Japanese Opinion Extraction System for Japanese Newspapers Using Machine-Learning Method

Toshiyuki Kanamaru, Masaki Murata, and Hitoshi Isahara
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{kanamaru, murata, isahara}@nict.go.jp

Abstract

We constructed a Japanese opinion extraction system for Japanese newspaper articles using a machine-learning method for the system. We used opinion-annotated articles as learning data for the machine-learning method. The system extracts opinionated sentences from newspaper articles, and specifies opinion holders and opinion polarities of the extracted sentences. The system also evaluates whether or not the sentences of the articles are relevant to the given topic. We conducted experiments using the NTCIR-6 opinion extraction subtask data collection and obtained the following accuracy rates using a lenient gold standard: opinion extraction, 42.88%; opinion holder extraction, 14.31%; polarity decision, 19.90%; and relevance evaluation, 63.15%.

Keywords: *opinion extraction, machine-learning method, newspaper articles*

1 Introduction

Recently, opinion extraction and sentiment analysis has been receiving a lot of attention in the field of natural language processing [4, 14]. As well as automatically detecting sentences in which an opinion is expressed [19], automatically judging the polarity of the opinion [12, 2, 18, 7] and specifying who holds that opinion (opinion sources) [16, 15, 6, 1] are topics now receiving more attention in the research community. This points to a need for more detailed opinion extraction.

We constructed a Japanese opinion extraction system for Japanese Newspaper articles. We used a machine-learning method for the system and opinion-annotated articles as learning data. The system extracts opinionated sentences from newspaper articles and specifies the opinion holders and opinion polarities of the extracted opinionated sentences. The system also judges whether the sentences of the articles are relevant to the given topic or not.

To evaluate our system, we took part in an opinion analysis pilot task at NTCIR-6. The pilot task had opinion extraction and application-oriented subtasks, each in three languages: Japanese, English, and Chinese. We participated in the Japanese opinion extraction subtask. We did experiments using a Japanese data collection distributed by the task organizers, and verified the performance of our system.

As a result, we obtained the following accuracy rates using a lenient gold standard: opinion extraction was 42.88%, opinion holder extraction was 14.31%, polarity determination was 19.90%, and relevance judgment was 63.15%.

2 Data set for this study

In this study, we used a Japanese test collection distributed in the opinion analysis pilot task at NTCIR-6 as a data set. The test collection format is described in more detail in the “Overview of Opinion Analysis Pilot Task at NTCIR-6” [13], described by task organizers.

In this data set, there were 30 topics, and a maximum of 20 documents were selected for each topic. There were 490 documents in the Japanese test collection, with 15,279 sentences. Of the 30 topics, 4 were provided as sample data, and the remaining 26 were used as open data for experiments and evaluations.

Four annotation categories such as opinionated sentences, opinion holders, relevant sentences, and opinion polarities were annotated to each sentence in the documents in the Japanese test collection. Opinionated sentences and relevant sentences had a binary value, and there were three values for opinionated polarities: positive (POS), negative (NEG), and neutral (NEU). The values of opinion holders were string and multiple.

3 Opinion extraction system used in this study

Our system consisted of four components: 1) an opinion extraction component, 2) a component to specify the opinion holder, 3) a component to judge the relevance to the given topic, and 4) a component to classify the polarities of the opinion sentences.

The system classifies an input sentence in the opinion extraction and polarity classification components using only the machine-learning method. We used a support vector machine as the basis of our machine-learning method because support vector machines are more effective than other methods in many research areas [9, 17, 11]. We used the sample data of the Japanese test collection as learning data. Each sentence in the documents had four kinds of annotation tags, and we used three kinds of annotation: opinionated sentences, opinion holders, and opinionated polarities.

In the polarity classification component, the system outputs three values, positive (POS), negative (NEG), and neutral (NEU). However, the support vector machines are only capable of handling data consisting of two categories. Data consisting of more than two categories is generally handled using the pair-wise method [9].

Pairs from two different categories (N(N-1)/2 pairs) were constructed for data consisting of N categories with this method. The best category was determined by using a two-category classifier (in this study, a support vector machine[1] was used as the two-category classifier), and the correct category was finally determined on the basis of "voting" on the N(N-1)/2 pairs that resulted from the analysis with the two-category classifier.

In the opinion holder specification component, the system specifies the opinion holder of the opinionated sentence using the machine-learning method and heuristics rules.

In the relevance to the given topic evaluation component, the system judges whether the sentences of the articles are relevant to the given topic or not by calculating the score of words that are included in each sentence.

