Multi-Document Summarization Reflecting Information Needs on Subjectivity

Yohei Seki
Toyohashi University of Technology
Aichi, 441-8580, Japan
seki@ics.tut.ac.jp

Koji Eguchi and Noriko Kando
National Institute of Informatics
Tokyo, 101-8430, Japan
{eguchi,kando}@nii.ac.jp

Masaki Aono
Toyohashi University of Technology
Aichi, 441-8580, Japan
aono@ics.tut.ac.jp

Abstract

In this paper, we present our experiments on improving multi-document summarization by reflecting information needs on subjectivity. Subjectivity is an essential aspect for better understanding of information needs. Our approach is based on sentence extraction, weighted by sentence type annotation, and combined with polarity term frequencies. From the DUC 2005 dataset, which focused on summarization for English documents, we selected 10 topics expressing information needs for subjective information and evaluated our results with two types of evaluation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BE (Basic Elements). For 10 topics, we found improvements of 10.2 % in the ROUGE-2 score, when compared to the baseline system with no analysis of topics. With failure analysis, we found the topics with improvements of ROUGE and BE scores contained effective subjective keywords.

Keywords: *DUC, Multi-Document Summarization, Information Needs on Subjectivity.*

1 Introduction

The purpose of this study is to clarify the effects of information needs on subjectivity for multi-document summarization. We have previously proposed the multi-document summarizer *v-SWIM*, which focuses on the facts, opinions, and knowledge described in documents, and have experimented on Japanese document sets [17, 19]. We reformulated this approach for application to English summarization, at DUC 2005 [20]. In this paper, we detail the analysis of the results at DUC 2005, so as to clarify the effects of keywords that express information needs for a subjectivity-sensitive task.

We suppose topics in the DUC 2005 dataset are written statements of user's information needs. We selected 10 topics expressing information needs for subjective information (which means expressive author's or authority's subjectivity), such as "benefits", "advantages", "positive or negative factors", "commentary", and so on.

We assume that "sentence types" in source documents can be significantly related to the types of users' information needs in actual information-seeking processes. Our proposed method automatically annotates the sentence type, such as subjective/objective, for every sentence in a source document, by using a support vector machine (SVM) [9, 8] and decision tree [15], which is a supervised machine-learning technique. We also counted the polarity term frequencies for subjective sentences, and built a summarizer to reflect information needs on subjectivity, using by these clues.

We evaluated our approach using two types of evaluation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [10], and BE (Basic Elements) [6]. We compared the summaries from our proposed system, which uses automatically identified sentence types in the source documents, with summaries from our baseline system, which does not differentiate sentence types. We analyzed the results in detail and found effective keywords for subjectivity-sensitive tasks.

This paper is organized as follows. In Section 2, we briefly introduce related work for subjective sentence categorization. In Section 3, we explain the DUC 2005 experiments and its datasets. Section 4 describes our proposed system and experimental results using DUC2005 dataset. Section 5 presents the analysis, which assess the effectiveness of our approach based on information needs analysis for subjectivity in topics. Section 6 contains discussions. Finally, we present our conclusions in Section 7.

2 Related Work

In this section, we briefly introduce related work on subjective sentence categorization.

2.1 Subjective Sentence Categorization

Related work for subjective sentence categorization can be divided into three groups, as follows:

1. Sentence categorization mainly focused on subjectivity and objectivity ([24, 3, 16]).
2. Sentence categorization mainly focused on semantic orientation/polarity ([13]).
3. Sentences categorization based on subjectivity, with subjective sentences being annotated with semantic orientation/polarity ([26, 25]).

In this paper, we took the first strategy and automatically categorize subjectivity in sentences. We did not categorize semantic orientation/polarity in sentences because positive samples were less in the training data. Instead, we count the polarity term frequencies in the subjective sentences to discriminate the semantic orientation/polarity in them which reflect information needs.

