Evaluation of Phrase-Representation Summarization based on Information Retrieval Task

Mamiko OKA Yoshihiro UEDA
Industry Solutions Company,
Fuji Xerox Co., Ltd.
430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa, Japan, 259-0157
oka.mamiko@fujixerox.co.jp Ueda.Yoshihiro@fujixerox.co.jp

Abstract

We have developed an improved task-based evaluation method of summarization, the accuracy of which is increased by specifying the details of the task including background stories, and by assigning ten subjects per summary sample. The method also serves precision/recall pairs for a variety of situations by introducing multiple levels of relevance assessment. The method is applied to prove phrase-represented summary is most effective to select relevant documents from information retrieval results.

Introduction

Summaries are often used to select relevant documents from information retrieval results. The goal of summarization for such "indicative" use is to serve fast and accurate judgement. We have developed the concept of the "at-a-glance" summary, and its realization in the Japanese language - "phrase-representation summarization" - to achieve this goal (Ueda, et al. 2000). We have conducted an evaluation experiment to verify the effectiveness of this summarization method.

There are two strategies for evaluating summarization systems: intrinsic and extrinsic (Jing, et al. 1998). Intrinsic methods measure a system's quality mainly by comparing the system's output with an "ideal" summary. Extrinsic methods measure a system's performance in a particular task. The aim of the phrase-representation summarization method is fast and accurate judgement in selecting documents in information retrieval. Thus, we adopted a task-based method to evaluate whether the goal was achieved. Task-based evaluation has recently drawn the attention in the summarization field, because the assumption that there is only one "ideal" summary is considered to be incorrect, and some experiments on information retrieval were reported (Jing, et al. 1998) (Mani, et al. 1998) (Mochizuki and Okunura 1999). However, there is no standard evaluation method, and we consider that there are some shortcomings in the existing methods. Thus, we have developed an improved evaluation method and carried out a relatively large experiment.

In this paper, we first give an overview of the phrase-representation summarization method. We then consider the evaluation method and show the result of an experiment based on the improved method to demonstrate the effectiveness of phrase-representation summarization.

1 Phrase-Representation Summarization

Most automatic summarization systems adopt the "sentence extraction" method, which gives a score to every sentence based on such characteristics as the frequency of a word or the position where it appears, and selects sentences with high scores. In such a way, long and complex sentences tend to be extracted. However, a long and complex sentence is difficult to read and understand, and therefore it is not a suitable unit to compose a summary for use in selecting documents.

To avoid the burden of reading such long and complex sentences, we have developed the phrase-representation summarization method, which represents the outline of a document by a series of short and simple expressions ("phrases") that contain key concepts. We use the word "phrase" to represent the simplicity characteristic[1] in a word.

The phrase-represented summary has the following characteristics.

(1) At-a-glance comprehension
Because each unit is short and simple, the user is able to grasp the meaning at a glance.

(2) Adequate informativeness
Unlike extracted sentences, phrases created by this method are not accompanied by information unnecessary for relevance judgement.

(3) Wide coverage of topics
Units composing a summary are relatively short, and point various positions of the original text. Therefore, even a generic summary includes various topics written in a document.

A phrase-represented summary is generated as follows.

1. Syntactic analysis to extract the relationships between words
2. Selection of an important relation (two word sequences connected by an arc) as a "core"
3. Addition of relations necessary for the unity of the phrase's meaning (*e.g.*, essential cases)
4. Generation of the surface phrase from the selected relations

An important relation is selected by considering both the importance of a word and that of a relation between words. For example, predicate-argument relations are considered important and noun-modifier relations are given low importance scores. Steps [2] to [4] are repeated until specified amount of phrases are obtained. Before selecting a new "core," the scores for the already selected words are decreased to suppress overuse of the same words.

Fig. 1 shows a sample summary created from a news article[2] put on WWW. The underlined words constitute the core relation of each phrase.

