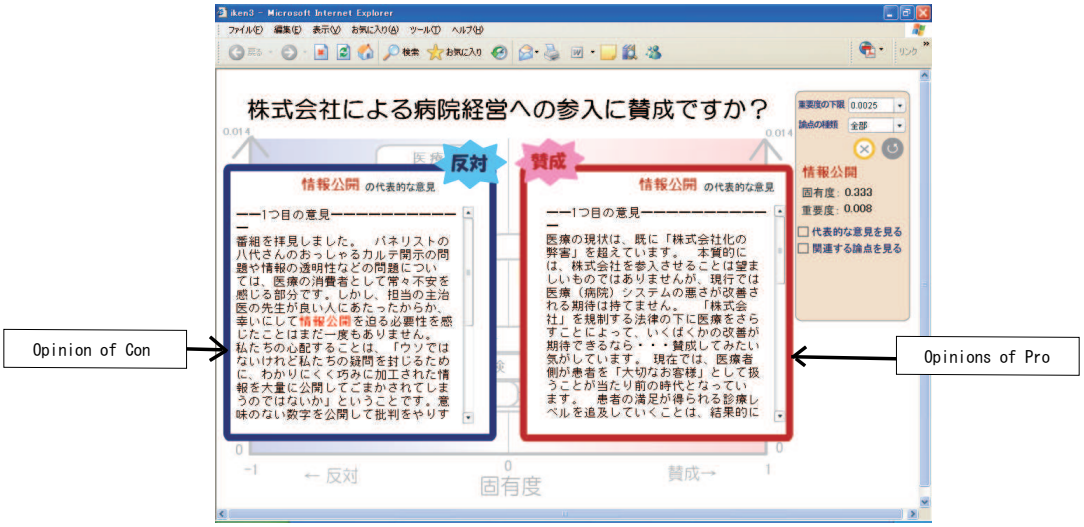

Figure 3: Example of presenting representative opinions for “information disclosure”.

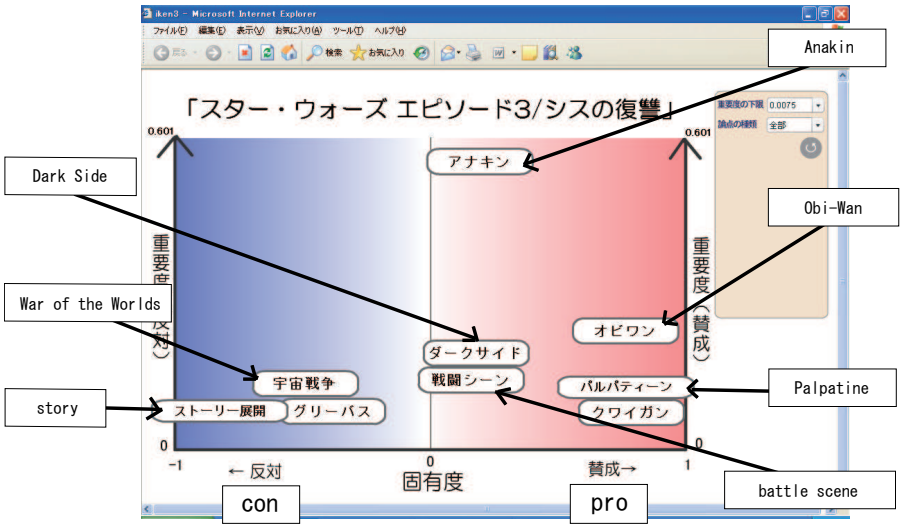

Figure 4: Example of summarizing reviews with multipoint scale rating for “Star Wars: Episode III”.

A supports S. We calculate $P(S|A)$ as the number of opinions that are classified into S and that include A, divided by the number of opinions that include A.

x_A ranges from -1 to 1. A is classified into one of the following three categories depending on the value of x_A:

- if A appears in the pros more frequently than in the cons, x_A is a positive number,
- if A appears in the pros and cons equally often, x_A is zero,
- if A appears in the cons more frequently than in the pros, x_A is a negative number.

