Extraction of Opinion Sentences using Machine Learning: Hiroshima City University at NTCIR-7 MOAT

Daisuke Kobayashi Hidetsugu Nanba Toshiyuki Takezawa
Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194 JAPAN
Phone: +81-82-830-1584
kobayashi@nlp.its.hiroshima-cu.ac.jp, {nanba, takezawa}@hiroshima-cu.ac.jp

Abstract

We propose a machine learning-based method for extracting opinion sentences using 13 features including about 760,000 of sentence-final expressions. We submitted two systems to the Japanese Subtask of the MOAT at NTCIR-7 Workshop, and obtained F-values of 0.5615 and 0.3319 using lenient gold standard, and 0.5213 and 0.3561 using strict gold standard, respectively.
Keywords: *Sentence-final expressions, Machine learning, Opinion Extraction.*

1 Introduction

Recently, opinion extraction and sentiment analysis have been considered as a major research topics in the NLP community. We have studied automatic extraction of opinion sentences using the TSC4 corpus, which was created for the purpose of opinion summarization studies [1]. We used machine learning for opinion extraction, and obtained precision of 0.8382 and recall of 0.8184. To confirm the effectiveness of our method using another data set, we participated in the MOAT task at NTCIR-7 [2].

The remainder of this paper is organized as follows. Section 2 explains how to extract trend information from newspaper articles. Section 3 reports on the experiments, and discuss the results. We present some conclusions in Section 4.

2 Opinion Extraction

We used machine learning to build our system. In this section, we describe several features used for machine learning.

2.1 Sentence-final Expressions

For the extraction of opinion sentences, we focused on sentence-final expressions, because useful clues, such as modality, tense, and aspect, appear there. In the past NTCIR workshop, many groups used sentence-final expressions for opinion extraction [3] [4]. Kanamaru et al. [3] applied machine learning method for opinion extraction. They used words and uni-gram to 10-gram strings at the end of sentences as the features for machine learning. Mizuguchi et al. [4] also focused on sentence-final expressions. They prepared two kinds of lists of sentence-final expressions. One is a list of expressions that appeared frequently in 585 opinion sentences frequently, and the other is a list of expressions that appeared in 2,167 non-opinion sentences. Both opinion and non-opinion sentences were provided by the organizers of NTCIR-6. In our work, we prepared several lists of sentence-final expressions, which were created from three corpora: "TSC4", "NTCIR-6", and "News". In the following, we describe the detail of the lists from these corpora.

2.1.1 Lists Produced from TSC4 and NTCIR-6 Corpora

The "TSC4" corpus comprises 12,000 Japanese sentences from the databases of Mainichi, Yomiuri, Asahi, and Nikkei newspapers written in 2003 and 2004. Human assessors annotated opinion tags to each opinion sentence in the corpus. Using the corpus, we created a list of sentence-final expressions by extracting uni-gram, bi-gram, tri-gram, and 4-gram of sentence-final expressions from opinion sentences. Finally, we obtained about 14,000 expressions.

We also prepared two other kinds of lists from "NTCIR-6" corpus [5]. "NTCIR-6" is a corpus distributed at the opinion analysis pilot task in the NTCIR-6 workshop. Our lists were created by extracting uni-gram, bi-gram, tri-gram, and 4-gram of sentence-final expressions from opinion and non-opinion sentences in the corpus. Finally, we obtained 142 expressions for opinion, and 198 for non-opinion, respectively.

The number of expressions from NTCIR-6 is much smaller than that from TSC4. We created both lists manually. We spent a couple of months on creation from the "TSC4" corpus, but we could spend only a few weeks on the "NTCIR-6" corpus.

2.1.2 Lists Produced from News Corpus

In addition to the lists described in Section 2.1.1, we prepared two sentence-final expressions lists using "News" corpus. Generally, objective facts are mentioned in most newspaper articles, while authors' subjective opinions are described in editorials. We therefore collected sentence-final expressions from both newspaper articles and editorials in "News" corpus using the following procedure.