... acquire chemical toner business[3]
Fuji Xerox ... acquires chemical toner business of Nippon Carbide Industries Co., Inc. ...
... new chemical toner that contributes to reduce cost in laser printers and to lower energy consumption ...
... strengthen...supplies business ...
manufacturing facilities of Hayatsuki Plant, ...
... uniform...each particle ...

Fig.1: A sample summary

2 Evaluation Method

2.1 Summarization Methods to be Compared

In this experiment, we compare the effectiveness of phrase-represented summaries to summaries created by other commonly used summarization methods. From the viewpoint of the phrase-represented summary, we focus the comparison of the units that constitute summaries. The units to be compared with phrases are sentences (created by the sentence extraction method) and words (by the keyword enumeration method). We also compare "leading fixed-length characters," which are often used as substitutes for summaries by WWW search engines. The generation method for each summary is described as follows.

(A) Leading fixed-length characters: extract the first 80 characters of the document body.

(B) Sentence extraction summarization: select important sentences from a document. The importance score of each sentence is calculated from the simple sum of the importance scores of the words in a sentence (Zechner 1996).

(C) Phrase-representation summarization: described in Chapter 1.

(D) Keyword enumeration summarization: list up important words or compound nouns.

[1] The word "phrase" as used here is not used in the linguistic sense, but an expression for "short" and "simple." In Japanese, there is no rigid linguistic distinction between a "phrase" and a "clause."

[2] The original text in Japanese and its outline in English can be seen in the following URL.
http://www.fujixerox.co.jp/release/2000/0224_purchase.html (in Japanese)
http://www.fujixerox.co.jp/headline/2000/0308_ntonerbiz.html (in English)

[3] This phrase lacks the subject because the original sentence lacks it. Cases are usually omitted in Japanese if they can be easily inferred.

In (B), (C), and (D), the same method of calculating the importance scores of words is used in common, and lengths of summaries are kept to be 60 to 80 characters.

As you can see each summary is generic, *i.e.* not created for any specific queries. Because the phrase-representation summarization method is applied to Japanese, we examine the effectiveness of these four methods in Japanese.

2.2 Previous Work

The best-known example of task-based evaluation on information retrieval is the ad hoc task in the TIPSTER Text Summarization Evaluation Conference (SUMMAC) (Mani, et al. 1998). Hand (1997) details the proposed task-based evaluation under TIPSTER. Jing, et al. (1998) describe how various parameters affect the evaluation result through a relatively large task-based experiment. Evaluation conferences like SUMMAC are not yet held for Japanese summarization systems[4]. Mochizuki and Okumura (1999) applied the SUMMAC methodology to Japanese summarization methods for the first time. Most previous experiments are concerned with SUMMAC, accordingly the methods resemble each other.

2.3 Framework of Evaluation

The framework of task-based evaluation on information retrieval is shown in Fig. 2.

Task-based evaluation in general consists of the following three steps:

(1) Data preparation: Assume an information need, create a query for the information need, and prepare simulated search results with different types of summaries.

(2) Relevance assessment: Using the summaries, human subjects assess the relevance of the search results to the assumed information needs.

(3) Measuring performance: Measure the accuracy of the subjects' assessment by comparing the subjects' judgement with the correct relevance. The assessment process is also timed.

