# Multi-Document Summarization with Subjectivity Analysis at DUC 2005

**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 team TUT/NII results at DUC 2005 and additional experiments on improving multi-document summarization. Summarization systems have typically focused on the factual aspect of information needs. Subjectivity analysis is another 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. We selected 10 topics related to subjectivity with analysis of "narratives", and investigated improvements of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BE (Basic Elements) scores with our approach. In addition, the factual aspect of information needs was also investigated.

## 1 Introduction

The purpose of our study is to build a multi-document summarizer on the basis of user-specified summary viewpoints. 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 (Seki et al., 2004a; Seki et al., 2005). We reformulated our approach for application to English summarization, at DUC 2005. We attempted to clarify the usability of user context information for a subjectivity-sensitive task. This has not been investigated to date, although user context information for topic focusing has featured in past studies in this research area. This research also provides an enlightening discussion on the association between input factors and purpose factors in summarization tasks.

Subjectivity usually refers to some aspects of language description that were formed to express an author's or an authority's opinions, evaluations, and speculations (Wiebe et al., 2004). Although subjectivity analysis research has been mainly applied to measuring perceptions of reputations of commercial products or movie titles on the web, subjectivity analysis of newspaper articles is also important for information analysis in some domains, such as politics. This study attempts to clarify the feasibility of this.

We assume that "sentence types" in source documents can be significantly related to the types of users' information needs in actual information-seeking processes. The "sentence type" (Teufel and Moens, 2002; McKnight and Srinivasan, 2003; Seki et al., 2004b) is defined as the role or type of information of a sentence in a document structure. We focused on sentence types for an investigation of subjectivity, which was defined as identifying whether a sentence expressed a positive or negative attitude.

We suppose topics in the DUC 2005 dataset are written statements of user's information needs. We selected 10 topics, in which "narratives" contained information needs focused on subjective information (which means expressive author's or authority's subjectivity), such as "benefits", "advantages", "positive or negative factors", "commentary", and so on. 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) (Joachims, 2002; Joachims, 2004), 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 these clues.

We evaluated our approach using two types of evaluation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2005a), which was automatic evaluation using n-gram co-occurrences and BE (Basic Elements) (Hovy et al., 2005), which was another automatic evaluation using a syntactic parser to detect a head-BE and a single dependent. 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.

This paper is organized as follows. In Section 2, we explain our multi-document summarization system. Section 3 details the official evaluations at DUC 2005. Section 4 presents additional experiments with subjectivity analysis. Finally, we present our conclusions.

## 2 System Overview

The TUT/NII team's 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 "narratives", which were given by DUC 2005 organizers as one type of user context information, named entity and subjectivity information was used as a weight in selecting sentences to extract.

### 2.1 Query-Biased Multi-Document Summarization Using Content Words in Narratives and Titles

The algorithm of the TUT/NII team's system was tested by the SOKEN team at NTCIR-4 TSC (Seki et al., 2004c) and worked well in comparison to other participants (Hirao et al., 2004).

Many clustering-based multi-document summarization frameworks (Stein et al., 2000; Hatzivassiloglou et al., 2001; Mana-Lopez et al., 2004; Radev et al., 2004) have been proposed. Those projects focused on making the topic structure explicit. By detecting similarities in topic structure, their systems could avoid redundant information in summaries. These methods have four principal aspects: (1) clustering algorithms; (2) cluster units; (3) sentence extraction strategy; and (4) the number of clusters.

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. Ward's clustering is a clustering procedure that seeks to form the partitions in a manner that minimizes the loss associated with each grouping. Information loss is defined in terms of an error sum of squares criterion. 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 (TF) 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 was fixed as 20 clusters, based on the number of extracted sentences. (We set this 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:

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

   Terms were stemmed using OAK (Sekine, 2002).

   (b) Clusters were ordered by the similarity between content words in "titles" and "narratives", provided for each topic by DUC 2005 organizers, and the cluster feature vectors.

   (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" were used as statements to express the information needs (This process will be explained in Section 2.2). The weight scheme is expressed in expression (2).

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

   *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 TF values in the cluster.

