How small a distinction among summaries can the evaluation method identify?

Yoshio Nakao
Fujitsu Laboratories Ltd.
KEN60, 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-8588, Japan
ynakao@jp.FUJITSU.com

Abstract

This paper describes the summarization methods used by the team FLAB (gid040) for text summarization tasks in the NTCIR-2 workshop. The focus is on the effectiveness of an extrinsic evaluation based on relevance assessment in information retrieval with reference to the evaluation results obtained for a task B. The team FLAB submitted two types of summaries for the task: baseline summaries and thematic hierarchy based summaries. Statistical analysis of the results of five types of summaries comprising these two types of submitted summaries and three types of official baseline summaries suggests that the difference between the former two was too small for the evaluation to identify. This suggests that the task might better be performed under the condition that only a few subjects could complete the given task, e.g., with only a certain short time limit given for assessment.

Keywords: *text summarization, topic identification, information retrieval, thematic hierarchy.*

1 Introduction

This paper describes the summarization methods used by the team FLAB (gid040) for text summarization tasks in the NTCIR-2 workshop. The discussion focuses on the effeciveness of the task B evaluation, which was an extrinsic evaluation based on relevance assessment in information retrieval with reference to the evaluation results.

At least two issues are involved in a task-based evaluation. One is the need to establish an objective measure for text summarization techniques, the other is the need to construct an ideal summary to support relevance assessment. Since the former was the main focus of the text summarization evaluation project reported here, the team FLAB submitted two sets of summaries for each text summarization task. These summaries were generated by different summarization methods, and my objective was to determine how the evaluation measured the difference between the summaries.

The summarization methods to be compared used the following two algorithms: a keyword-based sentence extraction algorithm[7] and a thematic-hierarchy-based (TH-based) sentence extraction algorithm[6].

My main interest during this workshop was to measure the effect of topic identification based on the thematic hierarchy. I was particularly interested in measuring smaller differences of summaries with task-based evaluation than Tombros et al.[9] measured in comparing a query-biased summary with one comprising the first few sentences of a source text. For this purpose, I tailored a TH-based summarization algorithm and a baseline algorithm that did not use the thematic hierarchy of the text for either of the tasks performed (i.e., tasks A or B).

The remainder of this paper is organized as follows: Section 2 describes the two sentence extraction algorithms separately. Section 3 describes the text summarization method used for each evaluation task and reports the evaluation results obtained. Section 4 discusses the effectiveness of the evaluation, and concluding remarks are given in Section 5.

2 Sentence extraction algorithms

2.1 Keyword-based sentence extraction algorithm

The keyword-based sentence extraction algorithm uses a given set of keywords (e.g., keywords in a title or a query) as the seeds of a summary. It generates the summary by extracting sentences that contain a large number of keywords from the text. The feature of this algorithm is that it extracts a small set of sentences that includes as many different keywords as possible. Intuitively speaking, it generates such a summary as in which every keyword occurs once. The main purpose of the algorithm is to provide a summary list of retrieved documents for IR systems. It was initially

designed to summarize newspaper articles and to generate a brief summary that contains information that complements the article headline.

Figure 1 shows the basic algorithm for keyword-based sentence extraction[7]. The algorithm uses a seed list that initially consists of given keywords. It evaluates sentences in a document with the following four scores relevant to the occurrence of keywords in the seed list:

1. Type of keywords included in the sentence,
2. Total number of keywords included in the sentence,
3. Type of content words other than keywords included in the sentence, and
4. Total number of content words other than keywords included in the sentence.

The algorithm checks these scores in that order and selects the sentence with the highest scores. In general, it selects the sentence that most widely covers the keywords. The seed list is then modified by removing the keywords included in the selected sentence. The procedure repeats until the seed list becomes empty.

seed list ← given keywords
summary ← ϕ
while seed list is not empty
do
  Evaluate sentences based on seed list.
  Select a sentence with the maximum score
    (exit if no sentence can be selected).
  Put the selected sentence into the summary.
  Remove the keywords appearing in the selected sentence from the seed list.
od

Figure 1. Keyword-based sentence extraction algorithm.

