Opinion-focused Summarization and its Analysis at DUC 2006

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 approach to opinion-focused summarization, its results with the DUC 2006 data, and additional analysis. We extend our approach previously proposed from DUC 2005 to achieve summarization responding multiple questions, assuming that given narrative consists of multiple questions, by segmenting the "narrative" into questions. Our new approach is based on sentence extraction, where sentence type annotation is used for weighting, and frequencies of terms with sentiment polarities are taken into account if question types are appropriate for this. In addition, we selected 15 topics related to opinion-focused summarization and analyzed sentences in original source documents which correspond to model summaries.

1 Introduction

The purpose of our study is to build a multidocument summarizer on the basis of user-specified summary viewpoints. We have previously proposed the multidocument summarizer *v-SWIM*, which focuses on the facts, opinions, or knowledge described in documents, and we have experimented on Japanese document sets (Seki et al., 2005a). We reformulated our approach for English summarization, and presented the results at DUC 2005 (Seki et al., 2005b). In addition to this, we assessed the improvement rates in ROUGE (Lin, 2005) and BE (Hovy et al., 2005) scores for 10 subjectivity-related topics using subjective sentence extraction strategy. Subjectivity usually refers to some aspects of language description that express the author's or an authority's opinion, evaluation, or speculation (Wiebe et al., 2004). Although subjectivity analysis research has been mainly applied to date to measure the perceptions of the reputation of commercial products or movie titles on the Web, subjectivity analysis on newspaper articles is also important for information analysis in some domains, such as a political domain. This study attempts to clarify the feasibility of this in the context of text summarization.

We changed the summarization strategy for DUC 2006 to produce summaries discriminating multiple questions within the "narratives". We also extended the subjectivity annotation framework by expanding synonyms of subjectivity terms using WordNet (Miller et al., 2005). The reason for this change was to assess the effect of taking into account subjectivity more accurately and figure out how sensitive this effect is to subjectivity features of the questions. This strategy is more sensitive to questions and we got better responsiveness compared with our DUC 2005 system.

For post-DUC verification, we also selected 15 topics as opinion-focused topics[1] and analyzed model summaries in detail. We first created alignment between sentences in source documents and those in each of four model summaries according to judgments by one annotator. Then, we analyzed sentences in the model summaries and source documents from the viewpoints of sentence-level subjectivity. With this analysis, we clarified text structure of opinion-focused summaries and propose a new summarization strategy not only with opinionated sentence extraction, but also with the purpose for opinion-focused summarization.

This paper is organized as follows. In Section 2, we explain our multidocument summarization system. Section 3 details the official evaluations on the DUC 2006 data. Section 4 presents post-DUC analysis for opinion-focused summarization. Finally, we present our conclusions.

[1] In DUC 2006, we focused on opinion-asking topics such as "Why...important?". This selection criteria was almost overlapped with DUC 2005, but slightly different. Therefore, we changed terms as "opinion-focused" instead of "subjectivity-related" in DUC 2005.

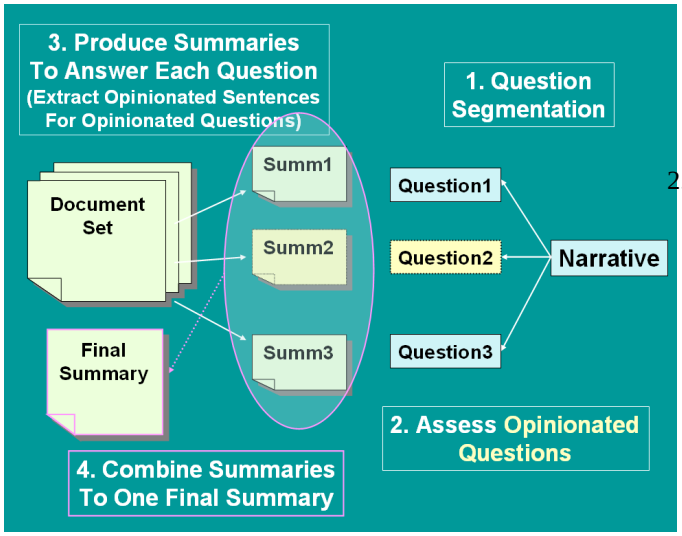

Figure 1: System Overview

2 System Overview

In DUC 2006, we changed our DUC 2005 summarization strategy slightly to assess the effect of opinionated sentence extraction. Our system overview is shown in Figure 1.

This summarization process consists of four steps: (1) "narrative" segmentation into questions; (2) assessment of opinionated questions; (3) production of summaries to answer each question; and (4) combination of summaries to form one summary. To produce summaries to answer questions, we divide the "narrative", which was given by DUC 2006 organizers as one type of user context information, into sentences.