1. Collect newspaper articles on the front page and editorials from News corpus. As a News

corpus, we used Mainichi newspapers published in the 14 years between 1993 and 2006.
2. Extract N-grams of sentence-final expressions from newspaper articles and editorials.
3. Rank the N-grams from both newspaper articles and editorials using cost criteria[1] [6].
4. Re-rank both N-grams using the following equation.
   - New score of an sentence-final expressions in editorials = score of the expression in editorials – score of the expression in newspaper articles
   - New score of an sentence-final expressions in newspaper articles = score of the expression in newspaper articles – score of the expression in editorials

Here, step four degraded sentence-final expressions that appear in both newspaper articles and editorials frequently. Finally, we obtained 351,914 sentence-final expressions for opinion (from editorials), and 402,054 for non-opinion (from newspaper articles).

We summarize the number of sentence-final expressions obtained from each corpus in Table 1.

Table 1. The number of sentence-final expressions

	TSC4	NTCIR-6	News
opinion	14,000	142	351,914
non-opinion		198	402,054

2.2 Other Features

In addition to sentence-final expressions, we used the following features for machine learning.

- **Sentence position:** The position of each sentence in articles or editorials. The values are normalized between 0.0 and 1.0 by dividing by the number of sentences in the article.
- **Sentence length:** The number of characters and words (morphemes)[2] in each sentence.
- **Subjects:** A subject in a sentence.
- **Subject types:** Organization, person's name, or location[3].
- **Tense:** Past, present, or future[4].
- **The number of parentheses:** Whether a sentence contains statements of organizations or person or not.
- **Frequencies of each part of speeches in a sentence:** Mainly used for detecting non-opinion sentences containing only symbols or particular marks.

2.3 Utilization of Features in Surrounding Sentences

From the results of our pilot study, we found that short sentences tend to obtain low precision, because the information contained in such sentences are insufficient. We therefore used features in the surrounding sentences as well as the features in each target sentence[5].

3 Experiments

3.1 Submission

To confirm the effectiveness of our approach, we submitted two systems: "HCU-1" and "HCU-2" to the formal run of the Japanese Subtask at NTCIR-7 MOAT [2]. Both systems used the "NTCIR-6" corpus and a sample data provided by the organizers of the MOAT in advance of the formal run, as training data. Each sentence in the data was annotated by three assessors. If at least two assessors identified a sentence as opinion, the HCU-1 system regarded it as a positive example, while the HCU-2 required all assessors to identify a sentence as opinion for it to be considered positive.

3.2 Data and Evaluation

We used the data provided by MOAT organizers for evaluation. The data consists of 18 topics including 249 documents in total. All the submitted systems were evaluated in terms of precision, recall, and F-measure using two gold standards: lenient (two of three assessors have the same annotation) and strict (all three assessors must have the same annotation) [2].

3.3 Results and Discussions

The experimental results are shown in Table 2. The results showed that HCU-1 obtained higher precisions than HCU-2 in both lenient and strict evaluations. However, the number of opinion sentences extracted by HCU-1 is much smaller than HCU-2.

[1] Simple term-frequency method tends to rank shorter terms higher. To solve this problem, the cost criteria ranks terms using the following equation.

Score(term) = frequency * length of the term

[2] To count the number of words in a sentence, we used a Japanese morphological analyzer MeCab. (http://mecab.sourceforge.net/)

[3] To identify the subject types, we used a Japanese parser CaboCha (http://chasen.org/~taku/software/cabocha/). The parser can be used as a named entity recognizer.

[4] To identify the tense of each sentence, we used a Japanese parser KNP (http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html).

[5] We used YamCha software package, which specializes in text chunking based on TinySVM software (http://chasen.org/~taku/software/yamcha/). We used a polynomial kernel of degree 2, and used a value of 9 for the window size.