   The two underlined predicates, *N(s)* and *S(s)*, are optional weight predicates based on analysis of "narratives", as discussed in Section

2.2. *N(s)* is the frequencies of named entity tags, matched against the information type from analysis of "narratives", *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.

### 2.2 Information Needs Analysis using Narratives

We analyzed "narratives", expressing information needs from factual and subjective aspects. In this section, we present an overview of sentence extraction processes by using these analyses of "narratives". For factual information needs, named entity tags in sentences were used to weight sentences. We explain this in Subsection 2.2.1. For subjective information needs, subjective sentences were weighted as detailed in subsection 2.2.2.

#### 2.2.1 Sentence Extraction with Named Entity

We analyzed "narratives" and categorized information types using interrogative words combined with keywords. For example, if "when" appeared in a "narrative", the information type was categorized as "TIME". If a "narrative" contained the interrogative word "which" and keywords such as "country", information type was categorized as "COUNTRY".

Named entity information was tagged in the original documents using the OAK system (Sekine, 2002). After matching named entity tags to information types, sentences were weighted based on the frequencies of the named entity elements in the sentences.

#### 2.2.2 Sentence Extraction with Subjectivity

We also tagged the subjective information in sentences, i.e., whether they were subjective. This information was tagged using SVM$^{light}$ (Joachims, 2004). Features were based on polarity type frequencies using adjective entries (Hatzivassiloglou and Wiebe, 2000) and General Inquirer (Stone, 2000). As training data, we utilized the Multi-Perspective Question-Answering Corpus (Wiebe et al., 2005). We selected 10 topics (d360, d383, d385, d404, d413, d654, d671, d683, d694, and d699), in which "narratives" contained information needs focused on subjective information, such as "benefits", "advantages", "disadvantages", "positive or negative factors", "commentary", "discuss", "pros and cons", and "arguments". For these sets, subjective sentences were weighted and extracted.

## 3 Evaluation

In this section, we present four types of evaluations of the TUT/NII team, as required by official submissions to DUC 2005: (1) linguistic quality questions; (2) pyramid evaluation; (3) responsiveness; and (4) ROUGE and BE.

### 3.1 Linguistic 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. 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.

Table 1: Quality evaluation for the TUT/NII team

| 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 |

### 3.2 Pyramid Evaluation

In DUC 2005, DUC participants were asked to participate in a pyramid evaluation, proposed by Columbia University members (Nenkova and Passonneau, 2004). The pyramid method is a manual method for summarization evaluation to address a problem that different humans choose different content when writing summaries. The pyramid method addresses the problem by using multiple human summaries to create a gold-standard and by exploring the frequency of information in the human summaries in order to assign importance to different facts. Of 31 participants, 24 teams systems agreed (plus one baseline system) and were evaluated. The results for our system (processed scores) are shown in Table 2. Note that only 20 of the 50 topics were evaluated. Six topics (d324, d400, d407, d426, d633, and d695) were evaluated

[1] Initial parameters were set empirically and optimal values were discussed in Section 4.

[2] Average of all the 50 topics in DUC2005. For Tables 3 and 4, this is the same.

by several teams and the results averaged. “Score” equals the weight of the summary content units normalized by the weight of an ideally informative summary consisting of the same number of content units as the peer. “Modified score” is very closely related to the original pyramid score, but uses a different normalization factor. The normalization factor for the modified score is the ideal weight of a summary with the expected content unit size of a human summary. The expected size is calculated as the average content unit size of the human summaries used to build the pyramid. In other words, according to Kathleen McKeown in DUC 2005 mailing list discussions, “score” is closer to the notion of “precision” and “modified score” is closer to the notion of “recall”.

