Phrase-representation Summarization Method and Its Evaluation

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

Abstract

We have developed a summarization method that creates a summary suitable for the process of sifting information retrieval results. Unlike conventional methods that extract important sentences, this method constructs short phrases to reduce the burden of reading long sentences. We developed an improved task-based evaluation method and applied to prove the effectiveness of phrase-represented summary. In the task-based evaluation in TSC, phrase-represented summary provided the fastest judgement among systems that achieved almost the same accuracy. However, the experiment method has some problems, and we must try to design tasks closer to the real world IR in the future TSC.

Keywords: *Phrase-representation summarization, Phrase-represented summary, At-a-glance, Indicative, Task-based evaluation*

1 Introduction

Summaries are used to select relevant document from information retrieval results. The goal of summarization for such "indicative" use is to provide fast and accurate judgement.

Most automatic summarization systems adopt the "sentence extraction" method. It gives a score to every sentence on the basis of its characteristics, such as word frequency, the position in which it appears, etc. and selects sentences with high scores.

The sentences collected in such a way tend to be so long and complex that the reader must reconstruct the structure while reading them. Reading such sentences involves some annoyance.

Our aim is to reduce this burden by providing an "at-a-glance" summary. Phrase-representation summarization is a method to create the "at-a-glance" summary for the Japanese language.

In this paper, we present the concept, the algorithm, and evaluation of the efficacy of the summary produced by a prototype based on this method. The results and the problems of NTCIR Text Summarization Workshop (TSC) are also mentioned.

2 The concept

Examples of an "at-a-glance" summary are headlines of news articles. The headline provides information for judging whether a reader should read the article or not and, in this sense, it is really "indicative." The characteristics are brevity (short in length) and simplicity (less embedded sentences).

We use "phrases" to represent the simplicity[1] and set our goal to create phrase-represented summaries. They provide a reader with an outline of the document, avoiding reading stress by enumerating short phrases containing the important words and concepts composed from these words.

The method we adopted to achieve this goal is to construct such phrases from the relations between words.

The phrase-represented summary has the following characteristics.

(1) At-a-glance comprehension

Because each unit is short and simple, a 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.