2.2 Features for Subjectivity Categorization

For subjectivity categorization using a supervised machine-learning technique, several types of clues have been used as features. The frequencies of content words, such as nouns or unknown words, are not effective features for categorizing subjective sentences. In this section, we introduce several features used for subjectivity categorization.

2.2.1 Orientation Words

Hatzivassiloglou et al. [3] utilized adjectives to decide if a sentence contains subjective information. They proposed a framework for extracting the orientation adjectives in previous research [2], and found that plus/minus orientation, plus/minus gradability, and dynamic adjectives were effective in extracting subjective sentences. These adjective entries are available for download from a Web site [4]. In [26], they utilized word unigram/bigram/trigram, part-of-speech frequencies, and orientation adjectives, as features for categorizing subjective sentences.

In this paper, we utilized orientation words from adjective entries [4] and the General Inquirer [22] as features.

2.2.2 Co-occurring Words

Subjectivity information has been categorized not only on a keyword such as an orientation adjective, but also on the combination of co-occurring words in a sentence. Riloff et al. [16] differentiated strong clues for subjectivity from weak clues for subjectivity, and surveyed syntactic patterns to extract subjective sentences.

```
<topic>
<num> d301i </num>
<title> International Organized Crime </title>

<narr>
Identify and describe types of organized crime that
crosses borders or involves more than one country.
Name the countries involved.
...
</narr>

<granularity> specific </granularity>
</topic>
```

Figure 1. Information needs in DUC 2005 dataset

3 DUC 2005 Experiments

The DUC (Document Understanding Conference) [12] is a series of evaluation in the area of text summarization, and has been held annually since 2000 by NIST (National Institute of Standards and Technology).

3.1 Dataset

The DUC 2005 dataset consists of three components: (1) 50 topics, (2) document sets relevant to each of the 50 topics, and (3) reference summaries of each document set.

For the DUC 2005 task, NIST will give assessors a list of old TREC topics. Assessors pick a TREC topic that they found interesting. Then, they formulate a DUC topic, which is a request for information about the aspects of the TREC topic that interest them.

An example of DUC topic statement is shown in Figure 1. The topic statement is expressed with the <num> tag (document set ID), the <title> tag, the <narr> tag, and the <granularity> tag. The statement related to information needs is written with the <narr> tag. The <granularity> tag specifies a two-valued user profile (specific or general).

Assessors read at least 50 of the documents (using $WebAssess$) and mark each one as "relevant" if they think it is relevant to the TREC topic. If they did not find at least 25 relevant documents, they pick another topic. Then, each topic contains a document set consisting of 25–50 relevant documents selected from Financial Times of London and Los Angels Times.

Assessors who developed the DUC topic also created a less-than-250-word summaries that met the need expressed in the topic. In total, four references summaries were prepared for each of 30 topics, and nine reference summaries for each of 20 topics. These multiple reference summaries were used in the evaluation.

3.2 Evaluation Metrics

NIST manually evaluated the linguistic well-formedness of each submitted system-produced summary, using a set of quality questions. In DUC 2005, linguistic quality was evaluated with five criteria: (1) grammaticality; (2) non redundancy; (3) referential clarity; (4) focus; and (5) structure and coherence.

NIST also manually evaluated the relative responsiveness to its topic of each submitted summary. In DUC 2005, responsiveness was evaluated by three schemes: (1) a raw responsiveness score assigned by NIST assessors; (2) a scaled responsiveness score computed as the sum of the scaled responsiveness scores proportional to the number of summaries for the topic; and (3) as in (2), but using only the automatic summaries (ignoring the human summaries in scaling responsiveness).

NIST automatically evaluate each submitted summary using ROUGE-1.5.5 [10], which enables automatic evaluation via n-gram co-occurrences.

In addition, the Pyramid evaluation [14] of how well each submitted summary agrees in content with the manually created reference summaries were carried out cooperatively by the participating groups under the leadership of Columbia University team. BE [6] is another automatic evaluation, which uses a syntactic parser to detect a head-BE and a single dependent and was used at DUC 2005 evaluation by Hovy et al. [7].