Table 2. Experimental result

		P	R	F
Lenient	HCU-1	0.6190	0.5138	0.5615
	HCU-2	0.7754	0.2111	0.3319
Strict	HCU-1	0.4894	0.5577	0.5213
	HCU-2	0.6544	0.2446	0.3561

3.3.1 Typical Errors of Our System

The followings are typical errors of our systems.

(1) The lack of sentence-final expressions (40%)
(2) Over-extraction of opinion sentences (20%)
(3) Titles and headlines (15%)
(4) Others (25%)

In the following, we describe each of these errors.

(1) The lack of sentence-final expressions (40%)

Although we prepared many of sentence-final expressions, some opinion sentences could not be extracted because they lacked sentence-final expressions. One such case follows.

Table 3. Example of expressions that are not contained in our lists

Not contained expressions	Similar expressions in our lists
否めない (cannot deny)	否定できない (cannot deny)
非難する (blame)	責める (blame)

We compared these expressions with our lists, and found that similar expressions are contained in our lists. Recently, several methods to detect synonyms or similar expressions have been proposed, and their effectiveness was confirmed in various applications, such as query expansion for information retrieval [7] and evaluation in text summarization [8]. The low recall by our methods might be improved applying these methods to our lists.

(2) Over-extraction of opinion sentences (20%)

Following three examples were identified as non-opinion sentences by human assessors, while our systems extracted them as opinion sentences, because sentence-final expressions of these sentences are contained in our lists.

[origintal]
同じ気持ちを抱いているに違いない。
[translation]
(They must feel the same way.)

[orignial]
「わが国は事務総長提案に賛同できない
[translation]
("Our country cannot agree with the proposal by a director-of-the-executive-office".)

[orignial]
構造改革を進めるべきだ。
[translation]
(We should promote our structural reform.)

(3) Titles and headlines (15%)

Some assessors annotated some titles or headlines as opinion, but our systems could not extract them, because cue expressions do not appear at the end of these sentences. The following are two examples that assessors judged as opinion sentences.

[original]
◇納得できない−川田龍平さん
[translation]
(◇I cannot be convincing. – Mr. Ryuhei Kawada)

[orignial]
◇政府と協調必要−−白川浩道・UBSウォーバーグ証券チーフエコノミスト
[translation]
(◇ Cooperation with the government is necessary – Hiromichi Shirakawa, a chief economist of UBS Warburg)

3.3.2 Error Analysis for each Topic

Evaluation results for each topic are shown in Table 4. Topics N05 (Kosovo civil war), N12 (El Nino), and N14 (greenhouse gas) gave low F-values. All of these topics tend to contain many statements or conversational sentences in comparison with other topics. Generally, opinion sentences are written in the present tense, but these sentences are often written in the past tense, and our systems could not identify them as opinion sentences. We show an example as follows.

[original]
NOAAのトム・カール博士は「１９００年以降、陸地の気温の上昇率は海上より２０％も高く、最近の高温傾向は人間の活動による温室効果ガスの増加と関係していると言える」と指摘した。
[translation]
(Dr. Tom Carl of NOAA pointed out "Since 1900, the percentage of rise of land temperature is more than 20% higher than that of sea temperature. It can be said that this phenomenon in these few years relates to the increase of greenhouse gas owing to the human activity.")

3.3.3 Effects of Sentence-final Expressions

To confirm the effects of each list of sentence-final expressions, we removed one of these lists, conducted machine learning, and calculated recall and precision values. When the list from "TSC4" corpus was removed, the recall score fell by about 2 %, while the precision score was almost the same. When the list from "NTCIR-6" corpus was removed, the precision and recall scores fell by 5% and 2%, respectively. When the list from the "News" corpus was removed, both precision and recall scores fell by about 10%.

From these results, we can conclude that the "News" corpus is the most effective among three lists. The list from the "TSC4" corpus contributes to improve recall. Although the size of the list from the

"NTCIR-6" corpus is very small, it significantly improved the precision.