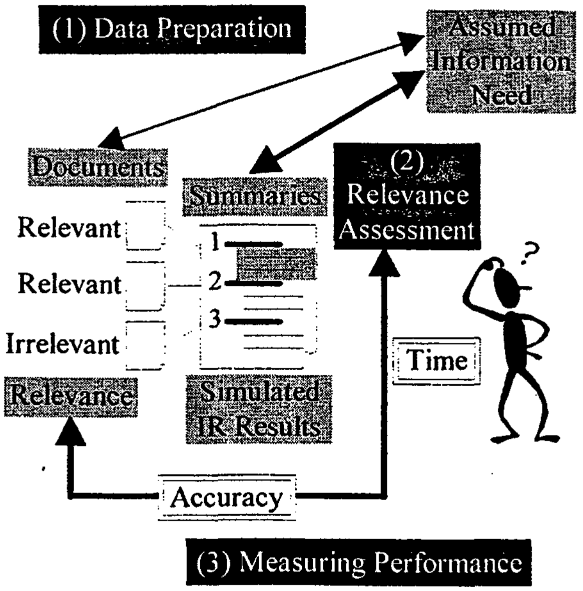

Fig.2: Framework of Task-Based Evaluation

We designed our evaluation method through detailed examination of previous work. The consideration points are compared to the SUMMAC ad hoc task (Table 1). A section number will be found in the "*" column if we made an improvement. Details will be discussed in the section indicated by the number in the next chapter.

3 Improvements

3.1 Description of Questions

To assess the relevance accurately, the situation of information retrieval should be realistic enough for the subjects to feel as if they really want to know about a given question. The previous experiments gave only a short description of a topic. We consider it is not sufficiently specific and the interpretation of a question must varied with the subjects.

We selected two topics ("moon cake" and "journey in Malay. Peninsula") and assumed three questions. To indicate to the subjects, we set detailed situation including the motivation to know about that or the use of the information obtained for each question. This method satisfies the restriction "to limit the variation in assessment between readers" in the MLUCE Protocol (Minel, et al. 1997).

[4] It is planning to be held in 2000. Further information is in the following URL. http://www.rd.nacsis.ac.jp/~ntcadm/workshop/ann2p-en.html

For each topic, ten documents are selected from search results by major WWW search engines, so that more than five relevant documents are included for each question. The topics, the outline of the questions, the queries for WWW search, and the number of relevant documents are shown in Table 2. The description of Question-a2 that was given to the subjects is shown in Fig. 3.

> One day just after the mid-autumn festival, my colleague Mr. *A* brought some moon cakes to the office. He said that one of his Chinese friends had given them to him. They looked so new to us that we shared and ate them at a coffee break. Chinese eat moon cakes at the mid-autumn festival while Japanese have dumplings then. Someone asked a question why Chinese ate moon cakes, to which nobody gave the answer. Some cakes tasted sweet as we expected; some were stuffed with salty fillings like roasted pork. Ms. *B* said that there were over fifty kinds of filling. Her story made me think of a question:
>
> What kinds of filling are there for moon cakes sold at the mid-autumn festival in Chinese society?

Fig. 3: An example of question (Question-a2)

3.2 Number of Subjects per Summary Sample

In the previous experiments, one to three subjects were assigned to each summary sample. Because the judgement must vary with the subjects even if a detailed situation is given, we assigned ten subjects per summary sample to reduce the influence of each person's assessment. The only requirement for subjects is that they should be familiar with WWW search process.

3.3 Relevance Levels

In the previous experiments, a subject reads a summary and judges whether it is relevant or irrelevant. However, a summary sometimes does not give enough information for relevance judgement. In actual information retrieval situations, selecting criteria vary depending on the question, the motivation, and other circumstances. We will not examine dubious documents if sufficient information is obtained or we do not have sufficient time, and we will examine dubious documents when an exhaustive survey is required. Thus, here we introduce four relevance levels L0 to L3 to simulate various cases in the experiment. L3, L2, and L1 are considered relevant, the confidence becomes lower in order. To reduce the variance of interpretation by subjects, we define each level as follows.

L3: The answer to the given question is found in a summary.

L2: A clue to the answer is found in a summary.

L1: Apparent clues are not found, but it is probable that the answer is contained in the whole document.

L0: A summary is not relevant to the question at all.

If these are applied to the case of the fare of the Malay Railway, the criteria will be interpreted as follows.

L3: An expression like "the berth charge of the second class is about RM15" is in a summary.

L2: An expression like "I looked into the fare of the train" is in a summary.

L1: A summary describes about a trip by the Malay Railway, but the fare is not referred in it.

3.4 Measures of Accuracy

In the previous experiments, precision and recall are used to measure accuracy. There are two drawbacks to these measurements: (1) the variance of the subjects' assessment makes the measure inaccurate, and (2) performance of each summary sample is not measured.