3.3 Task

The DUC 2005 task was a complex question-focused summarization, which required summarizers to collect information from multiple documents that answered a question or set of questions [1]. The summaries met the information needs expressed in the topics, and they were also consistent with the granularity field in them.

The system-produced summary could be no longer than 250 words (space-delimited tokens). Summaries over the size limit were truncated.

4 Our Proposed System

4.1 System Overview

Our proposed system was based on sentence extraction, using document clustering techniques for paragraph units to remove redundant information. In addition, in order to generate summaries sensitive to the topic, subjective information was used as a weight in selecting sentences to extract.

The algorithm of our proposed system was tested at NTCIR-4 TSC [18], and worked well in comparison to other participants [5]. We chose Ward's clustering algorithm as it obtained the best results in the pretest in which comparing the different clustering algorithms of complete linkage, group average, or Ward's method, on the same document collection.For the cluster unit, we used paragraphs rather than sentences because of the sparseness of vector spaces when using sentence vectors. The detailed algorithm is described as follows:

1. Paragraph Clustering Stage
   (a) Source documents were segmented into paragraphs, and then term frequencies were indexed for each paragraph.
   (b) Paragraphs were clustered based on Euclidean distances between feature vectors based on term frequency, using Ward's method. In DUC 2005, the summary size was 250 words. A sentence contained 22.58 words on average. For all the 50 document sets in DUC 2005 dataset, a document set contained 455.02 paragraphs, on average. In the official submission, the number of clusters for paragraphs were fixed at 20 clusters, based on the number of extracted sentences. (We set this the number of clusters because, if one sentence contained 25 words on average, sentences would be extracted from half the clusters, similar to queries represented by content words in "narratives" or "titles".)
2. Sentence Extraction Stage
   (a) The feature vectors for each cluster were computed with term frequencies (TF) and inverse cluster frequencies in expression (1).

$$\mathrm{TermFrequency} * log(\frac{\mathrm{TotalClusters}}{\mathrm{ClusterFrequency}}). \quad (1)$$

   Terms were stemmed using OAK [21].
   (b) Clusters were ordered by the similarity between content words in "titles" and "narratives" of each topic.
   (c) Sentences in each cluster were weighted, based on content words in "narratives" and "titles", heading words in the cluster, and TF values in the cluster. In addition, "narratives" in the topics were used for analysis of the information needs on subjectivity. (This process will be explained in Section 5.2). The weight scheme is expressed in expression (2).

$$\begin{aligned} W(s) = \quad & L(s) \times \\ & (a_1 \times Q(s) + a_2 \times H(s) + a_3 \times T(s) \\ & + a_4 \times \underline{N(s)} + a_5 \times \underline{S(s)}), \end{aligned} \quad (2)$$

where: *L(s)* is the weight based on the location of the sentence *s* in the document; *Q(s)* is the number of content words in "narratives" and "titles" appearing in sentence *s*; *H(s)* is the number of heading words appearing in sentence *s*; and *T(s)* is the sum of TF values of words appearing in sentence *s*.

The two underlined predicates, *N(s)* and *S(s)*, are optional weight predicates based on analysis of topics, as discussed in Section 5.2. *N(s)* is the frequencies of named entity tags, matched against the information type from analysis of topics. it S(s) = 1 if sentence s is subjective, otherwise *S(s)* = 0.

a_1 to a_5 are parameters. In DUC 2005, they were set as follows: a_1 = 0.4; a_2 = $\frac{1}{\text{total number of heading words in the cluster}}$; a_3 = 1; a_4 = 0.4; a_5 = 20. [1].

(d) One sentence was extracted from each cluster, in cluster order, ordered by the similarity between content words in "narratives" and "titles", and the cluster feature vectors, to reach the maximum number of words allowed (250 words).