The calculation of the y-coordinate of A, y_A depends on which of the above categories applies to A. If A appears in standpoint S more frequently than in its opposite, we define y_A as the probability that a point at issue randomly selected from the opinions classified into S is A.

We calculate y_A as the frequency of A in the opinions classified into S, divided by the total frequencies of points at issue in the opinions classified into S. Thus, y_A ranges from 0 to 1.

However, if A appears in the pros and cons equally often, we use the average of the values of y_A for both standpoints.

General words, which are usually high frequency words, tend to have high values for y_A. Therefore, we discard the words whose y_A is above a predefined threshold. We empirically set the threshold at 0.02.

Table 1 shows example points at issue for the topic "privatization of hospitals by joint-stock companies" and their values of x_A and y_A. In Table 1, points at issue, which have been translated into English, are classified into the three categories (i.e., pro, neutral, and con) according to x_A and are sorted according to y_A in descending order, for each category.

In Table 1, "improvement" is the most important in the pro category, and "medical corporation" is the most important in the con category. In the pro category, many people expect that the quality of medical treatment will be improved if joint-stock companies make inroads into the medical industry. However, in the con category, many people are concerned about the future of existing medical corporations.

Table 1: Examples of points at issue and their coordinates for "privatization of hospitals by joint-stock companies".

Point at issue	x_A	y_A
improvement	0.33	9.2×10^{-3}
information disclosure	0.33	7.9×10^{-3}
health insurance	0.60	5.3×10^{-3}
customer needs	0.50	3.9×10^{-3}
cosmetic surgery	0.00	2.6×10^{-3}
medical corporation	-0.69	4.4×10^{-3}
medical institution	-0.64	3.6×10^{-3}
medical cost	-0.60	3.2×10^{-3}
profit seeking	-0.78	3.2×10^{-3}

4.3 Ranking Opinions

Given a set of opinions from which a point at issue has been extracted, our purpose now is to rank the opinions in order of importance. We assume that representative opinions contain many content words that occur frequently in the opinion set. In our case, content words are nouns, verbs, and adjectives identified by morphological analysis.

We calculate the score of a content word w, $s(w)$, as the frequency of w in the opinion set. We calculate the importance of an opinion by the sum of $s(w)$ for the words in the opinion. However, we normalize the importance of the opinion by the number of words in the opinion because long opinions usually include many words.

5 Experiments

5.1 Method

The effectiveness of our system should be evaluated from different perspectives. First, the effectiveness of each component of our system should be evaluated. Second, the effectiveness of the system as a whole should be evaluated. In this second evaluation, the evaluation measure is the extent to which the decisions of users can be made correctly and efficiently.

As a first step in our research, in this paper we perform only the first evaluation and evaluate the effectiveness of the methods described in Section 4. We used the following Japanese websites as the source of opinions, in which pros and cons are posted for specific topics.

(a) BS debate[5]

(b) ewoman[6]

[5] http://www.nhk.or.jp/bsdebate/

[6] http://www.ewoman.co.jp/

(c) Official website of the prime minister of Japan and his cabinet[7]

(d) Yomiuri online[8]

For evaluation purposes, we collected the pros and cons for five topics. Table 2 shows the five topics, the number of opinions, and the sources. For topic #4, we used the opinions collected from two sources to increase the number of opinions.

In Table 2, the background of topic #5 should perhaps be explained. When using escalators, it is often customary for passengers to stand on one side (either left or right) to allow other passengers to walk past them. However, some people insist that walking on escalators, which are moving stairs, is dangerous.

Graduate students, none of who was an author of this paper, served as assessors, and produced reference data. The output of a method under evaluation was compared with the reference data.

For each topic, two assessors were assigned to enhance the degree of objectivity of the results. Final results were obtained by averaging the results over the assessors and the topics.