2.2 TH-based sentence extraction algorithm

The TH-based sentence extraction algorithm[6] decomposes a source text into segments of approximately the same size based on lexical cohesion[1] measured by term repetitions. It then extracts two or three sentences from the beginning portion of each segment that probably indicate the contents of their subsequent parts. (Hereafter, these sentences that the TH-based algorithm extracts are referred to as "boundary sentences".) The boundary sentences for a segment typically consists of a heading and a topic introducing sentence. In a prior experiment[6] using larger texts as test data, about half of the first boundary sentences were found to be identical to the headings in the original text.

The feature of this algorithm is that it can flexibly extract topics of various grading according to the required summary size. For example, if a ten sentence summary is required, it decompose the source text into about five segments of approximately the same size and extracts two or three sentences from each segment. The main purpose of the algorithm is to provide a one-page summary of a very long document (e.g., a 100-page book) and to indicate major topics in that document with example sentences that include appropriate keywords related to the topics.

The basic part of the TH-based summarization algorithm is thematic hierarchy detection. The feature of the thematic hierarchy detection algorithm is that it decomposes a text into segments of approximately the same size and thus can systematically detect thematic textual segments of different sizes, ranging from segments slightly smaller than the entire text to segments of about one paragraph.

The thematic hierarchy detection algorithm decomposes a text in a similar way as the TextTiling algorithm[2] does. The algorithm calculates a cohesion score at fixed-width intervals in a source text. A cohesion score is calculated based on the lexical similarity of two adjacent blocks of a fixed size by the following formula:

$$c(b_l, b_r) = \frac{\Sigma_t w_{t,b_l} w_{t,b_r}}{\sqrt{\Sigma_t w_{t,b_l}^2 \Sigma_t w_{t,b_r}^2}} \quad (1)$$

where b_l and b_r are the textual block in the left and right windows, respectively, w_{t,b_l} is the frequency of term t for b_l, and w_{t,b_r} is the frequency t for b_r. It then detects thematic boundaries according to the minimal points of a four-item moving average (arithmetic mean of four consecutive scores) of the cohesion score series.

The algorithm repeats this procedure with varying window width. The smaller the window width used is, the smaller the segments are that will be detected (see [6] for more details). For the evaluation tasks, the window width was set at a minimum of 40 words and then doubled each time (e.g., 40, 80, 160, ..., and 640 word widths) until the width exceeded half the document size.

The resulting thematic hierarchy of a text consists of several layers that individually correspond to the segmentation detected with a specific window width. That is, the root node corresponds to the entire text, and nodes on the bottom layer are atomic segments

detected with the minimum window width. A node in an upper layer corresponds to a segment detected with a larger window width and comprises one or more segments that were detected with a slightly smaller window width.

For a given summary size, the algorithm extracts boundary sentences from the root node to the nodes on the bottom layers as much as the given size allows. That is, the boundary sentences in the beginning portion of the source text are initially extracted as the boundary sentences of the entire text[1]. Other boundary sentences of segments in lower layers are extracted in order of their layer lavel in the hierarchy and their appearance in the text as far as the size limitation allows.

3 Summarization methods and evaluation results

At the NTCIR-2 workshop, participating summarization systems were evaluated based on the following three subtasks: two subtasks (tasks A1 and A2) for intrinsic evaluation and one subtasks (task B) for extrinsic evaluation. Newspaper articles (or editorials) were used as test data for all the evaluation subtasks.

In tasks A1 and A2, participating summarization systems were evaluated by comparing their submitted summaries with human-made summaries. Given a set of 30 source articles and a set of summary length specifications, each participant submitted a set of summaries. Participating summarization systems were then evaluated based on similarity between the submitted summaries and human-made summaries (task A1) and on relative ranking in comparison with two human-made summaries (in free style and by important part extraction) and an official baseline summary (task A2).