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

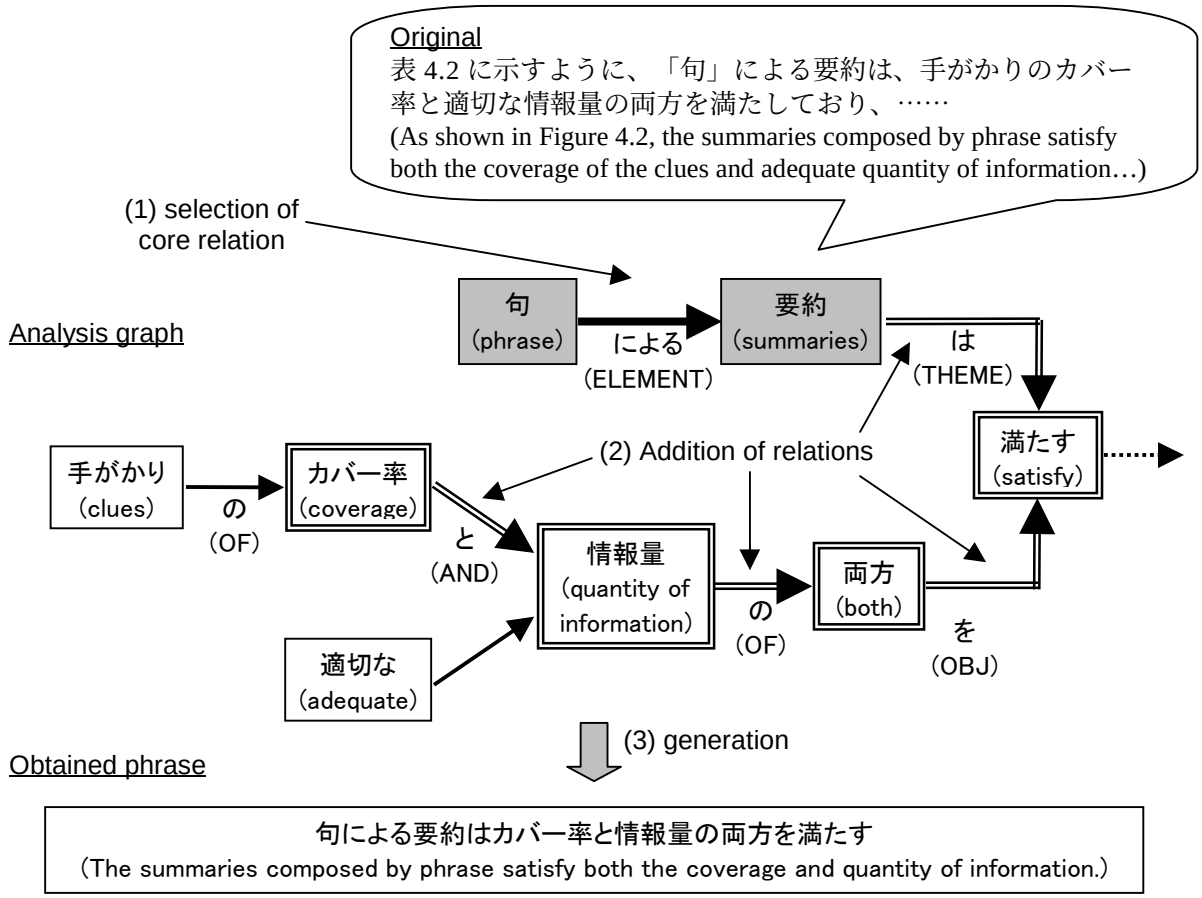

Figure 1: Outline of phrase construction.

3 Summarization method

3.1 Outline of phrase construction

Here we give the outline of the algorithm to construct a phrase using the example shown in Figure 1[2]. The method consists of the following three steps:

(1) Selecting a core relation from a text analyzed the relationships between words.
(2) Adding relations necessary for the unity of the phrase's meaning.
(3) Generating a surface phrase from the selected relations.

On the first step, a core relation of a phrase is selected from the given text. The sentences in the text are analyzed to produce directed acyclic graphs (DAGs) constructed from relation units. Each unit consists of two nodes (words) and an arc (relation between the words). Each node is not only a single word but also can be a word sequence (noun group). In Figure 1, the arc connecting the two shaded nodes represents the core relation.

The core relation alone carries insufficient information to convey the content of the original document. On the second step, additional relations are attached to specify the information the phrase supplies into further detail. In Figure 1, the nodes and arcs with double line are attached.

On the third step, a short phrase can be generated from the selected nodes and arcs in the graph.

Phrase-represented summary enumerates such short phrases to give the readers enough information to grasp the outline of a document. This algorithm is explained in the next section.

3.2 Outline of summary generation

A phrase-represented summary consists of several phrases, and there are two algorithms to generate a summary. One is based on important relations, another is based on important phrases.

Using the relation-based algorithm, an important relation is selected first and it is used as the core of a phrase. The required number of core relations are first selected. Using the phrase-based algorithm, all possible phrases are created first, and important

[2] In this paper, the examples are represented in Japanese, because TSC evaluates Japanese text summarization. However, translated words or particle functions in English are attached as much as possible. Applicability of the phrase-representation summarization to other languages is discussed in [1].

phrases are selected from them. Here, we introduce the flow of the phrase-based algorithm, because the summaries submitted to TSC are generated by this algorithm. The flow is shown in Figure 2.

The other method, the relation-based algorithm is introduced in [1].

3.3 Further description of each step

Here, we describe each step in Figure 2.

Relation Analysis

Syntactic analysis is applied to each sentence in the document to produce a DAG of the relations of words. We use a simple parser based on pattern matching [2], one of whose rules always judges each case dependent on its nearest verb. Some of the misanalysis will be hidden by "ambiguity packing" in the "additional relation attachment" step.

Core relation selection

A relation unit (two nodes and an arc connecting them) is selected as the core of a phrase.

Because different core relations can produce the same phrase by attaching additional relations (in the next step), it is not necessary to select every relation as a "core".

In the current implementation, most relations are selected as the core and omitting multiple phrases in the phrase selection step, because we are in the process of considering what is the good phrase by comparing the phrases generated by the system.