(e) Conjunctions, such as "And", "But", "However", at the beginning of a sentence were removed, and the initial character of a sentence was capitalized.

4.2 Evaluation

In this section, we present three types of evaluation of our proposed system, as required by official submissions to DUC 2005: (1) linguistic quality; (2) responsiveness; and (3) ROUGE and BE.

4.2.1 Linguistic Quality

The results for our system are shown in Table 1[2]. They show that our system removes redundant information very well, being ranked second out of 31 systems. Referential clarity turned out to be acceptable, being ranked seventh, partly because our system removed conjunctions, such as "And", "But", "However", at the beginning of sentences.

4.2.2 Responsiveness

Results for our system's average scores and ranks for responsiveness evaluation are shown in Table 2.

[1] Initial parameters were set empirically, and optimal values were discussed in Section 5.

[2] Average of all the 50 topics in DUC2005. For Tables 2 and 3, this is the same.

Table 1. Quality evaluation

Quality Criterion	Score	Rank (of 31 systems)
Grammaticality	3.74	21
Non-redundancy	4.72	2
Reference	3.3	7
Focus	3.06	19
Coherence	2.12	12
Average	3.39	11

Table 2. Responsiveness

	Responsiveness		
	Raw	Scaled	
		(all summaries)	(system summaries only)
Score	2.40	16.82	16.63
Rank (of 31 systems)	18	14	13

4.2.3 ROUGE and BE

ROUGE [10] and BE [6] can be evaluated automatically, and they can be used for re-evaluation. Official evaluations were based on chunking results for our submitted summaries. Because the chunker used was not provided to us, we re-evaluated our submission by chunking sentences from the original documents using OAK [21]. The results of the official evaluation and our re-evaluation are shown in Table 3. Note that BE [6] was not used as an official evaluation tool in DUC 2005 and the result was cited from Hovy's paper [7]. In BE, several types of parser could be used to evaluate summaries and we selected the MiniPar parser [11].

Table 3. ROUGE and BE scores

Evaluation Metrics	Official		Re-evaluation
	Scores	Rank (of 31 systems)	Scores
ROUGE-SU4	0.11117	19	0.11115
ROUGE-2	0.05726	19	0.05722
BE	0.02077	20	0.0223

5 Analysis

Starting with our official submission, we improved the system by tuning these parameters: (1) the number of clusters; and (2) query vectors using "narratives" and "titles" in topics. We set up our system with optimal parameters as a baseline. Then, we did experiments to investigate the effectiveness of our subjectivity analysis.

5.1 Baseline System Optimizing Number of Clusters and Title Weights

Depending on the number of clusters, our system scores changed drastically. Our submission was based on a the number of 20 clusters. We changed this size from 20 to 70 in steps of 10 and evaluated ROUGE

and BE scores [20]. Our system produces summaries with the highest ROUGE-2 score when the number of clusters = 60.

We could have used different weights, so as to focus more on content words in "narratives". We changed the weights of content words in "titles" from 0 to 1 in steps of 0.1, and evaluated the resulting ROUGE and BE scores [20]. From these results, we found that a ratio of content words in "titles" to content words in "narratives" of 1 : 10 to made query vectors (and sentence weighting) perform best.

5.2 Information Needs Analysis for Subjectivity Using Topics

We analyzed topics, which expressed information needs from subjective aspects. In this section, we present an overview of sentence extraction processes, by using this analysis.

We selected 10 document sets (d360, d383, d385, d404, d413, d654, d671, d683, d694, and d699), in which topics contained information needs focused on subjectivity/sentiment information, such as "benefits", "advantages", "disadvantages", "positive or negative factors", "commentary", "discuss", "pros and cons", and "arguments".

We also categorized topics as "comment", "positive", or "negative" types. (In the official submission version, we only categorized them as "subjective" or "not subjective". We re-implemented our question analysis module for these experiments.) The results of the categorization are shown in Table 4.