In task B, participating systems were evaluated based on a relevance assessment in terms of information retrieval, in a similar manner as that used in the SUMMAC project[5] for the ad-hoc task evaluation. Each participant submitted a set of summaries of given documents (30 articles for dry runs and 50 articles for formal runs) prepared for several retrieval topics (10 topics for dry runs, and 12 topics for formal runs). Three subjects for each set of summaries then assessed the relevance of each document for a given topic with reference to submitted summaries. Finally, participating summarization systems were evaluated based on accuracy of judgement and time consumed for assessment.

[1]This does not mean that the first sentence of the source text is always extracted, but that at least one of the first few sentences of the source text is always extracted.

3.1 Task A

For the A1 and A2 tasks, two types of summaries were submitted as follows.

- TH-based summary: The TH-based summary for the task A was generated by a combination of the TH-based sentence selection algorithm and the keyword-based sentence extraction algorithm. For each article, sentences were extracted based on the TH-based algorithm up to 80% of the required summary size, and the remaining sentences were then extracted based on the keyword-based algorithm. For the keyword-based sentence extraction, the seed list of a summary was constructed with keywords (nouns, verbs, and adjectives for the formal runs; nouns for the dry runs) extracted from the headline of the source article and the boundary sentence candidates that had not been extracted because of the extraction amount restriction (i.e., 80% of the required size).

- Baseline summary: The baseline summary for the task A was generated by the keyword-based sentence extraction algorithm. For the formal runs, keywords extracted from the headline and the lead paragraph of the source article were used as seeds; for dry runs, keywords from the headline of the source article were used as seeds. To make a summary of a given size, the sentence extraction procedure (the main loop of the keyword-based algorithm) was stopped if the summary size exceeded the limit, or else it repeatedly executed the procedure while reconstructing the seed list. Once the main loop was completed (i.e., the seed list became empty), the seed list was reconstructed by adding extra keywords extracted from the sentences that had been selected by that time.

Tables 1 and 2 show the evaluation results of tasks A1 and A2 respectively. The tables also show two types of official baseline summaries provided by the NTCIR-2 committee. The TF-based summary is the one that was generated by extracting sentences with many high-frequency terms. The lead-based summary is the one that was generated by extracting the first few sentences of an article.

The figures in Table 1 are average identical rates of sentences extracted by a specific system (the first column) to those extracted manually. In the dry runs, the figure of the TH-based summary exceeded that of the baseline summary. However, this may be due to the fact that a TH-based summary always includes the beginning portion of the source text without depending on how the thematic hierarchy of the source text was detected (see the last portion of the previous section). To compete with this feature, the baseline summary in the formal runs was generated by weighing the lead

Table 1. Task A1 results

Summary type	F-score	
	Formal run	Dry run
TH-based	.416	.540
Baseline	.449	.536
TF-based	.391	.525
Lead-based	.434	.554

Table 2. Task A2 results

Summary type	Impression score				Cosine	
	20R	20C	40R	40C	free	ext.
Formal run						
TH-based	3.00	3.17	2.73	3.03	.526	.561
Baseline	2.97	3.10	3.10	3.13	.522	.552
TF-based	3.20	3.27	2.77	3.07	.516	.549
Lead-based	–	–	–	–	.481	.513
Dry run						
TH-based	2.63	3.00	2.63	2.97	–	.586
Baseline	3.23	3.60	2.73	3.03	–	.546
TF-based	3.37	3.70	3.27	3.40	–	.569
Lead-based	–	–	–	–	–	.582

paragraph of the source article. As a result, the figure of the baseline summary of the formal runs exceeded that of the TH-based summary and also exceeded that of the lead-based summary, which was one of the official baselines provided by the NTCIR-2 committee and was made by extracting a given amount of sentences from the beginning of the source article.