Table 2: Pyramid evaluation for the TUT/NII team

| Document Set ID | Processed_Pans | |
|---|---|---|
| | Score | Modified Score |
| 311 | 0.4327 | 0.3285 |
| 324 (Average) | 0.3370 | 0.2768 |
| 324 (Reduced) | 0.3900 | 0.3482 |
| 345 | 0.1507 | 0.1209 |
| 366 | 0.1556 | 0.1157 |
| 376 | 0.2268 | 0.1447 |
| 391 | 0.1774 | 0.1375 |
| 393 | 0.1319 | 0.1034 |
| 400 (Average) | 0.2835 | 0.1907 |
| 407 (Average) | 0.2837 | 0.2188 |
| 413 | 0.3254 | 0.2715 |
| 422 | 0.1897 | 0.1310 |
| 426 (Average) | 0.1699 | 0.1000 |
| 431 | 0.0860 | 0.0584 |
| 435 | 0.1398 | 0.1024 |
| 632 | 0.0957 | 0.0698 |
| 633 (Average) | 0.1445 | 0.1323 |
| 654 | 0.1644 | 0.1791 |
| 671 | 0.1792 | 0.1293 |
| 683 | 0.1562 | 0.1087 |
| 695 (Average) | 0.2959 | 0.2590 |
| Avg. (Unreduced) | 0.2063 | 0.1589 |
| Rank (Unreduced) of 24 systems | 13 | 13 |
| Rank (Reduced) of 24 systems | 3 | 4 |

### 3.3 Responsiveness

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). Results for the TUT/NII team's average scores and ranks are shown in Table 3.

### 3.4 ROUGE and BE

ROUGE (Lin, 2005a) and BE (Hovy et al., 2005) are automatic evaluation tools 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 (Sekine, 2002). The results of the official evaluation and our re-evaluation are shown in Table 4. Note that BE (Hovy et al., 2005) was not used as an official evaluation tool in DUC 2005. In BE, several types of parser could be used to evaluate summaries and we selected the MiniPar parser. (Lin, 2005b).

Table 3: Responsiveness for the TUT/NII team

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

Table 4: ROUGE and BE scores for the TUT/NII team

| 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 | - | - | 0.0223 |

### 3.5 Topic-by-topic Evaluation with Multiple Evaluation Metrics

We investigated our results using topic-by-topic evaluation. The ranks for each topic are shown in Table 5. We only show the 20 topics evaluated by pyramid metrics due to space limitation.

Table 5: Topic-by-topic evaluation for the TUT/NII team

| Topic | Rank | | | | |
|---|---|---|---|---|---|
| | ROUGE | | Scaled | Pyramid | |
| | 2 | SU4 | Responsiveness | Score | Modified Score |
| D311 | 11 | 16 | 1 | 15 | 16 |
| D324 | 18 | 9 | 11 | 15 | 15 |
| D345 | 8 | 10 | 12 | 19 | 19 |
| D366 | 23 | 23 | 20 | 15 | 15 |
| D376 | 17 | 18 | 1 | 13 | 13 |
| D391 | 7 | 11 | 2 | 14 | 14 |
| D393 | 14 | 12 | 3 | 12 | 9 |
| D400 | 23 | 24 | 12 | 5 | 5 |
| D407 | 6 | 9 | 4 | 7 | 5 |
| D413 | 5 | 5 | 1 | 4 | 4 |
| D422 | 20 | 22 | 9 | 5 | 7 |
| D426 | 16 | 18 | 4 | 21 | 21 |
| D431 | 26 | 18 | 1 | 19 | 18 |
| D435 | 26 | 26 | 26 | 19 | 17 |
| D632 | 18 | 16 | 2 | 9 | 8 |
| D633 | 19 | 25 | 2 | 13 | 12 |
| D654 | 17 | 16 | 13 | 11 | 10 |
| D671 | 19 | 15 | 6 | 7 | 5 |
| D683 | 28 | 20 | 5 | 19 | 20 |
| D695 | 4 | 6 | 4 | 2 | 1 |
| Avg. | 19 | 19 | 13 | 13 | 13 |

In this table, D407, D413, and D695 were ranked as single figures for all evaluation metrics. For these topics, the “narratives” required three questions, which also