Table 4. Document sets utilizing subjectivity weights

Subjectivity Type	Document Set
Comment	d404,d683,d694,d699
Positive	d360,d383,d385,d413, d654,d671,d694,d699
Negative	d385,d654,d699

5.3 Subjectivity Analysis

5.3.1 Subjective Sentence Categorization

We tagged the subjective information in sentences, i.e., whether they were subjective. This information in source documents was automatically annotated using SVM^{light} [9] and C5.0 [15]. Features were based on polarity type frequencies using adjective entries [4] and the General Inquirer [22]. As training data, we utilized the Multi-Perspective Question-Answering Corpus [23].

To assess the effectiveness of our subjective information annotation framework, we conducted a fivefold cross validation using the Multi-Perspective Question-Answering (MPQA) corpus [23]. This corpus contains 535 documents (10,657 sentences in total). Following Riloff's research [16], we categorized sentences as either subjective or objective, and 5,572 sentences were annotated as subjective for this corpus. We then divided these document sets into five groups of 107 documents each. For our machine-learning technique, we used the frequency of the following nine features:

1. Polarity plus type adjectives in a sentence;
2. Polarity minus type adjectives in a sentence;
3. Gradability plus type adjectives in a sentence;
4. Gradability minus type adjectives in a sentence;
5. Dynamic adjectives in a sentence;
6. Strong positive words in a sentence;
7. Strong negative words in a sentence;
8. Weak positive words in a sentence;
9. Weak negative words in a sentence.

For features 1 to 5, we used adjective entries collected by Hatzivassiloglou et al. [4], which contained 1,914 word entries. For features 6 to 9, we utilized the General Inquirer [22], which contained 1,168 word entries. Using SVM^{light} and C5.0 with these features, the macro-average values of accuracy, precision, and recall for fivefold cross validation of automatic subjectivity annotation for the MPQA corpus are shown in Table 5. For SVM^{light} learning options, kernel function was set as polynomial type and cost was set as 1.2. C5.0 was used with boosting option.

Table 5. Results of fivefold cross validation test of automatic subjectivity annotation (macro-average value)

SVM			C5.0 (with boosting)		
Accuracy	Precision	Recall	Accuracy	Precision	Recall
0.602	0.610	0.657	0.614	0.636	0.605

5.3.2 Multi-Document Summarization Reflecting Information Needs on Subjectivity

We used the 10,657 sentences in the MPQA corpus as training data, and automatically annotated all sentences in the DUC 2005 source documents as subjective or not subjective.

For the "comment" type, subjective sentences were weighted. For the "positive" and "negative" types, frequencies of polarity plus type adjectives, gradability plus type adjectives, and strong positive words in a sentence (or polarity minus type adjectives, gradability minus type adjectives, and strong negative words in a sentence) were weighted to produce summaries.