Then, we automatically annotate opinionated properties in each question. Based on this information, we produce summaries that answer each question based on our approach used for DUC 2005. Following Stoyanov's hypothesis (Stoyanov et al., 2005), we suppose that opinionated questions relate to opinionated sentences in documents.

Each summary size was almost 250 words divided by the number of questions in the "narrative". Each summary was then combined into a single summary. All sentences in each summary are ordered chronologically within the source documents .

The summarization algorithm is based on sentence extraction with the paragraph-clustering algorithm that we used for DUC 2005 (Seki et al., 2005b). The detailed algorithm is described as follows:

1. Paragraph Clustering Stage
   (a) Source documents are firstly segmented into paragraphs, and then term frequencies (TF) are indexed for each paragraph.
   (b) Paragraphs are then clustered based on the Euclidean distance between feature vectors based on term frequency, using Ward's method.
2. Sentence Extraction Stage
   (a) The feature vectors for each cluster are computed using term frequencies (TF) and inverse cluster frequencies:

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

   Terms are stemmed using OAK (Sekine, 2002).
   (b) Clusters are ordered by the similarity between content words in "titles" and "narratives", provided for each topic by DUC 2006 organizers, and the cluster feature vectors.
   (c) Sentences within each cluster are weighted, based on content words in the "narratives" and "titles", heading words within the cluster, and TF values of the cluster. In addition, the "narratives" are used as statements to express information needs. The weight scheme is shown in equation (2):

$$\begin{aligned} W(s) = & \quad L(s) \times \\ & (a_1 \times Q(s) + a_2 \times H(s) + a_3 \times T(s) \\ & + a_4 \times \underline{N(s)} + a_5 \times \underline{S(s)} \\ & + a_6 \times \underline{Pos(s)} + a_7 \times \underline{Neg(s)}). \end{aligned} \quad (2)$$

   The weight $L(s)$ is 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 number of TF values in the cluster. The four underlined predicates, $N(s)$, $S(s)$, $Pos(s)$, and $Neg(s)$, are optional weight predicates based on analysis of **each question** in the "narrative"[2]. The term $N(s)$ is the frequency of named entity tags, matched against the information type from the analysis of the "narrative" and $S(s) = 1$ if sentence s is subjective, otherwise $S(s) = 0$. The $Pos(s)$ and $Neg(s)$ terms are the positive and negative term frequencies, respectively, of using adjective entries (Hatzivassiloglou and Wiebe, 2000) and the General Inquirer (Stone, 2000) in sentences.

[2]This is different from DUC 2005.

The coefficients a_1 to a_7 are parameters. For DUC 2006, they are set as follows: $a_1 = 0.8$; $a_2 = \frac{1}{\text{total number of heading words in the cluster}}$; $a_3 = 1$; $a_4 = 0.2$; $a_5 = a_6 = a_7 = 6$. [3].

(d) One sentence was extracted from each cluster, in cluster order, ordered by the similarity between content words in the "narratives" and "titles", and the cluster feature vectors, so as to reach the maximum number of words allowed divided by the number of questions ($\frac{250}{n}$ 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.

3 Evaluation

In this section, we present four types of evaluations of the TUT/NII team, as required for official submissions to DUC 2006: (1) responsiveness; (2) linguistic quality questions; (3) pyramid evaluation; and (4) ROUGE and BE scores.

3.1 Responsiveness

For the DUC 2006 data, responsiveness was evaluated by two schemes: (1) a content responsiveness score assigned by NIST assessors, based on the amount of information in the summary that helps to satisfy the information need expressed in the topic; and (2) the overall responsiveness score assigned by NIST assessors, based on both the readability of the summary and the amount of information in the summary that helps to satisfy the information need expressed in the topic. Results for the TUT/NII team's average scores and ranks are shown in Table 1. These results improved on the DUC 2005 results, for which our responsiveness scores were ranked 13th or 14th.

Table 1: Responsiveness for the TUT/NII team

	Responsiveness	
	Content	Overall
Score	2.82	2.42
Rank (of 34 systems)	7	6

3.2 Linguistic Quality Questions

For the DUC 2006 data, linguistic quality was evaluated using five criteria: (1) grammaticality; (2) nonredundancy; (3) referential clarity; (4) focus; and (5) structure and coherence. The results for our system are shown in Table 2[4]. Compared with the results in DUC 2005, the rank in the redundancy elimination dropped from the second rank to 17th rank. The reason for this seemed to be that the change in our system's algorithm introduced redundancy between answer summaries from different questions. This might be also the reason for the drop in the ranks from automatic evaluation.