3.1 Features (information used in classification)

The features the system used with the machine-learning method were different for each component.

In the opinion extraction component and the opinion holder specification component, the system used words and 1-gram to 10-gram strings at the ends of input sentences as the features. The words we used were only adjectives, adverbs, conjunctions, auxiliary verbs, and postfix verbs. We used the Japanese morphological analyzer, Chasen [10] to identify the parts of speech of the words.

The 1-gram to 10-gram strings features were constructed based on the characteristics of Japanese sentences. The Japanese sentences have a tendency to express the opinion using modality expression. Modality expression in Japanese sentences is often indicated by the verbs and the auxiliary verbs at the ends of sentences. The structure of the Japanese language is subject-object-verb (SOV), so verb phrases appear at the ends of sentences. Therefore, the strings at the ends of sentences were used as features.

In the component to classify polarities of the opinion sentences, the system used three kinds of features; words, the first 1, 2, 3, 4, 5, or 7 digits of the category numbers of words, and 1-gram to 10-gram strings at the ends of input sentences.

The reason why we used the 1-gram to 10-gram strings at the ends of input sentences as the features is the same as the opinion extraction component. The category number of word indicates a semantic class of words. A Japanese thesaurus, the *Bunrui Goi Hyou* [3], was used to determine the category number of each word. This thesaurus is 'is-a' hierarchical, in which each word has *a category number*. This is a 10-digit number that indicates seven levels of 'is-a' hierarchy. The top five levels are expressed by the first five digits, the sixth level is expressed by the next two digits, and the seventh level is expressed by the last three digits.

3.2 Specifying the opinion holder

In the opinion holder specification component, the system specifies the opinion holder in the following way.

First, the system uses the machine-learning method to classify sentences as opinionated, i.e., sentences in which an opinion is held by the writer or someone else. Next, the system handles only the sentences classified as having an opinion held by someone other than the writer.

The system extracts terms that are possible to be opinion holders, such as nouns related to human beings or organizations, human names, and organization names. If the sentence contains a quote and an expression such as "*to itta*" (said that), the system outputs the nearest extracted terms, except for the quoted part, as an opinion holder. If nothing is quoted in the sentence, the system outputs the nearest extracted terms, including the sentence, as an opinion holder.

[1] We used Kudoh's TinySVM software [8] as the support vector machine.

Table 1. Experimental results of opinion extraction

Standard	Answer	Proposed Answer	Correct Answer	Precision	Recall	F-measure
Lenient	2,974	1,397	937	67.07%	31.51%	42.88%
Strict	2,191	1,397	762	54.55%	34.78%	42.48%

Table 2. Experimental results of opinion holder extraction

Standard	Answer	Proposed Answer	Correct Answer	Precision	Recall	F-measure
Lenient	462	2,311	850	23.78%	10.23%	14.31%
Strict	893	1,177	431	13.25%	11.01%	12.03%

Table 3. Experimental results of relevance judgment

Threshold	Standard	Answer	Proposed Answer	Correct Answer	Precision	Recall	F-measure
0.01	Lenient	6,420	7,188	4,297	59.78%	66.93%	63.15%
	Strict	4,880	7,188	3,381	47.04%	69.28%	56.03%
0.05	Lenient	6,420	4,152	2,674	64.40%	41.65%	50.58%
	Strict	4,880	4,152	2,178	52.46%	44.63%	48.23%

Table 4. Experimental results of polarity decision

Standard	Answer	Proposed Answer	Correct Answer	Precision	Recall	F-measure
Lenient	2,793	1,397	417	29.85%	14.93%	19.90%
Strict	1,558	1,397	234	16.75%	15.02%	15.84%

3.3 Calculating the weight of words

In the component where the relevance to the given topic is judged, the system calculates the score of the words.

The sentence scores are calculated as follows.

We first extract terms in which the parts of speech are nouns and unknown words from the topic description and topic relevance fields[2] using the Japanese morphological analyzer, Chasen. Then, the score is calculated from the equation

$$Weight(s) = \sum_{\mathrm{term}t} log \frac{N}{df(t)} \tag{1}$$

$$Score(s) = \frac{Weight(s)}{Weight(top)} \tag{2}$$

where s is a sentence, t is a term extracted from the topic description and topic relevance fields, $df(t)$ is the number of documents in which t appears, N is the total number of documents, and $Weight(top)$ is the highest weight in the sentences included in each topic.

Finally, if the $Score(s)$ is higher than a threshold, the system judges that the sentence is relevant to the topic; otherwise, the system judges that the sentence is not relevant to the topic. We used two thresholds: 0.01 and 0.05.