The *cosine* columns in Table 2 list average scores of lexical similarity, i.e., cosine values of weighted term (term frequency multiplied by inverse document frequency) vectors, between a specific summary and a human-made summary. The *free* column lists those scores based on free-style summarization by persons and the *ext.* column lists those based on important-part extraction by persons. The TH-based summary scores were generally higher than the baseline summary scores. Specifically, in the dry runs the TH-based score was higher for both 20% summaries and 40% summaries. In the formal runs, however, the TH-based summary scores were higher than the baseline summary scores for 40% summary, but for 20% summary the opposite was true.

These results suggests that the TH-based summarization algorithm successfully identifies certain major (sub)topics that do not appears at the beginning of the source article, and that the sentence extraction strategy of TH-based summarization differs from that of human beings. These results may relate to the feature that the TH-based summarization algorithm tends to identify headings as boundary sentences.

The *impression-score* columns in Table 2 list average ranking scores (1 being best and 4 being worst) of a specific summary in comparison with two human-made summaries (in free style and by important part extraction) and an official baseline summary (lead-based summary). The *20R* column lists the resulting scores obtined through readability assessment of 20% summaries, and the *20C* column lists those obtained through content-appropriateness assessment of 20% summaries. The *40R* and the *40C* columns lists the same scores for 40% summaries. These scores also suggests that the TH-based summary is better able to successfully identify certain major (sub)topics. In addition, the difference in scores between TH-based summaries and baseline summaries appears to suggest that boundary sentence extraction improves the readability of summaries. However, the best TH-based summary score was only 2.73 at *40R*, where it ranked first six times, second two times, third 16 times, and fourth six times. This suggests that more elaborate techniques, such as sentence generation or information fusion, are required for summarization methods such as these to approach the level of human summarization.

3.2 Task B

For task B two types of summaries, TH-based and baseline summaries, were submitted. Both of them were tailored to support relevance assessment. Both summaries were query-biased ones generated by keyword-based sentence extraction algorithm based on given keywords (Figure 1).

The difference between the two summary types is keyword selection. The baseline method uses keywords (nouns, verbs, and adjectives for the formal runs, nouns for the dry runs) extracted from the headline and the lead paragraph[2] of the source article and the description and the narrative field of the query. The TH-based method uses keywords extracted from the headline of the article, the boundary sentences of the nodes in the layer immediately below the root node of the thematic hierarchy, and from the description field of the query. In short, the TH-based summary uses the boundary sentences in place of the sentences in the lead paragraph and does not use the narrative field of a query, which describes the requirement for retrieval in detail. As a result, the average length of baseline summaries and that of TH-based summaries are almost the same (see Table 3). The table also shows three types of official baseline summaries (or original text) provided by the NTCIR-2 committee. The TF-based summary is the one that was generated by extracting sentences with many high-frequency terms and query terms. The lead-based summary is the one that was generated by extracting the first few sentences of an article.

[2] In the dry runs, no keywords extracted from the lead paragraph were used.

Table 3. Summary length.

Summary type	Avg. length in characters (condensing rate)			
	Formal run		Dry run	
TH-based	263	(35%)	178	(38%)
Baseline	266	(35%)	153	(33%)
Full-text	819	(108%†)	463	(100%)
TF-based	254	(33%)		–
Lead-based	175	(22%)		–

† A full-text summary in the formal runs was attached with the headline of the soruce article.

Table 4 summarizes the result of the evaluation. The *level A* column lists recall (the *rec.* columns) and precision (the *prec.* columns) rates, and F-scores[3] (the *F* columns) using the level A (relevant) judgements as correct data; the *level B* column lists those using the level B (partially relevant) judgements as correct data. The *time* column lists the average time for assessment of a topic (50 articles assessed for formal runs, 30 articles for dry runs) and an article.