Table 2: Quality evaluation for the TUT/NII team

Quality Criterion	Score	Rank (of 34 systems)
Grammaticality	3.58	18
Nonredundancy	4.26	17
Reference	3.22	14
Focus	3.52	22
Coherence	2.42	16
Average	3.4	16

3.3 Pyramid, ROUGE, and BE Evaluation

For DUC 2006, 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 the issue that different humans choose different content when writing summaries. Of the 34 participants, 21 teams' systems agreed (plus one baseline system) and were evaluated. ROUGE (Lin, 2005) and BE (Hovy et al., 2005) are automatic evaluation tools and can be used for reevaluation. Official evaluations were based on chunking results for our submitted summaries. The results of the official evaluation are shown in Table 3.

Table 3: Pyramid, ROUGE, and BE scores for the TUT/NII team

Evaluation Metrics	Scores	Rank (of 34 systems)
ROUGE-2	0.073	21
ROUGE-SU4	0.129	21
BE	0.036	20
Pyramid	0.180	14 (of 21 systems)

3.4 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 4. Our manually selected opinion-focused 15 topics are shown in bold face. The results for responsiveness are fairly good.

4 Post-DUC Analysis

As we explained in the introduction, we selected 15 topics for opinion-focused summarization. We then ana-

[3] These parameters were chosen based on a parameter-tuning exercise conducted for DUC 2005(Seki et al., 2005b).

[4] Average of all the 50 topics in DUC2006. For Tables 1 and 3, this is the same except for pyramid evaluation.

Table 4: Topic-by-topic evaluation for the TUT/NII team

Topic	Rank										
	Responsiveness		Linguistic Quality					Pyramid	ROUGE		BE
	Content	Overall	Q1	Q2	Q3	Q4	Q5		2	SU4	
D0601	24	20	1	1	8	17	10	20	19	25	27
D0602	29	17	7	1	8	22	26	-	21	17	25
D0603	11	2	1	1	9	5	6	5	16	18	14
D0604	1	15	5	26	8	5	3	-	6	17	16
D0605	14	6	14	1	1	4	1	19	25	21	32
D0606	1	11	24	1	1	6	3	-	14	17	21
D0607	13	25	9	1	1	2	1	-	12	13	23
D0608	21	8	12	3	19	15	10	8	18	18	22
D0609	3	2	6	1	12	22	9	-	24	16	3
D0610	8	1	9	14	10	28	2	-	30	28	29
D0611	16	8	17	1	28	31	25	-	25	27	17
D0612	13	10	31	8	27	28	23	-	23	28	21
D0613	2	5	32	1	21	6	9	-	23	30	18
D0614	10	8	21	7	5	11	7	16	15	11	6
D0615	12	2	1	1	4	10	11	3	6	9	9
D0616	9	1	1	1	9	1	1	15	20	20	17
D0617	12	11	1	28	1	15	9	5	19	17	18
D0618	1	1	6	1	3	8	5	-	14	16	20
D0619	6	12	12	1	9	9	4	-	22	21	17
D0620	5	14	1	1	10	6	5	9	26	22	31
D0621	1	6	1	4	26	7	27	-	32	25	25
D0622	4	7	22	2	23	15	31	-	15	29	23
D0623	5	16	19	1	6	1	13	-	19	23	12
D0624	1	3	25	24	8	2	5	16	17	14	15
D0625	10	7	6	21	3	1	7	-	4	6	14
D0626	2	1	1	1	1	1	1	-	17	8	21
D0627	3	3	1	1	5	8	2	18	15	15	14
D0628	15	4	33	4	32	20	18	18	30	18	25
D0629	1	2	28	1	3	22	17	12	10	15	13
D0630	16	18	10	1	11	32	29	15	29	28	18
D0631	2	3	7	4	7	1	30	15	6	8	3
D0632	9	2	13	1	17	1	1	-	31	30	27
D0633	9	7	1	34	18	3	17	-	18	24	24
D0634	10	6	1	19	4	1	13	-	15	11	8
D0635	16	11	1	1	14	26	20	-	26	23	20
D0636	24	9	25	27	1	13	18	-	27	30	21
D0637	1	1	1	29	7	8	6	-	14	6	16
D0638	1	7	13	20	9	11	14	-	12	5	3
D0639	1	32	32	5	32	1	30	-	27	22	29
D0640	8	5	8	14	5	20	11	19	16	16	11
D0641	1	2	10	1	5	14	10	-	30	32	29
D0642	5	9	7	1	7	29	12	-	28	28	26
D0643	23	27	15	1	1	1	7	3	16	19	20
D0644	2	1	17	1	1	1	1	-	20	13	23
D0645	9	5	20	23	7	2	9	6	10	15	9
D0646	4	8	30	11	16	13	18	-	16	13	13
D0647	1	2	18	1	6	6	1	8	15	23	3
D0648	19	1	29	4	2	8	3	-	18	28	16
D0649	12	3	5	26	23	1	3	-	8	4	8
D0650	12	3	12	10	14	1	2	6	22	25	17

lyzed these topics in detail. First, we assessed our system's performance. Then, we analyzed model summaries in detail and clarified the current problem in producing opinion-focused summarization.