4 Experiments

The experimental results are listed in Tables 1 to 4. Tables 5 to 7 show the topic-by-topic results.

There were 26 topics in the experiments as well as two standards of evaluation. One evaluation used a strict gold standard, and the other used a lenient gold standard. Since all sentences were annotated by three assessors, the strict gold standard was that all three assessors had to have the same annotation, and the lenient gold standard was that two of the three assessors had to have the same annotation[3].

Table 1 lists the precision, recall, and f-measure for the extraction of opinionated sentences. There were 2,191 opinionated sentences in the strict gold standard and 2,974 opinionated sentences in the lenient gold standard. Our system output 1,397 sentences as opinionated sentences, and obtained an accuracy rate of 42.88% in the lenient gold standard evaluation. Table 5 lists the precision, recall, and f-measure for the extraction of opinionated sentences topic-by-topic. Our system obtained highly accuracy rate of 57.89% in the

[2] See [13] for more details of topic description and topic relevance fields.

[3] See [13] for more details of the evaluation.

Table 5. Experimental results of opinion extraction (topic-by-topic)

	Strict			Lenient		
Topic	Precision	Recall	F-value	Precision	Recall	F-value
004	31.58%	54.55%	40.00%	57.89%	57.89%	57.89%
005	58.00%	34.32%	43.12%	65.00%	29.41%	40.50%
006	40.00%	40.00%	40.00%	40.00%	22.22%	28.57%
007	67.86%	17.92%	28.35%	75.00%	18.58%	29.78%
008	50.00%	10.87%	17.86%	70.00%	10.77%	18.67%
009	41.25%	51.56%	45.83%	68.75%	46.22%	55.28%
010	55.00%	40.00%	46.32%	75.00%	33.33%	46.15%
011	37.84%	43.75%	40.58%	54.05%	40.82%	46.51%
012	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
014	67.07%	31.43%	42.80%	74.39%	28.11%	40.80%
015	61.07%	33.46%	43.23%	66.44%	31.53%	42.77%
016	41.67%	25.77%	31.85%	48.33%	23.77%	31.87%
017	45.00%	25.00%	32.14%	55.00%	19.30%	28.57%
018	38.98%	44.23%	41.44%	66.10%	39.39%	49.36%
019	61.79%	49.03%	54.68%	73.17%	45.23%	55.90%
020	42.86%	27.27%	33.33%	59.18%	21.17%	31.18%
021	60.00%	36.45%	45.35%	76.92%	34.97%	48.08%
022	60.44%	45.83%	52.13%	67.03%	39.10%	49.39%
023	37.50%	27.27%	31.58%	50.00%	22.22%	30.77%
024	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
027	64.00%	32.65%	43.24%	78.67%	28.64%	41.99%
028	65.15%	32.82%	43.65%	72.73%	29.63%	42.11%
029	58.33%	17.95%	27.45%	79.17%	18.45%	29.93%
030	61.70%	38.67%	47.54%	76.60%	34.29%	47.37%
031	45.79%	46.23%	46.01%	56.07%	41.96%	48.00%
032	8.33%	12.50%	10.00%	8.33%	7.69%	8.00%

lenient gold standard evaluation at topic 004. However, our system failed to extract correct oponinated sentences at topic 012 and 024.

Table 2 indicates the precision, recall, and f-measure for the specification of opinion holder of the opinionated sentences. There were 1,680 opinion holders in the opinionated sentences in the strict gold standard and 3,245 opinion holders in the opinionated sentences in the lenient gold standard. The number of opinion holders in the lenient gold standard was higher than the number of opinionated sentences. Our system output 1,396 opinion holders for the opinionated sentences, and obtained an accuracy rate of 14.31% in the lenient gold standard evaluation.

Table 3 indicates the precision, recall, and f-measure for the judgments of relevance to the topics. Using the strict gold standard, 4,880 sentences were relevant to the topics, whereas 6,420 sentences were relevant in the lenient gold standard. Our system output 7,188 sentences when the threshold was 0.01, and 4,152 sentences when the threshold was 0.05. Using 0.01 as a threshold, our system obtained an accuracy rate of 50.58% in the lenient gold standard evaluation. Table 6 lists the precision, recall, and f-measure for the judgments of relevance to the topics topic-by-topic. Our system obtained highly accuracy rate in the lenient gold standard evaluation at almost all topics except topic 023, 031, and 032.