5.2 Evaluation of Extracting Points at Issue

For each topic used in the experiments, the assessors read the opinions from both standpoints and extracted the points at issue. We defined the point at issue as the grounds for an argument. We did not restrict the form of the points at issue. Thus, the assessors were allowed to extract any continuous language units, such as words, phrases, sentences, and paragraphs, as points at issue.

Because our method is intended to extract points at issue exhaustively and accurately, we used recall and precision as evaluation measures for the extraction.

Recall is the ratio of the number of correct answers extracted automatically to the total number of correct answers. Precision is the ratio of the number of correct answers extracted automatically to the total number of points at issue extracted automatically.

Table 3 shows the results for each topic, in which "System" denotes the number of points at issue extracted automatically. In Table 3, "C", "R", and "P" denote the number of correct answers, recall, and precision, respectively, on an assessor-by-assessor basis.

[7] http://www.kantei.go.jp/

[8] http://www.yomiuri.co.jp/komachi/forum/

Looking at Table 3, we see that the results can vary depending on the topic and the assessor. However, recall and precision were approximately 50% and 4%, respectively, on average.

The ratio of agreement between assessors was low. When we used the points at issue extracted by one assessor as correct answers and evaluated the effectiveness of the other assessor in the extraction, the recall and precision ranged from 10% to 20% depending on the topic. To increase the ratio of agreement between assessors, the instruction for assessors needs to be revised for future work.

This was mainly because the viewpoint for a target topic and the language units to be extracted were different, depending on the assessor. Because our automatic method extracted points at issue exhaustively, the recall was high and the precision was low, irrespective of the assessor.

The ratios of noun phrases (including nouns) and verb phrases to the number of manually extracted points at issue were 78.5% and 2.0%, respectively. Although the ratio for verb phrases is relatively low, extracting both noun and verb phrases is meaningful.

The recalls of our method for noun phrases and verb phrases were 60.0% and 44.3%, respectively. Errors were mainly due to noun phrases that were not modeled in our method, such as noun phrases that include a relative clause.

5.3 Evaluation of Arranging Points at Issue

As explained in Section 4.2, in our system the points at issue are arranged in a two-dimensional space. The x and y axes correspond to the polarity and the importance of points at issue, respectively.

Because it is difficult for the assessors to judge the correctness of coordinate values in the two-dimensional space, we evaluated the effectiveness of arranging points at issue indirectly.

First, we evaluated the effectiveness of the calculation for the y-axis. We sorted the points at issue, which were extracted automatically (see Section 5.2), according to their importance. We evaluated the trade-off between recall and precision by varying the threshold of y_A. We discarded the points at issue whose y_A is below the threshold.

Note that while this threshold was used to determine the lower bound of y_A, the threshold explained in Section 4.2 (i.e., 0.02) was used to determine the upper bound of y_A and was used consistently irrespective of the lower bound threshold.

Table 2: Topics used for experiments.

Topic ID	Topic	#Opinions Pro	#Opinions Con	Source
#1	principle of result in private companies	57	29	(a)
#2	privatization of hospitals by joint-stock companies	27	44	(a)
#3	the summertime system in Japan	14	17	(b)
#4	privatization of postal services	28	20	(b), (c)
#5	one side walk on an escalator	29	42	(d)

Table 3: Recall and precision of extracting points at issue (C: # of correct answers, R: recall (%), P: precision (%)).

		Assessor A			Assessor B		
Topic ID	System	C	R	P	C	R	P
#1	1968	194	58.2	5.7	101	44.6	2.3
#2	1864	66	50.0	1.8	194	60.8	6.3
#3	508	43	48.8	4.1	43	60.5	5.1
#4	949	77	64.9	5.3	96	36.5	3.7
#5	711	91	30.0	3.8	75	18.7	2.0

Table 4 shows the results, in which the precision was improved to 50% by increasing the threshold. In Figure 2, users can change the threshold of importance by using the panel on the right side to control the number of points at issue presented in the interface. As a result, users can choose appropriate points at issue precisely.