Additional relation attachment

The information that the core relation carries is usually insufficient. Additional relations are attached to make the information the phrase supplies more specific and to give the reader sufficient information to infer the content of the original document. The following relations are a part of the relations to be attached.

(1) Mandatory cases

Relations that correspond to mandatory cases are attached to verbs. Mandatory case is defined for each verb except for those that share the common mandatory case list, which includes "は" (THEME), "も" (ALSO) and null-marker.

Ex.) パンダ(panda)が(AGENT)上野動物園(Ueno Zoo)に(DATIVE)贈られた(was presented)

In the examples in this section, underlined words consist of the core relation, and double underline represents the attached words.

(2) Noun modified by a verb

In Japanese, the "verb – noun" structure represents an embedded sentence, and the noun usually fills some gap in the embedded sentence. If the verb in the core relation (noun – verb) consists of such a verb – noun relation, the modified noun is also assumed to carry important information, even if it does not fill the mandatory case (though the case is not analyzed in the current algorithm). Thus the verb – noun relation is attached to the core.

Ex.) 日本中(all over Japan)を(OBJ)席巻する(dominating)パンダブーム(panda boom)

(3) Ambiguity packing

The analysis graphs often contain errors because the pattern-based parser does not resolve ambiguities. For example, the structure

V 0-THAT[3] N1 "の"(OF) N2 (Ving N1's N2)

is ambiguous in Japanese (V can modify either N1 or N2 but the parser always analyzes N1 as modified). If the V – N1 relation is selected as the core, the N1 – N2 relation is always attached to the core to include the possible V – N2 relation.

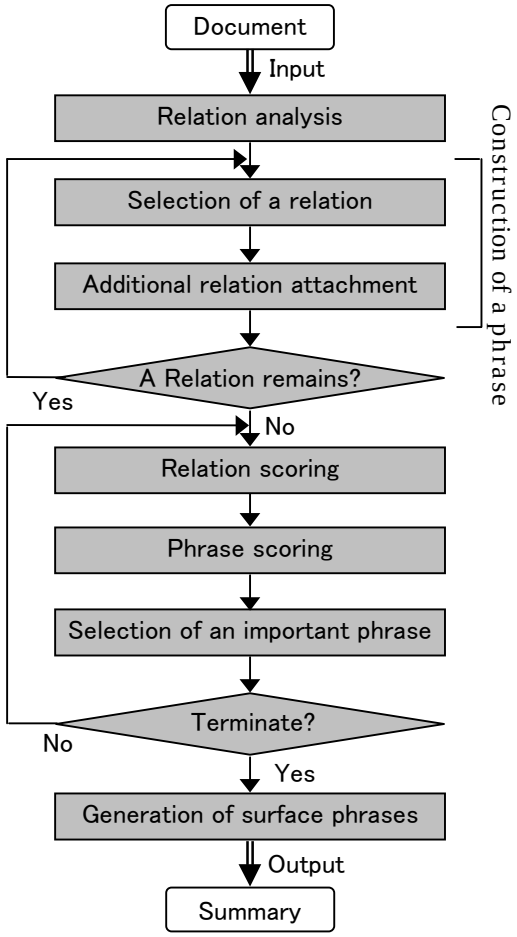

Figure2. Flow of the phrase-based algorithm.

[3] "0" shows that there are no particles or any other words connecting two words. Japanese does not require anything like relative pronouns.

Ex.) 椅子(chair)に(DATIVE)座った(sat)魯迅(Lu Xun)の(OF)像(statue)

(4) Modifiers of generic nouns

The concepts brought by generic nouns such as "もの" (thing), "こと" ("that" of that-clause), "場合" (case), "時代" (era) are not so specific that they usually accompany modifiers to be informative. Here such modifiers are attached to make them informative.

Ex.) 句(phrase)は(THEME)要約(summary)に(DATIVE)適した(suitable)単位(unit)である(is)

Relation scoring

An importance score is provided for each relation unit.

First, every word is scored by its importance. This score is calculated based on the tf*IDF value[4] [3]. In the summaries submitted in TSC, the score for the word in the headlines (Task A) or the queries (Task B) is weighted more than its tf*IDF value.