4.1 Our System's Performance in Opinion-focused Summarization

4.1.1 Detection of Opinionated Questions

First, we show our automatic detection results for our opinionated question analyzer. Our opinionated question detection algorithm consists of two steps:

1. Detection of opinionated questions
   (a) *Opinionated* questions were detected by using the subjectivity classifier we used for DUC 2005 (Seki et al., 2005b). Feature words were expanded using WordNet (Miller et al., 2005) to see if their synonyms or hypernyms were subjective terms registered in adjective entries (Hatzivassiloglou and Wiebe, 2000) and the General Inquirer (Stone, 2000) or not.
   (b) Based on several keywords defined in DUC 2005 (Seki et al., 2005b), several nonopinionated questions were categorized into opinionated questions.
2. Detection of questions asked with positive or negative attitudes
   Questions asked with *positive* or *negative* attitudes were detected using the criteria of whether hypernyms of query terms contained the "good" or "bad" concept using WordNet.

The results of automatic annotation for questions in 15 opinionated topics were as follows. $< O >$, $< P >$, and $< N >$ tags represent *opinionated*, *positive*, and *negative* type questions. Opinionated keywords are shown in bold face.

1. D0601A: Native American Reservation System — pros and cons
   $< O >$Discuss conditions on American Indian reservations or among Native American communities.$< /O >< P >< N >$Include **the benefits and drawbacks** of the reservation system.$< /N >< /P >$ Include legal privileges and problems.
2. D0603C: wetlands value and protection
   $< O >$Why are wetlands important?$< /O >< O >$Where are they threatened?$< /O >$ What steps are being taken to preserve them? $< O >$What frustrations and setbacks have there been? $< /O >$
3. D0604D: anticipation of and reaction to the première of Star Wars Episode I — The Phantom Menace
   $< O >$How did fans, media, the marketplace, and **critics** prepare for and **react** to the movie?$< /O >< O >$Include preparations and **reactions** outside the United States.$< /O >$
4. D0606F: impacts of global climate change
   $< O >$What are the most significant impacts said to result from global climate change?$< /O >$
5. D0609I: Israeli West Bank settlements
   What impact have Israeli settlements in the West Bank had on the Israeli/Palestinian peace process? $< O >$What are the **reactions** of both parties and of the international community?$< /O >$
6. D0610A: home-schooling — pros and cons
   $< O >< P >< N >$What are the advantages and disadvantages of home schooling?$< /N >< /P >< /O >< O >$Is the trend growing or declining? $< /O >$
7. D0615F: evolution/creationism debate
   $< O >$What are the various perspectives in the U.S. public debate regarding the teaching of evolution, creation science, or intelligent design in public school science classes? $< /O >< O >$What are the key points and counterpoints expressed by people who hold each of those perspectives?$< /O >$
8. D0619A: gays and the GOP
   $< O >$Discuss the relationship between gays (homosexuals) and the Republican party.$< /O >$ How are Republicans courting gays? How do they alienate gays? Include discussion of the Log Cabin Republicans.
9. D0623E: anti-smoking laws
   $< O >$Describe anti-smoking laws passed or rejected worldwide which prohibit smoking in public places or work places.$< /O >< O >$Include any **arguments** used **for or against** such laws.$< /O >$
10. D0624F: Stephen Lawrence
    $< O >$What is known about the murder of Stephen Lawrence, his killers, the actions of the government, and the **reactions of the public**?$< /O >$
11. D0628A: ADD/ADHD diagnosis and treatment
    Describe ADD/ADHD. How is it diagnosed? $< O >$What kind of treatments are there?$< /O >< O >$Discuss the **controversies** surrounding its treatment.$< /O >$
12. D0635H: capital punishment in Texas during Governor Bush's administration
    $< O >$How has the administration of Governor George W. Bush implemented capital punishment and how are those policies **viewed** outside of Texas? $< /O >$
13. D0636I: issues between the UAW and American automobile manufacturers
    $< O >$What are the key issues under **discussion** between the 3 major American automobile manufacturers and the United Auto Workers (UAW)?$< /O >$
14. D0641E: global warming
    $< O >$Describe theories concerning the causes and effects of global warming and **arguments** against these theories.$< /O >$
15. D0642F: Hugo Chavez
    $< O >< P >< N >$What have been the key policies and **outcomes (good or bad)** of the Venezuelan Presidency of Hugo Chavez?$< /N >< /P >< /O >< O >$What **supportive or critical statements** or actions have come from Venezuelans or leaders of other countries?$< /O >$

4.1.2 Improvement by Weighting Opinionated Sentences

We also did experiments comparing the results weighting opinionated sentences in source documents according to properties of questions (opinionated, positive, and negative) with the results without weighting opinionated sentences. The results are shown in Table 5.