Precision and recall are widely used to measure information retrieval performance. In the evaluation of summarization, they are calculated as follows.

$$\text{Precision} = \frac{\text{Documents that are actually relevant in } S}{\text{Documents that are assessed relevant by a subject } (S)}$$

$$\text{Recall} = \frac{\text{Documents that are assessed relevant by a subject}}{\text{Relevant documents}}$$

In the previous experiments, the assessment standard was not fixed, and some subjects tended to make the relevant set broader and others narrower. The variance reduces the significance of the average precision and recall value. Because we introduced four relevance levels and showed the assessment criteria to the subjects, we can assume three kinds of relevance document sets: L3 only, L3 + L2, and L3 + L2 + L1. The set composed only of the documents with L3 assessment should have a high precision score. This case represents a user wants to know only high-probability information, for example, the user is hurried, or just one answer is sufficient. The set including L1 documents should get a high recall score. This case represents a user wants to know any information concerned with a specific question.

Precision and recall represent the performance of a summarization method for certain question, however they do not indicate the reason why the method presents higher or lower performance. To find the reasons and improve a summarization method based on them, it is useful to analyze quality and performance connected together for each summary sample. Measuring each summary's performance is necessary for such analysis. Therefore, we introduce the *relevance score*, which represents the correspondence between the subject judgement and the correct document relevance. The score of each pair of subject judgement and document relevance is shown in Table 3.

By averaging scores of all subjects for every sample, summary's performances are compared. By averaging scores of all summary samples for every summarization method, method's performances are compared.

Table 1: Experimental Method

Consideration point	SUMMAC ad hoc task	Our method	
(1) Data preparation phase			
Document source	Newspaper (TREC collection)	WWW	
Question	Selected from TREC topics	Newly created, including the detailed situation	3.1
Number of questions	20	3	
Number of documents per question	50	10	
Summary type	User-focused summary	Generic summary	
Summarization systems or methods	11 systems	4 methods that utilize different units	
(2) Relevance assessment phase			
Subject	21 information analysts	40 persons who usually use WWW search	
Number of subjects assigned to each summary sample	1 or 2	10	3.2
Relevance levels	2 levels (Relevant or irrelevant)	4 levels (L0, L1, L2, L3)	3.3
(3) Performance measuring phase			
Measure of accuracy	Precision and recall	Precision and recall Relevance score	3.4

Table 2: Topics and Questions

	Topic	Outline of question	Query	Number of relevant documents
Q-a1	Moon cake	What is the origin of the Chinese custom to have moon cakes in the mid-autumn?	moon cake & mid-autumn	5
Q-a2		What kinds of fillings are there in moon cakes?		6
Q-b	Journey in Malay Peninsula	About the train between Singapore and Bankok: How much does it cost? How long does it take? What is the difference in the equipment by the class? (A document containing one of these information is regarded as relevant.)	Singapore & Bankok & railway	7

Table 3: Relevance Score

Document relevance	Relevant	Relevant	Relevant	Relevant	Irrelevant	Irrelevant	Irrelevant	Irrelevant
Subject judgement	L3	L2	L1	L0	L0	L1	L2	L3
Score	10	8	5	-2	2	-5	-8	-10

4 Experiment Results

4.1 Accuracy

4.1.1 Precision and Recall

The precision and recall are shown in Fig. 4, and the F-measure is shown in Fig. 5. The F-measure is the balanced score of precision and recall, calculated as follows:

$$\text{F-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Figures 4 and 5 show that the phrase-represented summary (C) presents the highest performance. It satisfies both the high precision and the high recall requirements. Because there are various situations in WWW searches, phrase-representation summarization is considered suitable in any cases.

4.1.2 Relevance Score

The relevance score for each question is shown in Fig. 6. The phrase-represented summary (C) gets the highest score on average, and the best in Question-a2 and Question-b. For Question-a1, though all summaries get poor scores, the sentence extraction summary (B) is the best among them.