As I had anticipated, all the evaluation scores of the TH-based summary (i.e., accuracy of the judgements with TH-based summaries) exceed those of the baseline summary. However, the difference is not significant. That is, statistical analysis of the results of five types of summaries (the two types of submitted summaries and three types of official baseline summaries) indicates that there is no significant difference among the TH-based, the baseline, full text, and the TF-based summaries. Consequently, the following section examines the results by using the lead-based summary as a baseline measure.

4 Discussion

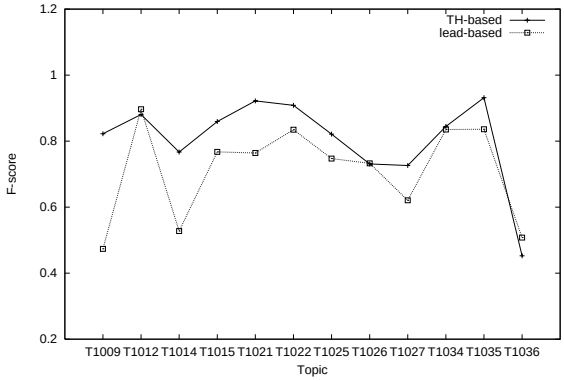
4.1 Statistical analysis

This section examines the evaluation results of task B using the lead-based summary as a baseline measure and discusses the effectiveness of task-based evaluation.

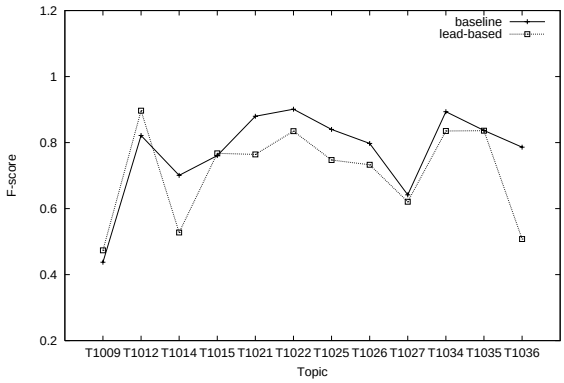
Figure 2 plots F-scores by topic for level-B judgement for four types of summaries (TH-based, baseline, full text, and TF-based) in comparison with the lead-based summary. These graphs, especially the part for the first three topics, show that these five types of summaries can be divided into two groups: 1) TH-based and TF-based summary group, and 2) baseline, full text, and lead-based summary group.

Table 5 also shows differences between the summary groups using two-factor analysis of variance (ANOVA)[3] concerning F-score distributions. It

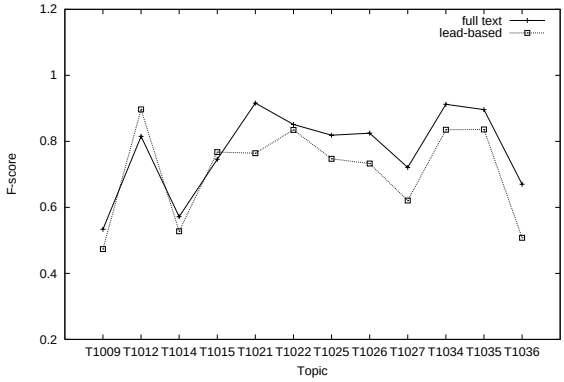
[3] $\frac{2*recall*precision}{(recall+precision)}$



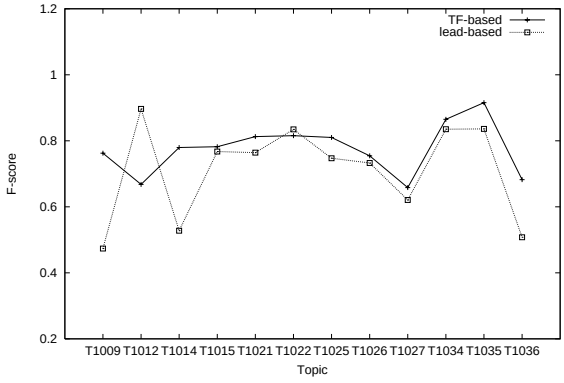
(a) TH-based summary vs. lead-based summary



(b) Baseline summary vs. lead-based summary



(c) Full-text article vs. lead-based summary



(d) TF-based summary vs. lead-based summary

Figure 2. Average F-score for topics (level B assessment).