4.2 Analysis of Model Summaries

We produced an experimental dataset of source documents that corresponding model summaries for 15 topics. This dataset was produced by a native English assessor who was a translator. The sentences in the source documents were segmented using OAK (Sekine, 2002).

Table 6: Polarity term frequencies per sentence averaged over model summaries, where the source documents correspond to the summaries, and where the source documents do not correspond

Document Set	Type	# of Sentences	Polarity Adj.		Gradability Adj.		Dynamic adj.	Strong Terms		Weak Terms	
			Plus	Minus	Plus	Minus		Pos.	Neg.	Pos.	Neg.
D0601	M	68	0.235	0.324	0.485	0.353	0.088	0.221	0	0.074	0.294
	S/C	102	0.333*	0.245*	0.627**	0.569	0.078	0.353*	0.029	0.127	0.304*
	S/NC	3173	0.226	0.146	0.381	0.482	0.06	0.227	0.004	0.167	0.157
D0603	M	58	0.328	0.155	0.483	0.31	0	0.31	0	0.103	0.052
	S/C	70	0.671**	0.314*	0.857*	0.614	0.157	0.671	0	0.314**	0.243
	S/NC	1199	0.384	0.148	0.548	0.634	0.096	0.385	0.005	0.158	0.153
D0604	M	62	0.194	0.016	0.226	0.242	0.016	0.081	0	0.032	0.048
	S/C	87	0.253	0.069	0.667**	0.345	0.034	0.092	0	0.172	0.103
	S/NC	2490	0.249	0.071	0.349	0.289	0.054	0.112	0.004	0.122	0.065
D0606	M	59	0.22	0.186	0.322	0.508	0.034	0.102	0	0.119	0.136
	S/C	110	0.309	0.218	0.473	0.755	0.036	0.209	0	0.082	0.164
	S/NC	1585	0.329	0.175	0.499	0.662	0.036	0.18	0.002	0.11	0.138
D0609	M	51	0.235	0.196	0.451	0.098	0.039	0.137	0	0.196	0.039
	S/C	55	0.255	0.164	0.509	0.309	0.036	0.291	0	0.182	0.018
	S/NC	938	0.154	0.106	0.34	0.391	0.027	0.232	0.01	0.093	0.05
D0610	M	52	0.481	0.058	0.519	0.654	0.115	0.462	0.058	0.058	0.115
	S/C	98	0.418	0.163	0.541*	0.582*	0.194*	0.255	0.01	0.102	0.133
	S/NC	3475	0.302	0.11	0.378	0.422	0.09	0.206	0.01	0.092	0.095
D0615	M	47	0.298	0.128	0.213	0.66	0.021	0.234	0	0.106	0.085
	S/C	57	0.386	0*	0.316	0.474	0.07	0.228	0.018	0.193	0.018**
	S/NC	3793	0.239	0.089	0.337	0.386	0.045	0.148	0.007	0.12	0.079
D0619	M	50	0.26	0.1	0.46	0.52	0.12	0.24	0.04	0.06	0.04
	S/C	62	0.452*	0.097	0.742**	0.806**	0.129	0.419	0.097	0.129	0.048
	S/NC	2540	0.252	0.104	0.336	0.492	0.111	0.257	0.05	0.126	0.099
D0623	M	57	0.175	0.07	0.228	0.667	0.105	0.175	0.018	0.263	0.105
	S/C	75	0.24	0.173	0.56**	0.653*	0.107	0.187	0**	0.227	0.12
	S/NC	1358	0.202	0.108	0.351	0.468	0.058	0.196	0.015	0.161	0.074
D0624	M	60	0.217	0.117	0.317	0.35	0.167	0.167	0	0.167	0.117
	S/C	91	0.374	0.385*	0.769*	0.56*	0.198*	0.264	0	0.297	0.242*
	S/NC	1236	0.292	0.234	0.552	0.387	0.104	0.187	0.002	0.238	0.133
D0628	M	57	0.351	0.07	0.263	0.281	0.193	0.193	0	0.088	0.246
	S/C	63	0.492**	0.159	0.571**	0.444*	0.19*	0.206	0	0.127	0.476**
	S/NC	2280	0.222	0.111	0.36	0.284	0.069	0.178	0.001	0.12	0.196
D0635	M	59	0.237	0.22	0.169	0.458	0.017	0.237	0.051	0.102	0.22
	S/C	83	0.313	0.181	0.337	0.675**	0.072	0.301	0.072	0.157	0.361
	S/NC	3448	0.26	0.152	0.33	0.392	0.086	0.198	0.025	0.177	0.278
D0636	M	50	0.18	0.08	0.24	0.36	0.1	0.28	0	0.1	0.08
	S/C	63	0.27	0.222	0.54	0.603	0.127	0.492**	0	0.175	0.127
	S/NC	2565	0.24	0.142	0.47	0.451	0.08	0.232	0.002	0.233	0.122
D0641	M	47	0.511	0.191	0.638	0.851	0.043	0.17	0.021	0.128	0.128
	S/C	72	0.389	0.139	0.625	0.792	0.083	0.111*	0	0.056	0.069
	S/NC	1456	0.339	0.159	0.521	0.705	0.06	0.203	0.008	0.098	0.131
D0642	M	66	0.167	0.152	0.333	0.455	0.061	0.152	0	0.197	0.136
	S/C	87	0.437**	0.172	0.529	0.713*	0.138	0.299	0**	0.138	0.207
	S/NC	1065	0.251	0.16	0.438	0.515	0.117	0.243	0.023	0.131	0.133