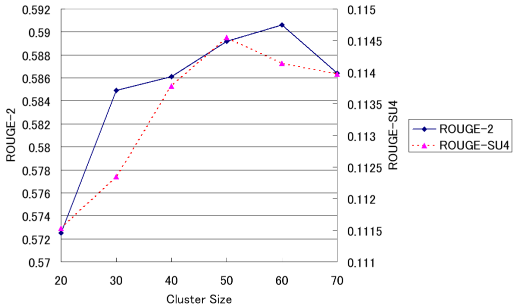

Figure 1: ROUGE score change with number of clusters

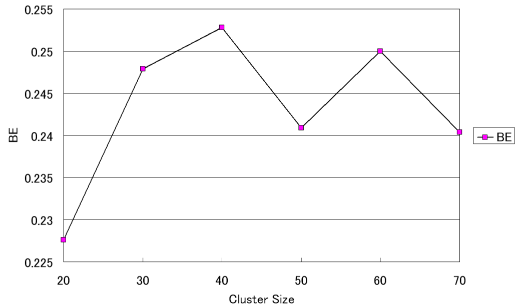

Figure 2: BE score change with number of clusters

consisted of several types of information needs. In contrast, D366 and D435 were ranked under the ranks of the average scores. The "narratives" contained the inferential/causal type information needs such as "potential", "caused", "factors", or "influencing" .

## 4 Additional Experiments for Future Improvements

Starting with our official submission, we performed additional experiments to improve our system for a future submission. Here, we discuss the effect of: (1) the number of clusters; (2) query vectors using "narratives" and "titles"; (3) named entity for factual information needs; and (4) subjectivity analysis.

### 4.1 Number of Clusters

Depending on the number of clusters, our system scores changed drastically. Our submission was based on 20 clusters. We changed this size from 20 to 70 in steps of 10 and evaluated ROUGE and BE scores. ROUGE scores are shown in Figure 1 and BE scores are shown in Figure 2. In these results, the effects of named entity analysis and subjectivity analysis, discussed in the following subsections, were removed. From Figure 1, our system produces summaries with the highest ROUGE-2 score with the number of clusters = 60 and with the highest ROUGE-SU4 score with the number of clusters = 50. From Figure 2, our system produces summaries with the highest BE score when the number of clusters = 40.

### 4.2 Query Vectors Using Narratives and Titles

In our official submission, to make query vectors, we treated content words appearing in "narratives" and in "titles" equivalently. However, we could have used different weights, so as to focus more on content words in "narratives". In Figure 3 and 4, 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. From these results, we found that a 1:10 ratio of content words in "titles" to content words in "narratives" made query vectors (and sentence weighting) perform best.

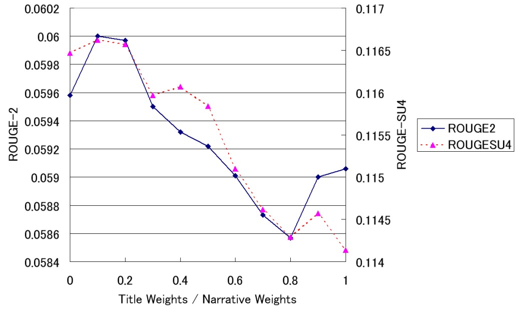

Figure 3: ROUGE score change with title weights (number of clusters = 60)

### 4.3 Named Entity Analysis

We implemented a question analysis module to automatically detect named entity (NE) types inferred from "narratives", as shown in Table 6. The "Person" NE type almost corresponded to "who"-type questions in "narratives". The "Country" NE type related to questions in "narratives" that asked about countries. NE tags were also annotated to source documents using OAK (Sekine,

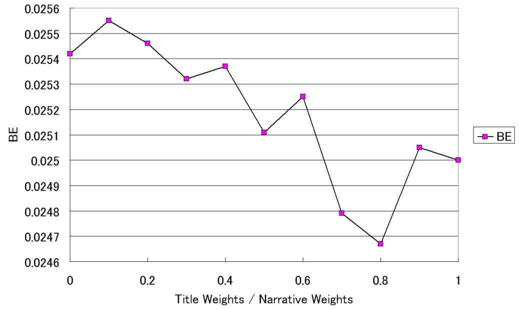

Figure 4: BE score change with title weights (number of clusters = 60)

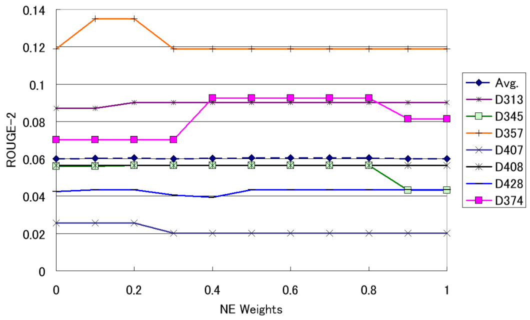

Figure 5: ROUGE-2 score change with named entity weights (number of clusters = 60, title weights = 0.1)