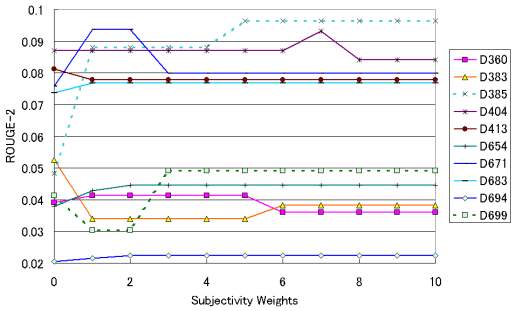

Figure 2. ROUGE-2 score change with subjectivity weights

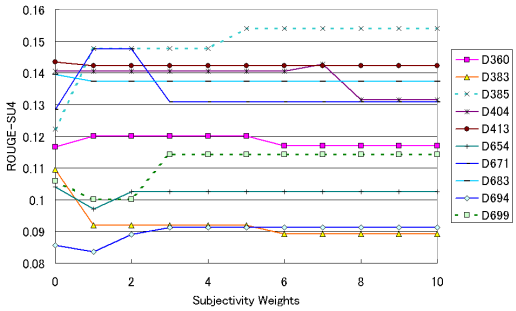

Figure 3. ROUGE-SU4 score change with subjectivity weights

6 Discussion

6.1 Effective Keywords for Subjectivity-Sensitive Tasks

For 10 subjective topics, we changed the subjectivity weights in expression (2) from 0 to 10 in steps of 1 and evaluated the ROUGE and BE scores. These results are shown in Figure 2, Figure 3, and Figure 4. Note that only the 10 topics appear in the graphs. This result was based on the automatic annotation using SVM^{light}. We also evaluated the summaries based on the automatic annotation using C5.0, and it showed the similar result.

In these figures, the best performance is when subjectivity weights = 5 (ROUGE-SU4) and 7 (ROUGE-2, BE). The distribution of topics for which scores were increasing, showed no change, or were decreasing, compared to the baseline system, when subjectivity weights = 7, is shown in Table 6. Document sets D360, D383, D413, and D671 were categorized "positive" type only, and most of the scores in these topics did not show any improvement. One reason for this, especially for D383 and D413, was that their topic statements contained types of information needs other than subjectivity. The outstanding improvements were shown in D385 and D699. The topics contained the keywords "positive and negative factors" or "pros and cons".

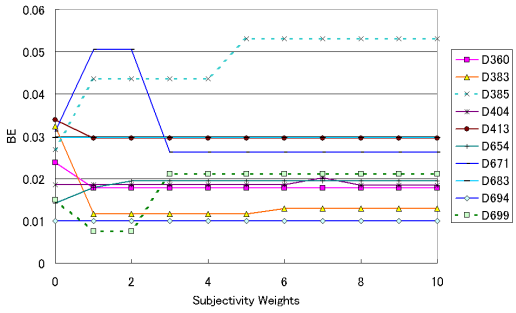

Figure 4. BE score change with subjectivity weights

6.2 Polarity in Reference Summaries

We analyzed reference summaries by counting polarity term frequencies for each summary. The reference summaries were constructed by either four or nine assessors. The results are shown in Table 7.

This table shows term frequencies for polarity plus/minus type adjectives ("Polarity Adj."), gradability plus/minus type adjectives ("Gradability Adj."), strong positive/negative words from the General Inquirer ("GI Strong"), and weak positive/negative words from the General Inquirer ("GI Weak"). The significance of term frequencies between the 10 subjective topics and other 40 topics were tested with a t-test. The results showed that frequencies of polarity plus type adjectives and strong positive words in subjective topics were significantly higher than in non-subjective topics.

The table also shows the polarity term frequencies for "Comment + Positive" type summaries (D694, D699) and "Positive + Negative" type summaries (D385, D654, D699). For the "Comment + Positive" type, the topics contained a description such as "**Discuss** making and using compost for gardening. Include different types of ... and **benefits**." (D694), or "What are the **pros** and **cons** of term limits? What are the similarities and differences among the **arguments** when ..." (D699). For the "Positive + Negative" type, the topics contained descriptions such as "What are the **advantages** and **disadvantages** of same-sex schools?" (D654). The summaries for these topics contained more polarity terms. Note, however, that the difference was not statistically significant because of the small number of topics.

Table 6. Distributions of topics with subjectivity weights = 7

	ROUGE-2	ROUGE-SU4	BE
Increasing	D385,D404,D654,D671,D683,D694,D699	D360,D385,D404,D671,D694,D699	D385,D404,D654,D699
No change	–	–	D683,D694
Decreasing	D360,D383,D413	D383,D413,D654,D683	D360,D383,D413,D671