$**$: statistically significant with t-test: $p < 0.01$, 95% confidence interval,
compared with TF of sentences in source docs not corresponding to model summaries
$*$: statistically significant with t-test: $p < 0.05$, 95% confidence interval,
compared with TF of sentences in source docs not corresponding to model summaries

Table 7: Polarity term frequencies per sentence as in Table 6, but using expanded polarity synonyms from WordNet

Document Set	Type	# of Sentences	Polarity Adj.		Gradability Adj.		Dynamic adj.	Strong Terms		Weak Terms	
			Plus	Minus	Plus	Minus		Pos.	Neg.	Pos.	Neg.
D0601	M	68	0.279	0.338	0.603	0.368	0.088	0.426	0.029	0.353	0.574
	S/C	102	0.382	0.314	0.745**	0.578	0.088	0.569*	0.039	0.451	0.608**
	S/NC	3173	0.299	0.201	0.476	0.519	0.063	0.417	0.017	0.407	0.331
D0603	M	58	0.431	0.172	0.5	0.31	0	0.431	0.052	0.293	0.207
	S/C	70	0.771*	0.457**	0.957**	0.614	0.157	0.829	0.043	0.671*	0.4
	S/NC	1199	0.508	0.205	0.644	0.671	0.096	0.651	0.016	0.452	0.334
D0604	M	62	0.21	0.016	0.242	0.258	0.016	0.113	0	0.097	0.113
	S/C	87	0.276	0.138	0.816**	0.391	0.046	0.207	0	0.414	0.402*
	S/NC	2490	0.328	0.122	0.439	0.317	0.061	0.247	0.008	0.276	0.222
D0606	M	59	0.271	0.203	0.356	0.508	0.034	0.186	0.051	0.22	0.322
	S/C	110	0.391	0.282	0.545	0.791	0.045	0.318	0**	0.409	0.373
	S/NC	1585	0.417	0.214	0.621	0.69	0.037	0.301	0.013	0.38	0.322
D0609	M	51	0.333	0.196	0.49	0.098	0.039	0.235	0	0.392	0.157
	S/C	55	0.364	0.164	0.545	0.364	0.036	0.473	0.036	0.491	0.145
	S/NC	938	0.251	0.118	0.405	0.412	0.03	0.396	0.027	0.333	0.109
D0610	M	52	0.519	0.077	0.558	0.673	0.115	0.577	0.096	0.269	0.212
	S/C	98	0.51*	0.204	0.622	0.592	0.194*	0.398	0.02	0.255	0.265
	S/NC	3475	0.361	0.163	0.46	0.451	0.092	0.333	0.024	0.268	0.199
D0615	M	47	0.489	0.128	0.404	0.681	0.021	0.447	0	0.362	0.191
	S/C	57	0.509	0.018**	0.474	0.561	0.088	0.544*	0.088	0.526*	0.14
	S/NC	3793	0.308	0.114	0.416	0.424	0.049	0.293	0.031	0.3	0.192
D0619	M	50	0.4	0.1	0.52	0.6	0.12	0.38	0.04	0.08	0.06
	S/C	62	0.726**	0.097	0.855**	0.823**	0.129	0.79*	0.097	0.194*	0.129
	S/NC	2540	0.316	0.128	0.394	0.552	0.115	0.481	0.073	0.317	0.213
D0623	M	57	0.263	0.105	0.281	0.719	0.105	0.228	0.053	0.526	0.246
	S/C	75	0.347	0.24	0.72**	0.747*	0.107	0.36	0.04	0.52	0.32
	S/NC	1358	0.285	0.151	0.455	0.517	0.06	0.37	0.029	0.414	0.272
D0624	M	60	0.25	0.217	0.35	0.4	0.167	0.183	0	0.35	0.283
	S/C	91	0.516	0.681*	0.901*	0.67**	0.198*	0.374	0.022	0.835*	0.505**
	S/NC	1236	0.361	0.387	0.64	0.454	0.112	0.299	0.009	0.57	0.275
D0628	M	57	0.421	0.105	0.404	0.298	0.193	0.491	0.035	0.246	0.561
	S/C	63	0.603**	0.27	0.714*	0.508*	0.19*	0.476	0.016	0.333	0.841**
	S/NC	2280	0.282	0.186	0.491	0.315	0.077	0.361	0.015	0.329	0.394
D0635	M	59	0.271	0.373	0.288	0.525	0.034	0.407	0.051	0.237	0.339
	S/C	83	0.373	0.301	0.446	0.88**	0.096	0.578	0.084	0.373	0.59
	S/NC	3448	0.299	0.193	0.387	0.497	0.088	0.376	0.035	0.369	0.455
D0636	M	50	0.22	0.08	0.32	0.36	0.1	0.6	0	0.36	0.28
	S/C	63	0.397	0.286	0.746	0.603	0.127	0.794**	0.016	0.46	0.444
	S/NC	2565	0.344	0.178	0.602	0.47	0.085	0.445	0.011	0.455	0.3
D0641	M	47	0.745	0.34	0.894	0.894	0.064	0.298	0.021	0.383	0.34
	S/C	72	0.514	0.222	0.736	0.833	0.097	0.292	0**	0.181	0.181
	S/NC	1456	0.429	0.22	0.626	0.736	0.073	0.338	0.019	0.266	0.28
D0642	M	66	0.227	0.197	0.424	0.47	0.061	0.303	0.015	0.348	0.288
	S/C	87	0.437	0.207	0.575	0.736*	0.138	0.494	0.011	0.379	0.333
	S/NC	1065	0.309	0.191	0.547	0.552	0.12	0.4	0.027	0.306	0.288