Table 4. Result of task based evaluation.

(a) Formal run

Summary type	Level A			Level B			Avg. time	
	F	rec.	prec.	F	rec.	prec.	for topics	for articles
TH-based	.768	.849	.741	.805	.752	.923	9'31"	11.4"
Baseline	.749	.824	.738	.775	.719	.913	9'16"	11.1"
Full-text	.751	.843	.711	.773	.736	.888	13'46"	16.2"
TF-based	.738	.798	.724	.776	.700	.913	8'44"	10.5"
Lead-based	.731	.740	.766	.712	.625	.921	7'32"	9.0"

(b) Dry run

Summary type	Level A			Level B			Avg. time	
	F	rec.	prec.	F	rec.	prec.	for topics	for articles
TH-based	.838	.915	.796	.857	.869	.864	5'40"	11.3"
Baseline	.822	.840	.838	.814	.786	.891	6'01"	12.0"
Full-text	.842	.913	.796	.867	.878	.874	8'46"	17.5"
TF-based	.794	.804	.827	.802	.757	.895	5'12"	10.4"
Lead-based	.773	.781	.813	.781	.744	.883	4'25"	8.8"

shows the interaction effects between the summary factor (i.e., summary type) and the topic factor for each pair of summary types. The underlined figures (percentage point values) correspond to significant interaction effects. Since a significant interaction effect indicates that the relationship between the F-scores of the corresponding pair of summaries depends a great deal on the topic factor, those pairs of summaries with significant interaction effects should have different features. That is, the difference between either the TH- or TF-based summary and the lead-based summary is larger than that between either the baseline or the full-text summary and the lead-based summary. In this experiment, both the baseline summary and the lead-based summary are the ones that mainly relate to the first few sentences of an article. Thus, the TH- and TF-based summaries probably include information that does not appear in the lead part of an article.

Table 7 shows the evaluation results of the significance of the difference shown in Table 4 based on the rank-sum test (Wilcoxon test). The reason for using the rank-sum test is that significant interaction effects were found (see Table 5) and that F-score distribution was not normal. The table indicates that there is no significant difference among summaries in level A judgements, but that there are significant differences between either the TH-based, the baseline, or the full-text summary and the lead-based summary in level B judgements. The difference between the TH-based and the lead-based summaries is particularly significant (1% level of significance). These results suggest that a TH-based summary includes more topics (or subtopics) than a lead-based or baseline summary, and that these topics are useful for relevance assessment in full-text searches.

In comparison with the results of the rank-sum test, the ANOVA results with respect to the main effects of the summary factor (Table 6) differs in that they show a significant difference between the TF- and lead-based summaries. The difference between the rank-sum test and ANOVA results suggests that F-scores of the TF- and the lead-based summaries greatly differs from each other, but that the medians of their F-scores are nearly identical (or not significantly different). This indicates that a TF-based summary probably provides information that is very different than that provided by the lead-based summary but that that information is not always useful for relevance assessment.

One possible interpretation of these results is that both the TH-based method and the baseline method can extract major topics of an article more accurately than the TF-based method can. The fact that the TH-based summary and the baseline summary made higher scores for the level A judgements than those made by the TF-based summary also supports to this interpretation.

However, we cannot safely judge the difference between a TH-based summary and a baseline summary solely on the basis of these statistical examinations, although the difference in the significance level shown in Table 7 suggests that a TH-based summary may possibly provide more useful information than a baseline summary can. One possible reason for this is that the relevance assessment task was too easy. This suggests that the task might better be performed under a condition in which only a few subjects could complete the given task, e.g., with only a certain short time limit given for assessment. The next section briefly discusses the relation between assessment accuracy and the time taken for assessment.