Table 7. Polarity term frequencies in reference summaries

Average frequencies per summary	Positive Term Frequencies				Negative Term Frequencies			
	Polarity Adj.	Gradability Adj.	GI Strong	GI Weak	Polarity Adj.	Gradability Adj.	GI Strong	GI Weak
Average on 50 topics	8.41	10.38	9.03	0.29	7.22	11.35	10.79	10.48
Average on 10 subjective topics	10.18*	12.08	10.76*	0.27	7.13	11.78	9.51	12.08
Comment + Positive (D694,D699)	13.25*	14	10.88	0.13	8.38	11.5	12.88	15.88
Positive + Negative (D385,D654,D699)	11.31	13.59	12.44	0.23	5.71	13.81	9.19	10.44

*: statistically significant with t-test: $p < 0.05$

6.3 Discussion on Pyramids

We analyzed SCUs (summarization content units) in [14] for subjective topics. In DUC 2005, only 20 topics were evaluated based on the Pyramid Method. For our 10 subjective topics, only four topics (D413, D654, D671, and D683) were evaluated. Of these topics, three topics contained fewer numbers of SCUs than average (= 120). This reflects the tendency for SCUs for subjective summarization to be longer. This tendency is particularly clear for D654, with this document set containing only 89 SCUs. The "narrative" in the topic for D654 was "What are the **advantages** and **disadvantages** of same-sex schools?". By contrast, D683 contained 132 SCUs. The "narrative" in this topic was "**Discuss** the events leading to the breakup of Czechoslovakia, ..." and its reference summaries contained many fact-based events. This was a different type of subjective summary in that it focused on the subjective aspect of the construction of summaries.

Another aspect was the mean SCU weight, which indicates how many assessors included the common SCUs. For subjective topics, D413 and D671 were above the average (= 1.90), and the other two sets, D654 and D683, were below the average. D413 and D671 contained the keyword "benefits" and so we may suppose that their summaries relating to the concept expressed by this term tend to share the elements common to different assessors.

7 Conclusions

In this paper, we discussed the effectiveness of our approach, which reflects information needs on subjectivity from topics. We also presented a detailed analysis of 10 topics from DUC 2005, with their reference summaries and pyramids. The results showed that more polarity terms were contained in summaries reflecting combinations of subjective keywords. The analysis of pyramids implied that there were several types of subjective summaries, and they tend to be longer than nonsubjective summaries.

Acknowledgments

This work was partially supported by the Grants-in-Aid for Exploratory Research (#16650053) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. We thank Karen Spärck Jones for her useful comments in the DUC 2005 workshop.

References

[1] H. T. Dang. Overview of DUC 2005. In *Proc. of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf. / Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, October 2005.

[2] V. Hatzivassiloglou and K. M. McKeown. Predicting the semantic orientation of adjectives. In *Proc. of Wksp. Intelligent Scalable Text Summarization at the 35th Ann. Meeting of the Assoc. for Computational Linguistics joint with the 8th Conf. of the European Chapter of the Assoc. for Computational Linguistics (ACL/EACL '97)*, pages 174–181, Madrid, Spain, 1997.

[3] V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proc. of the 18th Int'l Conf. on Computational Linguistics (COLING 2000)*, pages 849–855, Saarbrücken, Germany, 2000.

[4] V. Hatzivassiloglou and J. M. Wiebe. Lists of manually and automatically identified gradable, polar, and dynamic adjectives. gzipped tar

file, 2000. [cited 2005-8-26]. Available from: <http://www.cs.pitt.edu/ wiebe/pubs/coling00/coling00adjs.tar.gz>.

[5] T. Hirao, M. Okumura, T. Fukusima, and H. Nanba. Text Summarization Challenge 3: Text Summarization Evaluation at NTCIR Workshop 4. In *Proc. of the Fourth NTCIR Wksp. on Research in Information Access Technologies: Information Retrieval, Question Answering, and Summarization*. NII, 2004.

[6] E. Hovy, C.-Y. Lin, J. Fukumoto, K. McKeown, and A. Nenkova. Basic Elements (BE) Version 1.1 [online], 2005. [cited 2005-8-26]. Available from: <http://www.isi.edu/~cyl/BE/>.