4.2 Time

The time required to assess relevance is shown in Fig. 7. The time for Question-a is a sum of the times for Questions a1 and a2. In the Question-a case, phrase-represented summary (C) requires the shortest time. For Question-b, leading fixed-length characters (A) requires the shortest time, and this result is different from the intuition. This requires further examination.

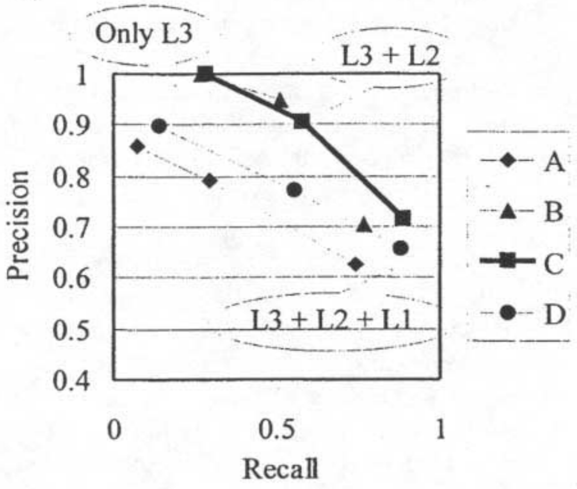

Fig. 4: Precision & Recall

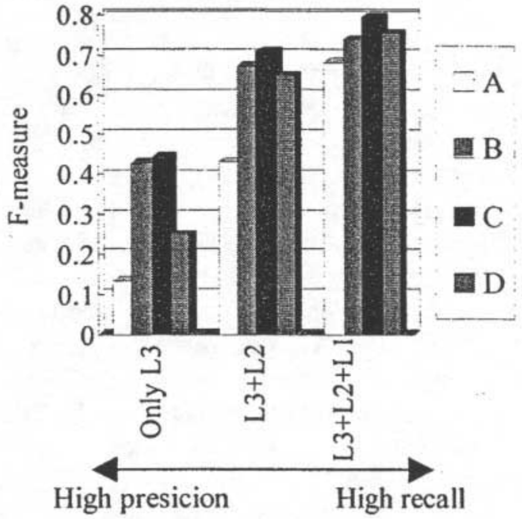

Fig. 5: F-measure

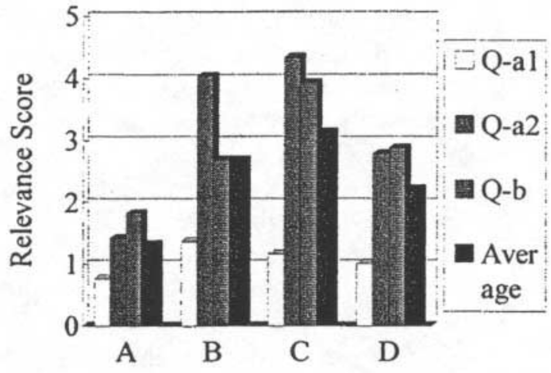

Fig. 6: Relevance Score

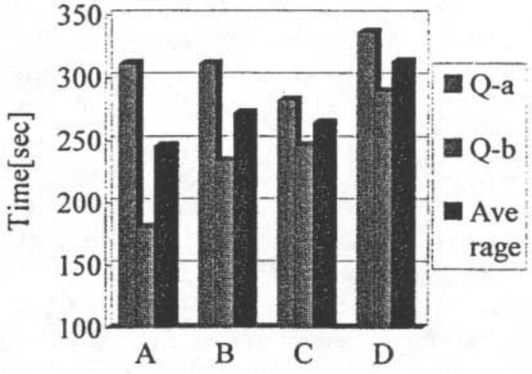

Fig. 7: Time

Table 4: Summaries Containing Clues

Question	Q-a1				Q-a2				Q-b			
Summarization method	A	B	C	D	A	B	C	D	A	B	C	D
Number of the relevant document	5				6				7			
Number of the summaries that contain clues	0	0	2	3	1	3	6	5	0	3	5	6
Average relevance score of the above summaries	-	-	7.2	6.77	5.4	9.43	6.83	5.56	-	7.73	8	6.43

5 Discussion

Here we analyze the experiment result from multiple viewpoints: the constituent unit of summaries and the characteristics of questions and documents in Section 5.1 and 5.2. We then discuss advantages of our experimental method in Section 5.3, and language dependency of the experiment result in Section 5.4.