Table 5. Interaction effect between summary type and topic.

Level A judgement

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	1.7	$(p > .05)$	2.1	$(p < .05)$
Baseline	1.2	$(p > .05)$	.81	$(p > .05)$
Full-text	1.9	$(p > .05)$	–	
TF-based	3.7	$(\mathbf{p < .01})$	2.0	$(p > .05)$

Level B judgement

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	2.1	$(p < .05)$	2.2	$(p < .05)$
Baseline	1.4	$(p > .05)$	.59	$(p > .05)$
Full-text	.57	$(p > .05)$	–	
TF-based	2.9	$(p < .05)$	1.6	$(p < .05)$

Table 6. Main effect of summary type.

Level A judgement

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	3.3	$(p > .05)$	.53	$(p > .05)$
Baseline	.60	$(p > .05)$	.007	$(p > .05)$
Full-text	.56	$(p > .05)$	–	
TF-based	.10	$(p > .05)$	.33	$(p > .05)$

Level B judgement

Summary type	F-value (percentage points)			
	Lead-based		Full-text	
TH-based	17	$(\mathbf{p < .01})$	1.6	$(p > .05)$
Baseline	7.1	$(p < .05)$	.004	$(p > .05)$
Full-text	5.5	$(p < .05)$	–	
TF-based	7.9	$(\mathbf{p < .01})$	.009	$(p > .05)$

Table 7. Rank-sum test results.

Level A judgement

Summary type	Z-value (percentage points)			
	Lead-based		Full-text	
TH-based	.63	$(p > .05)$	.23	$(p > .05)$
Baseline	.07	$(p > .05)$	.44	$(p > .05)$
Full-text	.73	$(p > .05)$	–	
TF-based	.35	$(p > .05)$	.87	$(p > .05)$

Level B judgement

Summary type	Z-value (percentage points)			
	Lead-based		Full-text	
TH-based	2.8	$(\mathbf{p < .01})$	1.1	$(p > .05)$
Baseline	1.9	$(p < .05)$	.005	$(p > .05)$
Full-text	1.8	$(p < .05)$	–	
TF-based	1.5	$(p > .05)$	.38	$(p > .05)$

Table 8. Detailed T1014 assessment information.

Summary type	Avg. time for articles	Recall	Precision
TH-based	12”, 45”, 10”	.88, .58, .82	.78, .83, .77
Baseline	23”, 20”, 34”	.70, .46, .67	.79, .83, .92
Full-text	18”, 18”, 21”	.49, .60, .24	.84, .91, .80
Lead-based	9”, 10”, 13”	.42, .36, .36	.74, .86, 1.0

4.2 Assessment time

In the formal runs, the full-text assessment accuracy made a relatively low score. The average F-score for level B full-text assessment was the second lowest; only that for the lead-based summary was lower. One possible reason for the low full-text F-score is that subjects using full text for relevance assessment were in such a hurry that they missed some minor topics related to queries. The following data seem to support this possibility.

The average assessment time for a full-text article in the formal runs was one second shorter than that in the dry runs, in spite of the fact that the average length of the target articles in the formal runs (about 760 characters) was much longer than that in the dry runs (about 460 characters). In addition, as Figure 2(c) shows, the pattern of the full-text F-scores is very similar to that of the lead-based summary scores. This may relate the experimental condition of the formal runs that a full-text article was presented with an article headline. With a headline, subjects may have judged without reading carefully the detail of the article.