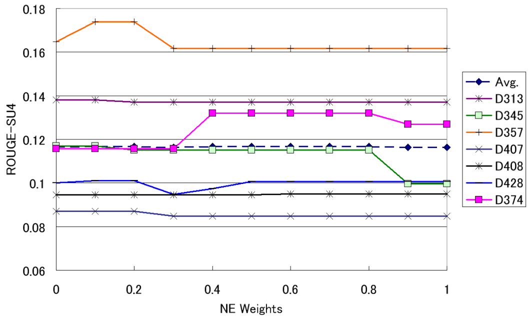

Figure 6: ROUGE-SU4 score change with named entity weights (number of clusters = 60, title weights = 0.1)

2002). For these topics, we changed the NE weights in expression (2) from 0 to 1 in steps of 0.1 and evaluated the ROUGE and BE scores. These results are shown in Figure 5, Figure 6, and Figure 7. Note that only topics for which scores changed within this interval appear in the graphs.

Table 6: Document sets utilizing named entity weights

| NE type | Document set |
|---|---|
| Person | d331,d374,d393,d436 |
| Country | d313,d357,d389,d435,d632 |
| Company | d345,d385 |
| Place | d407,d408,d428 |
| Time | d408 |
| Event | d436,d442,d683 |

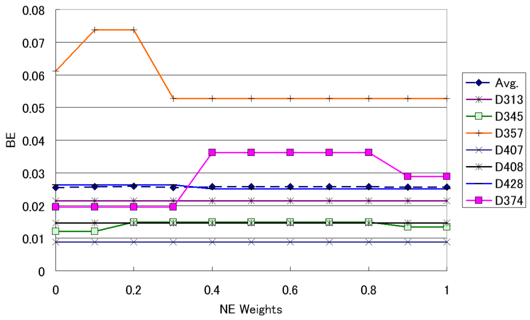

Figure 7: BE score change with named entity weights (number of clusters = 60, title weights = 0.1)

### 4.4 Subjectivity Analysis

Subjective information for sentences in source documents was automatically annotated using SVM$^{light}$ (Joachims, 2004). To assess the effectiveness of our subjective information annotation framework, we conducted a fivefold cross validation using the Multi-Perspective Question Answering (MPQA) corpus (Wiebe et al., 2005). This corpus contains 535 documents (10,657 sentences in total). Following Riloff's research (Riloff and Wiebe, 2003), 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. (Hatzivassiloglou and Wiebe, 2000), which contained 1,914 word entries. For features 6 to 9, we utilized the General Inquirer (Stone, 2000), which contained 1,168 word entries. Using SVM$^{light}$ 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 7.

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. We also categorized "narratives" as "comment", "positive", or "negative" types. (In the official submission version, we only categorized them as "comment" or

Table 7: Results of fivefold cross validation test of automatic subjectivity annotation (macro-average value)

| SVM | | |
|---|---|---|
| Accuracy | Precision | Recall |
| 0.602 | 0.610 | 0.657 |

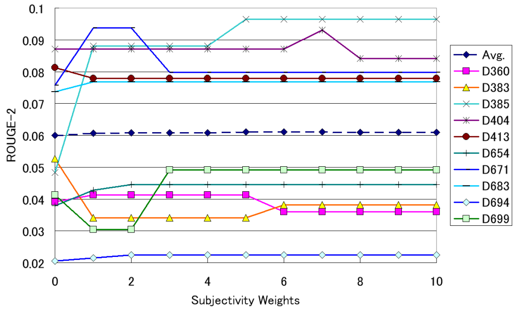

Figure 8: ROUGE-2 score change with subjectivity weights (number of clusters = 60, title weights = 0.1)

not. We re-implemented our question analysis module for these additional experiments.) The results of categorization are shown in Table 8.

Table 8: Document sets utilizing subjectivity weights

| Subjectivity Type | Document Set |
|---|---|
| Comment | d404,d683,d694,d699 |
| Positive | d360,d383,d385,d413,<br>d654,d671,d694,d699 |
| Negative | d385,d654,d699 |

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. For these 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 8, Figure 9, and Figure 10. Note that only the 10 topics appear in the graphs. In these figures, you can see that the results improved for several topics in Table 8.