4 Conclusions

We have proposed a machine learning-based method for extracting opinion sentences using about 760,000 of sentence-final expressions and some other features. We submitted two systems to the Japanese Subtask of the MOAT at NTCIR-7 Workshop. We obtained F-values of 0.5615 and 0.3319 using lenient gold standard, and 0.5213 and 0.3561 using strict gold standard, respectively.

5 Acknowledgements

The authors would like to express their gratitude to the organizers of the MOAT.

References

[1] M. Okumura, T. Hirao, and H. Nanba. TSC4: A Corpus for Opinion Summarization, and A Workshop using the Corpus. In Proceedings of the 11th Annual Meeting of the Association for Natural Language Processing, 2005. (in Japanese)

[2] Y. Seki, D. K. Evans, L.-W. Ku, L. Sun, H.-H. Chen, and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[3] T. Kanamaru, M. Murata, and H. Isahara. Japanese Opinion Extraction System for Japanese Newspapers Using Machine-Learning Method. in Proceedings of the 6th NTCIR Workshop Meeting, pp.301-307, 2007.

[4] H. Mizuguchi, M. Tsuchida, and D. Kusui. Three-Phase Opinion Analysis System at NTCIR-6. In Proceedings of the 6th NTCIR Workshop Meeting, pp.330-335, 2007.

[5] Y. Seki, D. K. Evans, L.-W. Ku, H.-H. Chen, and N. Kando. Overview of Opinion Analysis Pilot Task at NTCIR-6. In Proceedings of the 6th NTCIR Workshop Meeting, pp.265-278, 2007.

[6] K. Kita, Y. Kato, T. Omoto, and Y. Yano. A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria. Journal of Natural Language Processing, Vol.1, No.1, pp.21-33, 1994.

[7] H. Nanba. Query Expansion using an Automatically Constructed Thesaurus. In Proceedings of the 6th NTCIR Workshop, pp.414-419, 2007.

[8] L. Zhou, C.-Y. Lin, D.S. Munteanu, and E. Hovy. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp.447-454, 2006.

Table 4. Evaluation results for each topic

Topic ID	precision	recall	F-value	TOPIC
N03	0.5809	0.5267	0.5525	米国同時多発テロ＆（米国）経済 (The impact of 911 terrorist attacks on America's economy)
N04	0.6183	0.6807	0.6480	薬害エイズ＆ミドリ十字 (HIV-tainted blood scandal)
N05	0.1892	0.3500	0.2456	コソボの民族紛争 (Kosovo civil war)
N06	0.3714	0.6190	0.4643	ネパール (Incident with Nepal's ruling family (royalty))
N07	0.3968	0.6410	0.4902	中国系インドネシア人に対する暴動 (Attacks toward Chinese Indonesian people)
N08	0.5686	0.6744	0.6170	米国対マイクロソフト (Lawsuit American Government against Microsoft)
N09	0.3125	0.4878	0.3810	核実験 (Nuclear weapons tests)
N10	0.3710	0.5750	0.4510	シリアと中東和平プロセス (Suriyah in the Middle East Peace Process.)
N11	0.4706	0.4211	0.4445	AOL とネットスケープ (The relationship between AOL and Netscape)
N12	0.3800	0.2754	0.3194	エルニーニョ (El Nino)
N13	0.3846	0.6000	0.4687	中国とロシア (The relationship between China and Russia)
N14	0.3158	0.2400	0.2727	温室効果ガス (Greenhouse gasses)
N15	0.5429	0.5938	0.5672	NATO とポーランド (The relationship between NATO and Poland)
N16	0.6211	0.6082	0.6146	タイとアジア経済危機 (Thailand in the Asian economic crisis)
N18	0.5000	0.5429	0.5206	チェチェン紛争 (Chechin (Chechnia) civil war)
N19	0.4167	0.6081	0.4945	スハルト大統領 (Indonesian President Suharto)
N20	0.5922	0.5495	0.5701	北朝鮮のミサイル開発放棄 (Nuclear missile abandonment of North Korea)
N21	0.3750	0.5581	0.4486	アジアでの航空機墜落事故 (Airplane crashes in Asia)