For example, Table 8 shows detailed information on the assessment of topic T1014 “不良債権処理 (Handling bad loans made by financial institutions)” by three subjects. As the table shows, the full-text F-score of assessment of this topic was almost the same as that for the lead-based summary, and the average assessment time for a full-text article was shorter than that for either the TH-based or the baseline summary. However, as indicated by the average time difference between the TH-based and the baseline summaries in the dry runs shown in Table 4(b), it sometimes took longer to assess a shorter summary, particularly a very short one. Thus, Table 8 may include examples of such cases.

4.3 Problem related to query interpretation

There are other important factors to be considered, such as experimental conditions related to subjects or query. This section describes a problem related to a query interpretation that was found by chance in analyzing the results for the TH-based summary.

As shown in Figure 2(a), the lowest TH-based summary F-scores recorded among 12 topics were these for topic T1036 “労働時間短縮 (Reducing working hour)”. One reason for this appears to be a difference in interpretation of this topic between the subjects using the TH-based summary and those who had created the judgement data. Many assessment-target articles on this topic reported or discussed “裁量労働制 (free working hour system)”. However, there were no notes in the description or narrative field of this topic about the relevance to these articles since the system was not

popular when the topic was defined[4]. Accordingly, some subjects judged such articles as irrelevant.

5 Conclusion

This paper described the summarization methods used by the team FLAB for text summarization tasks in NTCIR-2, with the main focus on the effectiveness of extrinsic evaluation based on relevance assessment in information retrieval. Statistical analysis of the results obtained for five types of summaries, including TH-based and baseline summaries submitted by the team FLAB, suggests that the difference among these two types of summaries was too small for the evaluation method to identify. This suggests that the task might better be performed under a condition in which only a few subjects could complete the given task. However, other important factors remain to be considered, such as experimental conditions related to subjects or queries. It is my hope that more comprehensive analysis of the evaluation results, including the results of other systems, will reveal some important aspects of summarization to support relevance assessment, especially concerning the issue of preferable summary features to support relevance assessment.

References

[1] M. A. Haliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[2] M. A. Hearst. Multi-paragraph segmentation of expository text. In *Proc. of the 32nd Annual Meeting of Association for Computational Linguistics*, pages 9–16, 1994.

[3] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proc. of SIGIR'93*, pages 329–338. the Association for Computing Machinery, 1993.

[4] H. Kimoto, Y. Ogawa, T. Ishikawa, Y. Masunaga, T. Fukushima, T. Tanaka, H. Nakawatase, I. Keshi, J. Toyoura, T. Miyauchi, Y. Ueda, K. Matsui, T. Kitani, S. Miike, T. Sakai, T. Tokunaga, H. Tsuruoka, and T. Agata. Construction of a test collection for the evaluation of japanese information retrieval systems. In *Proc. of the 15th Symposium on Informatics*. the Science Council of Japan et al., 1998. (In Japanese. See also http://www.ulis.ac.jp/˜ishikawa/bmir-j2/).

[5] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrsi, T. Firmin, M. Chizanowski, and B. Sundheim. The TIPSTER SUMMAC text summarization evaluation (final report). Technical Report MTR 98W0000138, MITRE Corporation, Virginia, Oct 1998. (http://www.itl.nist.gov/div894/894.02/ related_projects/tipster_summac/final_rpt.html).

[6] Y. Nakao. An algorithm for one-page summarization of a long text based on thematic hierarchy detection. In *Proc. of the 38th Annual Meeting of Association for Computational Linguistics*, pages 302–309, 2000.

[4]The queries used for the task were a subset of those in the IREX IR text collection [8] and were originally developed by the BMIR project (February 1993 to March 1996) [4].

[7] R. Ochitani, Y. Nakao, and F. Nishino. Goal-directed approach for text summarization. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 47–50, 1997.

[8] S. Sekine and H. Isahara. IREX project overview. In *Proc. of the IREX Workshop*, pages 7–12. IREX Committee, 1999. (http://www.csl.sony.co.jp/person/sekine/IREX/).

[9] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proc. of SIGIR'98*, pages 2–10. the Association for Computing Machinery, 1998.