## 5 Conclusions

In this paper, we discussed our TUT/NII system and its evaluation at DUC 2005. Although our approach performed in the middle rank among participants at DUC 2005, we were able to improve our system by tuning the number of clusters and title weights. We also presented topic-by-topic evaluation and pursued the effectiveness of our approach utilizing named entity analysis and subjectivity analysis of "narratives" in additional experiments.

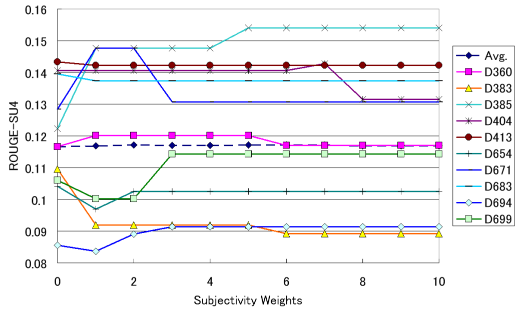

Figure 9: ROUGE-SU4 score change with subjectivity weights (number of clusters = 60, title weights = 0.1)

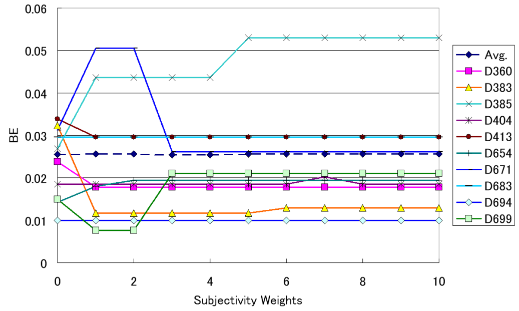

Figure 10: BE score change with subjectivity weights (number of clusters = 60, title weights = 0.1)

## 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

V. Hatzivassiloglou and J. M. Wiebe. 2000. Lists of manually and automatically identified gradable, polar, and dynamic adjectives. gzipped tar file. [cited 2005-8-26]. Available from: <http://www.cs.pitt.edu/ wiebe/pubs/coling00/coling00adjs.tar.gz>.

V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M. Y. Kan, and K. R. McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *Proc. of Workshop on Automatic Summarization at the Second Meeting of the North American Chapter of the Assoc. for Computational Linguistics (NAACL 2001)*, pages 41–49, Pittsburgh, PA, June.

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

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

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

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

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

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

M. J. Mana-Lopez, M. D. Buenaga, and J. M. Gomez-Hidalgo. 2004. Multidocument summarization: An added value to clustering in interactive retrieval. *ACM Trans. on Information Systems (TOIS)*, 22(2):215–241.

L. McKnight and P. Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proc. of the American Medical Informatics Assoc. (AMIA) Sympo.*, pages 440–444, Ottawa, Canada.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. of the 2004 Human Language Technology Conf. of the North American Chapter of the Assoc. for Computational Linguistics (HLT/NAACL 2004)*, The Park Plaza Hotel, Boston.

D. R. Radev, H. Jing, M. Stys, and D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

E. Riloff and J. M. Wiebe. 2003. 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.

Y. Seki, K. Eguchi, and N. Kando. 2004a. 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.

Y. Seki, K. Eguchi, and N. Kando. 2004b. Compact Summarization for Mobile Phones. In F. Crestani, M. Dunlop, and S. Mizzaro, editors, *Mobile and Ubiquitous Information Access*, volume 2954 of *Lecture Notes in Computer Science*, pages 172–186. Springer-Verlag, Heidelberg, Germany, February.

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

Y. Seki, K. Eguchi, and N. Kando. 2005. 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*, chapter 24, pages 317–336. Springer, Dordrecht, The Netherlands, October.

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

G. C. Stein, T. Strzalkowski, G. B. Wise, and A. Bagga. 2000. Evaluating summaries for multiple documents in an interactive environment. In *Proc. of the Second Int'l Conf. on Language Resources & Evaluation (LREC 2000)*, pages 1651–1657, Athens, Greece.

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

S. Teufel and M. Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

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

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