Then, the relation score is calculated as follows:

$$\text{Score} = S_{rel} * (S_1 + S_2)$$

Here, S_1 and S_2 are the scores of the two words connected by relations. The score of a word sequence is calculated by decreasing the sum of the scores of its constituent words according to the length of the word sequence.

S_{rel} is the importance factor of the relation. The relations that play central roles in the meaning, such as verb cases, are given high scores, and the surrounding relations, such as "AND" relations, are scored low.

Phrase scoring

An importance score is provided for each phrase. The basic algorithm is to sum up the scores of constituent relations in a phrase. To remove the influence of the phrase length, the total score is relaxed according to the number of constituent relations.

Phrase selection

The phrase with the highest score among all phrases is selected to compose a summary.

Terminative condition

Whether the summaries created so far are sufficient is judged. Currently, either the number of phrases or characters defines the condition.

Re-scoring of relations

If the condition is not satisfied, another phrase is selected. Before them, relation scores and phrase scores are re-calculated reducing the scores of the words used in the last phrase to avoid frequent use of the same words.

Score reduction is achieved by multiplying the predefined cut-down ratio R ($0 < R < 1$) by the scores of the words used.

Generation of surface phrases

The surface phrases are produced from selected relations to connect the surface strings of the nodes and their belonging words in the original order. In this step, the parts that are not selected in a summary phrase are replaced by "…". An example is as follows.

Ex.) …「句」による要約は、…カバー率と…情報量の両方を満たしており、…

4 Implementation

We developed a prototype of the summarization system based on this algorithm. The system is developed in Java and C++, and working on Windows 95/98/NT and Solaris 2.6.

The time consumed by summarization process is in proportion to the text length and it takes about 800 msec to generate a summary for an A4 sized document (2000 Japanese characters) using a Sun Enterprise 450[5].

5 Evaluation

5.1 Evaluation method for the phrase-represented summary

The aim of the phrase-representation summarization is to give fast and accurate judgement in selecting relevant documents from IR results. Thus, task-based evaluation on information retrieval [4] is adequate to evaluate the effectiveness of the summarization method.

We have conducted an evaluation experiment in 1999 [5], and participated in task-based evaluation (Task B) in TSC.

5.2 Our task-based evaluation in 1999

Task-based evaluation has recently drawn the attention in the summarization field, and some experiments on information retrieval were reported [6][7]. However, there is no standard evaluation method, and we consider that there are some shortcomings in the existing methods. Thus, we developed an improved evaluation method and carried out a relatively large-scale experiment.

[4] IDF is calculated from Mainichi Newspapers in 1995 (CD-ROM).

[5] Java , Solaris and Sun are the trademarks of Sun Microsystems. Windows is the trademarks of Microsoft and Intel, respectively.

Here we briefly introduce our task-based evaluation.

Evaluation method

We compared four types of summary: (a) leading fixed-length characters, (b) tf*IDF-based sentence extraction summaries [8], (c) phrase-represented summaries and (d) tf*IDF-based keywords. The evaluation criteria were the accuracy of subjects' relevance assessment and the time to assess.

The characteristics of our evaluation method are as follows:

(1) The summary length is regulated to sixty to eighty characters.
(2) To reduce the diversity of the assessment, we made the assumed situation realistic and specify it into details including the purpose of the search.
(3) We assigned ten subjects per summary sample to reduce the influence of each person's assessment.
(4) For subjects to assess the relevance, we introduced four relevance levels (from higher to lower: L3, L2, L1 and L0, which is judged to be irrelevant).

Experiment results

The relationship of the accuracy and the time is shown in Figure 3. To represent the accuracy, f-measures are used in Figure 3.

The result proves that the phrase-represented summaries serve accurate judgement in a relatively short time.