[7] E. Hovy, C.-Y. Lin, and L. Zhou. Evaluating DUC 2005 using Basic Elements. In *Proc. of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf. / Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, October 2005.

[8] T. Joachims. *Learning to Classify Text Using Support Vector Machines : Methods, Theory, and Algorithms*. Kluwer Academic Publishers, 2002.

[9] T. Joachims. SVM^{light} Support Vector Machine Version 6.01 [online], 2004. [cited 2005-8-26]. Available from: <http://svmlight.joachims.org>.

[10] C.-Y. Lin. ROUGE - Recall-Oriented Understudy for Gisting Evaluation - Version 1.5.5 [online], 2005. [cited 2005-8-26]. Available from: <http://www.isi.edu/~cyl/ROUGE/>.

[11] D. Lin. MINIPAR Home Page [online], 2005. [cited 2005-8-26]. Available from: <http://www.cs.ualberta.ca/~lindek/minipar.htm>.

[12] National Institute of Standars and Technology. Document Understanding Conferences (DUC) [online]. In *Document Understanding Conferences (DUC) website*, 2001-2005. [cited 2004-10-26]. Available from: <http://duc.nist.gov/>.

[13] B. Pang and L. Lee. Seeing stars: Explointing class relationships for sentiment categorization with respect to rating scales. In *Proc. of the 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL 2005)*, pages 115–124, Ann Arbor, Michigan, June 2005.

[14] R. J. Passonneau, A. Nenkova, K. McKeown, and S. Sigleman. Applying the Pyramid Method in DUC 2005. In *Proc. of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf. / Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, October 2005.

[15] R. Quinlan. Data Mining Tools See5 and C5.0 [online], 2004. [cited 2005-8-26]. Available from: <http://www.rulequest.com/see5-info.html>.

[16] E. Riloff and J. M. Wiebe. Learing extraction patterns for subjective expressions. In *Proc. 2003 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 105–112, Sapporo, Japan, July 2003.

[17] Y. Seki, K. Eguchi, and N. Kando. Analysis of Multi-Document Viewpoint Summarization Using Multi-Dimensional Genres. In *Proc. of AAAI Spring Sympo. on Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT 2004)*, pages 142–145, Stanford, CA, March 2004.

[18] Y. Seki, K. Eguchi, and N. Kando. User-focused Multi-document Summarization with Paragraph Clustering and Sentence-type Filtering. In *Proc. of the Fourth NTCIR Wksp. on Research in Information Access Technologies: Information Retrieval, Question Answering, and Summarization*, pages 459–466, June 2004.

[19] Y. Seki, K. Eguchi, and N. Kando. Multi-document viewpoint summarization focused on facts, opinion and knowledge (in press). In J. G. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text*, volume 20 of *The Information Retrieval Series*, chapter 24, pages 317–336. Springer, Dordrecht, The Netherlands, October 2005.

[20] Y. Seki, K. Eguchi, N. Kando, and M. Aono. Multi-Document Summarization with Subjectivity Analysis at DUC 2005. In *Proc. of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf. / Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, October 2005.

[21] S. Sekine. OAK System (English Sentence Analyzer) Version 0.1 [online], 2002. [cited 2005-8-26]. Available from: <http://nlp.cs.nyu.edu/oak/>.

[22] P. J. Stone. The General-Inquirer [online], 2000. [cited 2005-8-26]. Available from: <http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm>.

[23] J. M. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, and T. Wilson. MPQA: Multi-Perspective Question Answering Opinion Corpus Version 1.1, 2005. [cited 2005-8-26]. Available from: <http://nrrc.mitre.org/NRRC/02_results/mpqa.html>.

[24] J. M. Wiebe, T. Wilson, R. F. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

[25] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of the 2005 Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, B. C., 2005.

[26] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. 2003 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 129–136, Sapporo, Japan, July 2003.