5.1 Comparison of Constituent Unit

The units that constitute a summary may affect the judging process; if the unit is long, the number of units appeared in a summary may decrease and the summary contains fewer key concepts in the original document. We counted the number of the summaries that contain the clues to the given questions (see Table 4). The average numbers are 0.3, 2.0, 4.3 and 4.7 for (A) fixed-length characters, (B) sentences, (C) phrases and (D) words, respectively. The phrase-represented summary (C) and the keyword enumeration summary (D) widely cover the topics, and they are about twice as wide as the sentence extraction summary (B). The leading fixed-length characters (A) contain very few clues and this fact supports that this summary presents the worst performance (see Section 4.1).

In order to compare a summary's performance with a summary's quality, we calculate the average relevance score of summaries that contain clues. These scores are also shown in Table 4. The average score represents the informativeness of each summary. Table 4 shows that the sentence extraction summary (B) and the phrase-represented summary (C) get relatively high scores, but vary with the question. This is because sentences and phrases are sufficiently informative in most cases, but sentences tend to contain unnecessary information, and phrases tend to lack necessary information. The keyword enumeration summary (D) gets a relatively low score. This is because a word is not sufficiently informative to enable judgement of whether it is clue to the answer, and relations among words are lacked.

These analyses support the two characteristics of the phrase-represented summaries described in Chapter 1, that is, adequate informativeness and wide coverage of topics.

5.2 Influence of Question and Document

The most suitable summarization method may depend on the type of question and/or document. In the experiment results (see Section 4.1.2), the sentence extraction summary (B) and the phrase-represented summary (C) get the highest relevance score. Therefore, here we focus on those two summarization methods and consider the influence of questions and documents.

In selecting questions, we consider two factors may affect performance. One is which unit an answer is expressed in. Another is whether clue words easily come to mind.

If an answer is expressed as a relation of a predicate to its arguments, the phrase-representation summarization may be suitable. Question-a2 and Question-b are of this case. If an answer is expressed as compound relations, *e.g.*, reason-consequence relations or cause-result relations, the sentence extraction summarization may be required. And, if an answer is expressed in complex relations of sentences, any summarization method of the four is not suitable. Questions that ask historical background or complicated procedures are examples of this kind, *e.g.*, Question-a1.

As for another factor, if clue words easily come to mind, the phrase-represented summary is suitable for any unit in which an answer is expressed. This is because the clues are found more easily in short phrases than in long sentences.

In selecting documents, whether a question is relevant to the main topic of a document affects the performance, because we use generic summaries. By sentence extraction summarization, the answer is extracted as a summary only when the question is relevant to the main topic. Phrase-represented summary is able to cover topics more widely, for example, one of the main topics or detailed description of each topic (see Section 5.1). Because the characteristic of the document is independent of the question, which summaries cannot be predicted, and thus the phrase-represented summary will give better results.

Through these discussions, we conclude that the phrase-representation summarization is suitable for various cases, while the sentence extraction summarization is for only some restricted cases. Though the samples of questions and documents are relatively few in our experiment, it is sufficient to show the effectiveness of the phrase-representation summarization.

5.3 Advantages of our Experimental Method

Our experimental method has the following advantages.

(1) More exact assessment

(2) Serves precision/recall pairs for a variety of situations

(3) Helps further analysis of problems of a summarization method

5.3.1 More Exact Assessment

Our experimental method provides more exact relevance assessment in the following ways.