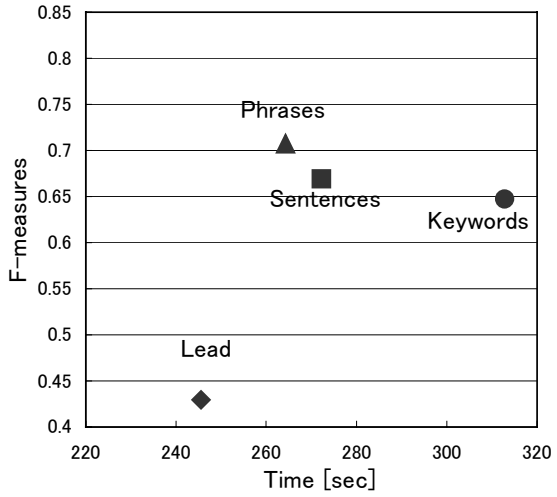

Figure 3. The relationship between f-measures and time.

5.3 Task-based evaluation in TSC

The relationship between f-measures and time in TSC-task B is shown in Figure 4. We use the result of "answer level B", because whether a topic is the subject of a document or not is independent of whether a user needs the document or not.

We can not simply compare all summaries in Figure 4 because of the diversity of the summary length.

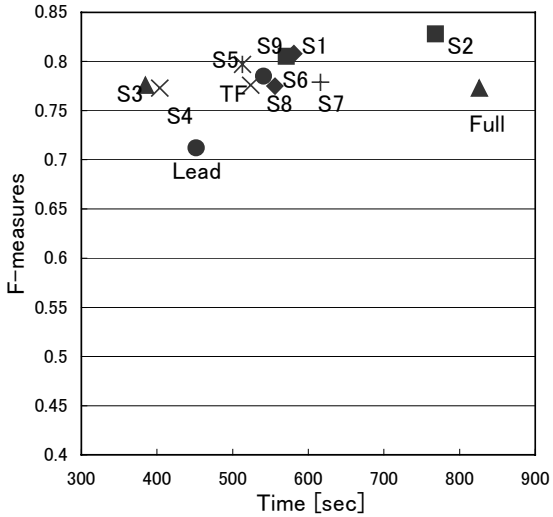

Figure 4. The relationship between f-measures and time.

The summary length is not regulated in TSC-task B. We submitted two different lengths of phrase-represented summaries: within 100 characters (System 3) and 150 characters (System 4) to fit it to the purpose of the task (IR sifting).

However, other participants submitted much longer summaries. The summary length of each system is compared in Figure 5. The broken line represents the average character numbers of all summaries except full text. Figure 5 shows that summaries by System 3 and System 4 are much shorter than the average length.

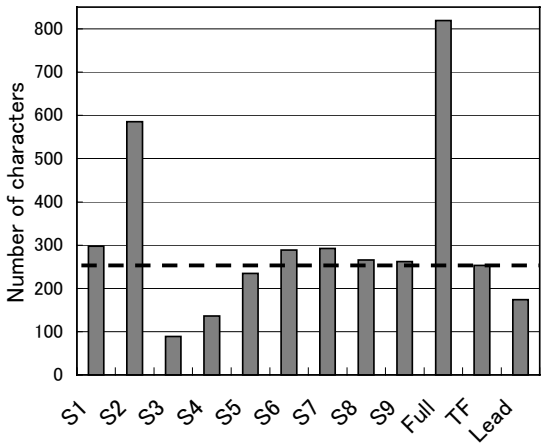

Figure 5. Summary length.

Generally speaking, longer text has larger information, and requires a longer time to read. The relationship between the summary length and the f-measures is shown in Figure 6, and the relationship between the summary length and time is shown in Figure 7. These figures support the above assumption.

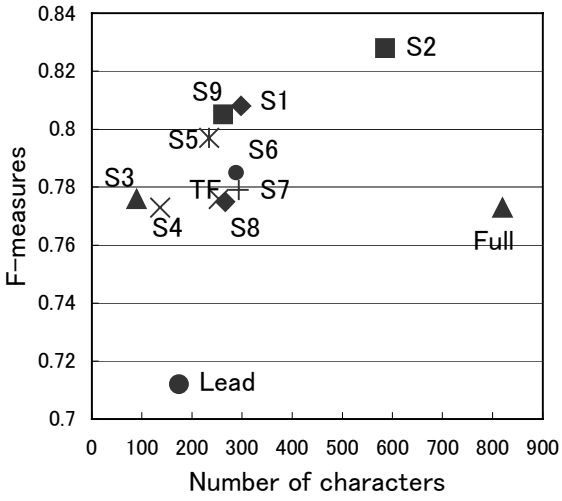

Figure 6. The relationship between summary length and f-measures.