$**$: statistically significant with t-test: $p < 0.01$, 95% confidence interval,
compared with TF of sentences in source docs not corresponding to model summaries
$*$: statistically significant with t-test: $p < 0.05$, 95% confidence interval,
compared with TF of sentences in source docs not corresponding to model summaries

Table 5: Changes in automatic evaluation scores by weighting opinions

Document Sets	Author	ROUGE		BE
		2	SU4	
D0601	A	0.0010	0.0012	-0.0001
D0603	C	0.0059	0.0018	0.0047
D0604	D	0.0078	0.0097	0.0017
D0606	F	0.0010	-0.0008	-0.0034
D0609	I	-0.0070	0.0000	-0.0006
D0610	A	-0.0039	-0.0039	0.0018
D0615	F	-0.0050	-0.0072	0.0033
D0619	A	-0.0059	-0.0084	-0.0031
D0623	E	0.0121	0.0054	0.0007
D0624	F	0.0118	0.0025	0.0030
D0628	A	0.0040	0.0107	0.0017
D0635	H	-0.0128	-0.0031	-0.0050
D0636	I	-0.0142	-0.0080	0.0004
D0641	E	-0.0139	-0.0155	-0.0050
D0642	F	-0.0166	-0.0092	-0.0054

4.2.1 Statistical Analysis

First, we counted the positive (respectively negative) term frequencies of the original sentences that corresponded to model summaries and those that did not correspond, respectively. The results are shown in Table 6 and Table 7. The first five items were counted using the adjective entries (Hatzivassiloglou and Wiebe, 2000). The latter four items were counted using General Inquirer (Stone, 2000). In Table 7, term frequencies were counted by expanding terms using WordNet (Miller et al., 2005).

To clarify the effectiveness of summarization parameters, we also applied the linear regression analysis by setting summarization parameters in each sentence as independent variables and by setting sentences aligned with model summaries or not as 1/0 in dependent variable. The results are shown in Table 8 and summarized as follows:

- Opinionatedness parameter was effective for 12 topics.
- Sentence position parameter was effective for 15 topics.
- Length parameter was effective for 15 topics.

This result shows that lengthy sentences were preferred to extract in opinion-focused summarization.

4.2.2 Content Analysis

For our proposed system, we have three discussion points: (1) question analyzer, (2) opinion extraction, and (3) combined effect of document genre.