Table 4 gives the precision, recall, and f-measure for the classification of polarity of opinionated sentences. The polarities of 1,558 and 2,793 opinionated sentences were respectively defined in the strict and lenient gold standards. Our system output the polarities of 1,397 opinionated sentences and obtained an accuracy rate of 19.90% in the lenient gold standard evaluation. Table 7 lists the precision, recall, and f-measure for the classification of polarity of opinionated sentences topic-by-topic. We can easily see the difference between the accuracy rate in the strict gold standard and one in the lenient gold standard. This suggestions that our system could correctly classify the polarity of opinionated sentence whose polarity is unclear, but failed to classify the polarity of opinionated sentence whose polarity is clear.

We found out the following details from the tables.

- Our system did not obtain high scores in opinion extraction. This result affected the holder extraction and polarity decision. The system output the opinion holder and polarity of opinion only

Table 6. Experimental results of relevance judgment (topic-by-topic)

	Strict			Lenient		
Topic	Precision	Recall	F-value	Precision	Recall	F-value
004	47.41%	93.22%	62.85%	62.07%	94.74%	75.00%
005	38.62%	67.35%	49.09%	43.34%	67.28%	52.72%
006	73.47%	97.30%	83.72%	75.51%	97.37%	85.06%
007	55.93%	70.21%	62.26%	76.69%	64.64%	70.15%
008	59.24%	68.68%	63.61%	65.40%	59.74%	62.44%
009	78.18%	89.79%	83.58%	83.84%	89.06%	86.37%
010	40.69%	81.38%	54.25%	56.72%	77.59%	65.53%
011	51.35%	90.48%	65.52%	70.27%	86.67%	77.61%
012	50.85%	100.00%	67.42%	57.63%	97.14%	72.34%
014	34.27%	88.74%	49.45%	36.06%	84.94%	50.63%
015	18.24%	49.56%	26.67%	26.71%	48.81%	34.53%
016	51.44%	69.42%	59.09%	68.71%	65.86%	67.25%
017	28.64%	96.72%	44.19%	51.94%	88.43%	65.44%
018	33.89%	90.99%	49.39%	40.94%	91.73%	56.61%
019	76.57%	85.04%	80.58%	91.65%	81.40%	86.22%
020	49.85%	70.95%	58.56%	65.60%	66.37%	65.98%
021	64.46%	67.79%	66.08%	72.40%	64.37%	68.15%
022	43.73%	74.58%	55.13%	65.69%	75.28%	70.16%
023	15.19%	85.71%	25.81%	16.46%	86.67%	27.67%
024	64.29%	85.71%	73.47%	78.57%	75.86%	77.19%
027	22.06%	80.15%	34.60%	38.24%	72.51%	50.07%
028	19.79%	45.24%	27.53%	29.17%	50.00%	36.84%
029	39.18%	55.37%	45.89%	84.21%	53.14%	65.16%
030	41.50%	44.85%	43.11%	48.30%	40.57%	44.10%
031	37.50%	0.74%	1.45%	37.50%	0.66%	1.30%
032	5.56%	25.00%	9.10%	5.56%	25.00%	9.10%

for sentences judged by the system to contain an opinion. In fact, the precision of holder extraction (23.78%) and polarity decision (29.85%) was much higher than the recall of holder extraction (10.23%) and polarity decision (14.93%) in the lenient gold standard.

- The 0.01 threshold obtained higher accuracy (63.15%) than that obtained by the 0.05 threshold (50.58%) in the relevance judgment in the lenient gold standard, and our system obtained high scores in relevance judgment. This output was independent of the result of the opinion extraction, so the recall of this is higher than others.

5 Conclusions

We constructed a Japanese opinion extraction system for Japanese newspaper articles using a machine-learning method. We used opinion-annotated articles as learning data for the machine-learning method. The system extracts opinionated sentences from newspaper articles and specifies opinion holders and opinion polarities of the extracted opinionated sentences. The system also determines whether the sentences of the articles are relevant to the given topic or not. We conducted experiments using the NTCIR-6 opinion extraction subtask data collection and obtained the following accuracy rates using a lenient gold standard: opinion extraction, 42.88%; holder extraction, 14.31%; polarity decision, 19.90%; and relevance judgment, 63.15%.

Targets for our future work include trying to increase the recall of opinionated sentences using various linguistic resources. For example, we constructed an adverb dictionary that relates to speaker attitudes [5], that we will use in our next experiments on opinion extraction.

Acknowledgements

We are grateful to all of the organizers of NTCIR-6, who gave us a chance to participate in the NTCIR-6 contest to improve and examine our opinion extraction system. We greatly appreciate the kindness of all those who helped us.