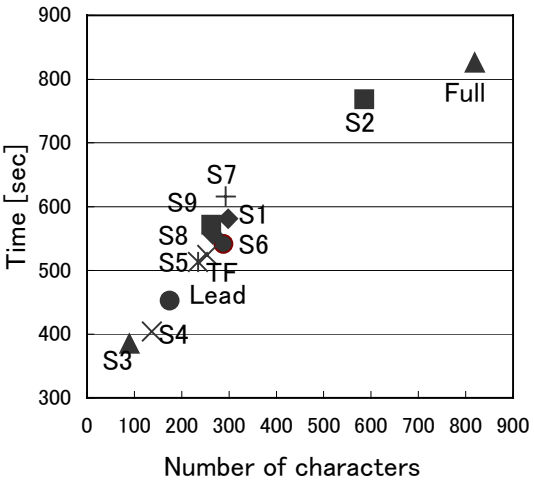

Figure 7. The relationship between summary length and time.

However, at least we can get the following information from these results.

(1) The phrase-represented summaries are much faster than the systems that get almost the same f-measures.

(2) The phrase-represented summaries are about 1/8 length of full text and get almost the same f-measures.

5.4 Toward future task-based evaluation

The comparison between our experiment and TSC-task B is shown in Table 1. From the experiences of designing an experiment method and participating in TSC-task B, we would like to make some suggestions for the future TSC.

If the task-based evaluation on information retrieval will be carried out also in the future, we must try to design tasks closer to the real situation. Here we discuss them from three aspects: summary, task and document.

Summaries to be read

In the standard WWW search engines, about 100 characters are used as the summary and the length must be adequate as a replacement of the document itself. Not only regulate the length, but also selecting an appropriate length must be needed.

Task

In the assumed IR on TSC, each topic is given as a simple concept (e.g. “performances at Kabuki-za”) and thus the subjects must judge the relevance whether the concept is included in the document or not. Such IR is largely different from the real world IR. We usually have some purpose and retrieval is just a mean to achieve the goal, and judge whether the document itself contains information to fulfill the purpose by reading the summary. The criteria depends on the situations, for example, “the searcher must make an survey of reviews of the performances at Kabuki-za,” or “the searcher wants to reserve tickets of coming performance at Kabuki-za”.

The f-measures of full texts support our view. In the task-based evaluation, it is assumed that subjects can judge relevance correctly by full texts, and the system can be evaluated by how good they can achieve the task using summaries instead of full texts. In the experiment result shown in Figure 4, the f-measures of the full texts are relatively low. It represents that many subjects cannot achieve their task even if they are given the full information.

Documents to be retrieved

The document source in TSC is newspaper articles, but usually we search various types of documents at once. In the WWW for example, there are various styles or forms of documents, such as essays, articles, papers, advertisements, mail archives or diaries. Though there are many problems (e.g. copyright) to obtain and use various types of document, we should try to make a test collection of inconsistent documents and use them in evaluations like TSC.

Table 1. Comparison of the experiment method.

	Our experiment in 1999	TSC-task B
Document source	WWW	Newspaper
IR question	Newly created with detailed situations	Selected topics from IREX
Number of questions	3	12
Number of documents per question	10	50
Summary length	60-80 characters	Not regurated
Subjects	40 persons who usually use WWW search	36 students
Number of subjects per summary sample	10	3
Relevance levels in subject's assessment	4 levels (L3: the answer to the question is found in a summary. L2: a clue to the answer is found in a summary. L1: clues are not found in a summary, but the document may be relevant. L0: irrelevant.)	2 levels (relevant or irrelevant)
Relevance levels of document	2 levels (relevant or irrelevant)	3 levels (Level A: given topic is subject of an article. Level B: given topic is not subject but written in an article. Level C: irrelevant.)

5.5 The evaluation of the phrase style

Though TSC-task A is for informative summaries, we have participated also in task A-2. In task A-2, the system results are compared with human-prepared summaries by two ways: one is content-based evaluation (task A-2-1) and another is subjective evaluation (task A-2-2).