(a) More detailed description of a question

We asked the subjects to assess the relevance of full documents to each question after the experiment. Result shows that 93% of the subject judgements match the assumed relevance, while only 69% match in the same kind of assessment in SUMMAC. The percentage that all judgements per document agreed the assumed relevance is 33%, while only 17% in SUMMAC. This is because the subjects comprehended the questions correctly by given detailed information about the situation.

(b) More subjects assigned per summary sample

We assigned ten subjects to each summary sample, while only one or two subjects were used in SUMMAC. We examined the difference of judgement between the average of ten subjects and the first subject of the ten. Result shows that 47% of the first subject's judgement differ more than one level from the average. This proves that the assessment varies from one subject to another, even if a detailed situation is given.

(c) Finer levels of relevance

We introduced four levels of relevance, by which ambiguity of relevance can be expressed better.

5.3.2 Serves precision/recall pairs for a variety of situations

According to the four levels of relevance, we assume three kinds of relevance document sets. This enables to plot the PR curve.

In evaluation conferences like SUMMAC, various summarization methods that are developed for different purposes must be compared. Using such a PR curve, each method can be compared in a criterion that matches its purpose.

5.3.3 Helps further analysis of problems of a summarization method

We have introduced the relevance score, which allows each summary to be evaluated. Using this score, we can analyze the extrinsic evaluation result and the intrinsic evaluation result connected together, for example, an evaluation result based on information retrieval task and that based on Q & A task using the same questions. Through such analyses, the text quality of summaries or the adequate informativeness can be examined. We ourselves got a lot of benefit from the analysis to find problems and improve the quality of the summary.

5.4 Language dependency

Though experiment method may be applied to any other languages, we must consider the possibility that our result depends on the language characteristics. Japanese text is written by mixing several kinds of characters; Kana characters (Hiragana and Katakana) and Kanji (Chinese) characters, and alphabetic characters are also used. Kanji characters are mainly used to represent concept words and Hiragana characters are used for function words. The fact that they play the different roles makes it easy to find the full words. Also Kanji is a kind of ideogram and each character has its own meaning. Thus, most words can be expressed by 1 to 3 Kanji characters to make short phrases (15 - 20 characters) sufficiently informative.

Though the basic algorithm to create phrase-represented summary itself can be applied to other languages by replacing its analysis component and generation component, similar experiment in that language is required to prove the effectiveness of the phrase-represented summary.

Conclusion

We proposed an improved method of task-based evaluation on information retrieval. This method can be used to evaluate the performance of summarization methods more accurately than is possible by the methods used in previous work. We carried out a relatively large experiment using this method, the results of which show that phrase-representation summarization is effective to select relevant documents from information retrieval results.

Acknowledgements

We would like to thank our company members who gave valuable suggestions and participated in the experiment.

References

Hand, T. F. (1997). "A Proposal for Task-based Evaluation of Text Summarization Systems." In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp31-38.

Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. (1998). "Summarization Evaluation Methods: Experiments and Analysis." In Intelligent Text Summarization. pp51-59. AAAI Press.

Mani, I., House, D., Klein,G., Hirschman, L.,Obrst, L., Firmin, T., Chizanowski, M., and Sundheim, B. (1998). "The TIPSTER SUMMAC Text Summarization Evaluation." Technical Report MTR 98W0000138, MITRE Technical Report.

Minel, J.-L., Nugier, S. and Piat, G. (1997). "How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN." In Proc. of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp.25-30.

Mochizuki, H and Okumura, M. (1999). "Evaluation of Summarization Methods based on Information Retrieval Task." In Notes of SIGNL of the Information Processing Society of Japan, 99-NL-132, pp41-48. (In Japanese)

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company, Inc.

Ueda, Y., Oka, M., Koyama, T. and Miyauchi, T. (2000). "Toward the "At-a-glance" Summary: Phrase-representation Summarization Method." submitted to COLING2000.

Zechner, K. (1996). "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences." In Proc. of COLING-96, pp. 986-989.