1. opinionated question analyzer
   Our opinionated question analyzer was based on the polarity terms, their synonyms, and predefined keywords. This approach sometimes leaded to miscategorization in case that "the topic of question" contained polarity terms. For example, the first question in D0623 contained polarity terms such as "pass" or "reject", but these terms were used as "topic" and the main asking content is the description (not opinion). In contrast to this, the second question in D0623 was to ask for opinions. To implement more accurate system, we must discriminate these cases.

2. opinion extraction
   For opinion-focused topics, the summaries were categorized into two groups: (A) asking sentiments such as evaluations for G. W. Bush's policies concerning punishment (D0635); and (B) asking comments such as arguments against theories concerning global warming (D0641). Our approach was based on polarity term frequencies and sometimes was not effective for the latter case. In future, we plan to extend our approach to solve this problem.

3. combined effect of document genre
   The "news story" document genre was annotated to English source documents beforehand. Following researches in (Seki et al., 2005a), we extend the opinion-focused summarization framework to English summarization. Sentences in the "news stories" document genre were biased negatively. We assessed the optimal weighting parameter as '1', using manual annotation of opinionated sentences in the DUC 2006 dataset. We also set the multiplying weighting parameter for sentences in the "news source" document genre as '0'. Based on these parameters, we investigated the feasibility of our approach using automatic annotation in the DUC 2006 dataset. The combined effect of sentence type and document genre in ROUGE and BE scores is shown in Table 9. We found that the combined weighting of sentence type and document genre was effective for opinion-focused summarization in English.

Table 9: Effect of document genre and sentence type in English summarization

System Type	ROUGE-2	ROUGE-SU4	BE
Baseline	0.07400	0.12886	0.03162
Upper Ceiling (Manual Annotation)	0.07711	0.13221	0.03423
System Results (Auto. Annotation)	0.07430	0.13093	0.03153

5 Conclusions

We participated in DUC 2006 to clarify the effectiveness of our opinion-focused summarization. For responsive-

Table 8: Summarization Effect Parameter for 15 topics

Document Set	Opinionated-ness using SVM	Position	Length	Similarity to Questions				heading terms	tfidf
				Q1	Q2	Q3	Q4		
D0601	0.015	-0.198	0.024	0.015	-0.006	-		-0.009	-0.016
D0603	-0.096	-0.244	0.030	0.035	-0.017	0.004	-0.004	0.096	-0.015
D0604	0.037	-0.211	0.070	0.001	0.022			-0.025	0.08
D0606	0.050	-0.273	0.077	0.027				0.032	0.004
D0609	0.018	-0.252	0.038	0.007	-0.034			0.09	-0.019
D0610	0.016	-0.175	0.032	0.032	-0.005			-0.018	-0.02
D0615	0.062	-0.147	0.067	-0.013	0.005			0.016	-0.03
D0619	0.027	-0.191	0.067	0.062	-	0.001	0.032	0.001	-0.001
D0623	0.052	-0.291	0.061	0.056	0.100			0.045	0.018
D0624	0.056	-0.296	0.030	-0.020				-0.016	0.043
D0628	0.002	-0.185	0.019	-	0.072	0.009	0.030	0.036	0.042
D0635	0.006	-0.154	0.024	-0.008				0.045	0.012
D0636	-0.044	-0.184	0.074	0.012				0.011	-0.036
D0641	0.039	-0.214	0.064	-0.010				-0.026	0.012
D0642	-0.002	-0.306	0.040	0.113	-0.053			-0.027	0.007

ness evaluation, our result was satisfactory. For intrinsic evaluations, our result was not so good. We selected 15 opinion-related topics and did the experiment to assess the effectiveness of our approach. We continued this analysis to clarify the problem.

Acknowledgments

This work was partially supported by the Grants-in-Aid for Young Scientists (B) (#18700241), Young Scientists (A) (#17680011) and the Grants-in-Aid for Exploratory Research (#16650053) both from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

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>.

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/>.

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

G. A. Miller, C. Fellbaum, R. Tengi, S. Wolff, P. Wakefield, H. Langone, and B. Haskell. 2005. WordNet [online]. [cited 2005-8-26]. Available from: <http://wordnet.princeton.edu/>.

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.

Y. Seki, K. Eguchi, and N. Kando. 2005a. Multi-document viewpoint summarization focused on facts, opinion and knowledge. In J. G. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theories and Applications*, chapter 24, pages 317–336. Springer, Dordrecht, The Netherlands, December.

Y. Seki, K. Eguchi, N. Kando, and M. Aono. 2005b. 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.

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

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

V. Stoyanov, C. Cardie, and J. Wiebe. 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proc. of the Human Language Technology Conf. / Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, October.

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