Second, we evaluated the effectiveness of the calculation for the x-axis. We evaluated the effectiveness of our method in a binary classification. For each point at issue extracted by an assessor, the assessor judged which of the two standpoints the point supports.

If a point at issue whose x-coordinate calculated by our method is positive (or negative), it was classified as pro (or con) automatically. We did not use the points at issue whose x-coordinate was zero for evaluation purposes.

Table 5 shows the results. While the number of target points at issue was different depending on the topic and the assessor, the difference in classification accuracy was marginal.

For each topic, we averaged the accuracy determined by each assessor and averaged the accuracies over the topic, which gave 95.6%. Overall, our method performs the binary classification for points at issue with a high accuracy.

Errors were mainly due to opinions that included arguments for both standpoints. For example, a person supporting a standpoint might suggest that he/she would support the other side under a specific condition. Points at issue classified incorrectly had usually been extracted from such contradictory opinions.

5.4 Evaluation of Ranking Opinions

To evaluate the effectiveness of our method in ranking opinions on a point-by-point basis, we used a method that sorts the opinions randomly as a control. We compared the accuracy of our method and that of the control. The accuracy is the ratio of the number of correct answers to the number of opinions presented by the method under evaluation.

For each point at issue extracted by an assessor, the assessor assigned the opinions to one of the following degrees:

- A: the opinion argues about the point at issue and is represented,
- B: the opinion argues about the point at issue but is not represented,
- C: the opinion includes the point at issue but does not argue about it.

We varied the number of top opinions presented by changing the threshold for the rank of opinions.

Table 6 shows the results, in which N denotes the number of top opinions presented. The column "Answer" refers to two cases: the case in which only the opinions assigned to "A" were regarded as correct answers, and the case in which the opinions assigned to "A" or "B" were regarded as correct answers. In either case, our method outperformed the control in ranking accuracy.

Although the accuracy of our method for "A" opinions was low, the accuracy for "A" and "B"

Table 4: Trade-off between recall and precision in extracting points at issue.

Threshold	0	0.002	0.004	0.006	0.008	0.010
Recall	0.48	0.17	0.11	0.04	0.03	0.02
Precision	0.04	0.14	0.21	0.31	0.33	0.50

Table 5: Accuracy for classifying points at issue.

	Assessor A		Assessor B	
Topic ID	#Points	Accuracy (%)	#Points	Accuracy (%)
#1	113	98.2	45	97.7
#2	33	91.0	118	94.1
#3	21	95.2	26	100
#4	50	92.0	35	91.4
#5	27	96.3	14	100

Table 6: Accuracy of ranking opinions.

Answer	Method	$N = 1$	$N = 2$	$N = 3$
A	Random	19%	28%	19%
	Ours	38%	32%	23%
A+B	Random	81%	83%	75%
	Ours	87%	87%	83%

opinions was high. This suggests that our method is effective in distinguishing opinions that argue about a specific point and opinions that include the point but do not argue about it.

6 Conclusion

In aiming to support users' decision making, we have proposed a method for summarizing and visualizing the pro and con arguments about a topic.

Our prototype system, called "OpinionReader", extracts points at issue from the opinions for both pro and con standpoints, arranges the points in a two-dimensional space, and allows users to read important opinions on a point-by-point basis. We have experimentally evaluated the effectiveness of the components of our system.

Future work will include evaluating our system as a whole, and summarizing opinions that change over time.

References

Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. 2004. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 264–271.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

Lun-Wei Ku, Li-Ying Lee, Tung-Ho Wu, and Hsin-Hsi Chen. 2005. Major topic detection and its application to opinion summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627–628.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of the 14th International World Wide Web Conference*, pages 324–351.

Tomoyuki Nanno, Toshiaki Fujiki, Yasuhiro Suzuki, and Manabu Okumura. 2004. Automatically collecting, monitoring, and mining Japanese weblogs. In *The 13th International World Wide Web Conference*, pages 320–321. (poster session).

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 264–271.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.