Table 7. Experimental results of polarity decision (topic-by-topic)

	Strict			Lenient		
Topic	Precision	Recall	F-value	Precision	Recall	F-value
004	0.00%	0.00%	0.00%	26.32%	83.33%	40.00%
005	15.00%	12.71%	13.76%	29.00%	24.58%	26.61%
006	0.00%	0.00%	0.00%	40.00%	100.00%	57.14%
007	10.71%	4.29%	6.13%	32.14%	12.86%	18.37%
008	10.00%	2.86%	4.45%	20.00%	5.71%	8.88%
009	11.25%	20.00%	14.40%	23.75%	42.22%	30.40%
010	16.25%	17.11%	16.67%	31.25%	32.89%	32.05%
011	10.81%	18.18%	13.56%	16.22%	27.27%	20.34%
012	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
014	36.59%	20.55%	26.32%	43.90%	24.66%	31.58%
015	20.13%	15.54%	17.54%	28.86%	22.28%	25.15%
016	5.00%	4.48%	4.73%	11.67%	10.45%	11.03%
017	5.00%	5.88%	5.40%	20.00%	23.53%	21.62%
018	10.17%	16.67%	12.63%	30.51%	50.00%	37.90%
019	25.20%	26.27%	25.72%	36.59%	38.14%	37.35%
020	10.20%	9.26%	9.71%	30.61%	27.78%	29.13%
021	21.54%	18.67%	20.00%	36.92%	32.00%	34.28%
022	14.29%	16.88%	15.48%	27.47%	32.47%	29.76%
023	0.00%	0.00%	0.00%	12.50%	33.33%	18.18%
024	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
027	17.33%	11.82%	14.05%	30.67%	20.91%	24.87%
028	16.67%	11.96%	13.93%	34.85%	25.00%	29.11%
029	16.67%	7.02%	9.88%	41.67%	17.54%	24.69%
030	25.53%	21.43%	23.30%	42.55%	35.71%	38.83%
031	14.95%	21.33%	17.58%	24.30%	34.67%	28.57%
032	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

References

[1] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 355–362, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[2] K. Dave, S. Lawrence, and D. M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *the Twelfth International World Wide Web Conference, WWW2003.*, Budapest, Hungary, May 2003.

[3] T. N. I. for Japanese Language. *Bunrui Goi Hyou*. Shuuei Publishing, 1964.

[4] M. Gamon and A. Aue. Proceedings of workshop on sentiment and subjectivity in text at the annual meeting of the Association of Computational Linguistics. Sydney, Australia, July 2006. Association for Computational Linguistics.

[5] T. Kanamaru, M. Murata, and H. Isahara. Construction of Adverb Dictionary that Relates to Speaker Attitudes and Evaluation of Its Effectiveness. In *the Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, pages 295–302, Wuhan, China, November 2006.

[6] S.-M. Kim and E. Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics.

[7] S.-M. Kim and E. Hovy. Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology Conference of the NAACL 2006, Main Conference*, pages 200–207, New York City, USA, June 2006. Association for Computational Linguistics.

[8] T. Kudoh. TinySVM: Support Vector Machines. http://cl.aist-nara.ac.jp/˜taku-ku/software/TinySVM/index.html, 2000.

[9] T. Kudoh and Y. Matsumoto. Use of Support Vector Learning for Chunk Identification. *CoNLL-2000*, pages 142–144, 2000.

[10] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese Morphological Analysis System ChaSen version 2.0 Manual 2nd edition. 1999.

[11] M. Murata, Q. Ma, and H. Isahara. Comparison of Three Machine-Learning Methods for Thai Part-of-Speech Tagging. *ACM Transactions on Asian Language Information Processing*, 1(2):145–158, 2002.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86, 2002.

[13] Y. Seki, D. K. Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin. Overview of Opinion Analysis Pilot Task at NTCIR-6, 2007.

[14] J. G. Shanahan, Y. Qu, and J. Wiebe. *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*. Springer-Verlag, New York, 2005.

[15] V. Stoyanov and C. Cardie. Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 336–344, 2006.

[16] V. Stoyanov and C. Cardie. Toward Opinion Summarization: Linking the Sources. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 9–14, Sydney, Australia, July 2006. Association for Computational Linguistics.

[17] H. Taira and M. Haruno. Feature Selection in SVM Text Categorization. In *Proceedings of AAAI2001*, pages 480–486, 2001.

[18] J. Wiebe and E. Riloff. Creating Subjective and. Objective Sentence Classifiers from Unannotated Texts. In *the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, 2005.

[19] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308, 2004.