The result of task A-2-2 shows that the acceptability of the phrase-represented summaries is not so highly evaluated by human. We suppose there are two reasons. One is that the form of phrase-represented summary extremely differs from that of which general human reminds as "summary". The other is that phrase-represented summary occasionally contains a phrase that is grammatically incorrect, whereas sentence extraction does not contains such errors, because each constituent is the original sentence itself.

One of the purposes that we have participated in task A-2 is to evaluate the phrase style. As described in Section 3.3, the parts not selected in a summary are replaced by "…". This style has both advantages and shortcomings. One of the advantages of using "…" is to allow reader to imagine the meaning even if the phrase is not perfect. One of the shortcomings is that "…" disturbs smooth reading and understanding.

We submitted two types of summary: (1) All parts not selected in a phrase are replaced by "…" (System 3). (2) Only the end of a phrase is replaced by "…" (System 4).

The result is that the (2) type summaries were better in both readability and similarity to the human summaries. We would like to take this into consideration in designing future systems.

5.6 The content-based evaluation

The result of task A-2-1, the similarity between the system results and the human-prepared summaries is shown in Figure 8. System 3 is omitted here, because System 3 and System 4 have almost the same contents.

In Figure 8, the phrase-represented summaries occupy relatively high positions among all systems.

One of the factors that raise the similarity comes from the summary creating method. The phrase-represented summaries are constructed so that they include the important concepts of the document, and this goal resembles that of the human free summaries. And each phrase is a part of the original sentence as the sentences of the important-part summaries. On the contrary, the sentences which most summarization systems output necessarily include not so important parts of the sentences to reduce the similarity score.

If we tune up our system for newspaper summaries or human-prepared summaries used in TSC dryrun,

the similarity score might be much higher. One reason why we avoided tune-up is that our target is not limited to one specific document type and tuning toward it is inappropriate. Another reason is that we consider there is no single "correct" summary and the human-prepared summary provided for TSC is just one sample and inappropriate as a target of tune-up.

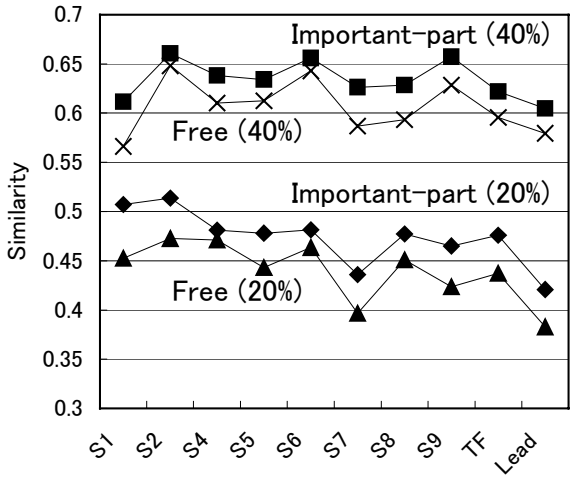

Figure 8. The Similarity between system results and human summaries.

6 Conclusion

We introduced the phrase-representation summarization method. The task-based evaluation in TSC shows that the summaries are effective for fast sifting, however there seems to be some problems in the experimental method. Toward the future TSC, we should try to design tasks closer to the real world IR.

References

[1] Ueda, Y., Oka, M., Koyama, T. and Miyauchi, T. Toward the "At-a-glance" Summary: Phrase-representation Summarization Method. *Proceedings of COLING-2000*: 878-884, 2000.

[2] Miyauchi, T., Oka, M. and Ueda, Y. Key-relation technology for text retrieval. *Proceedings of SDAIR'95*: 469-483, 1995.

[3] Salton, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[4] Hand, T. F. A Proposal for Task-based Evaluation of Text Summarization Systems. *Proceedings of ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*: 31-38, 1997.

[5] Oka, M. and Ueda, Y. Evaluation of Phrase-representation Summarization based on Information Retrieval Task. *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization*: 59 - 68, 2000.

[6] Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. Summarization Evaluation Methods: Experiments and Analysis. *Intelligent Text Summarization*: 51-59, AAAI Press, 1998.

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

[8] Zechner, K. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. *Proceedings of. COLING-96*: 986-989, 1996.