A Multiple-Document Summarization System introducing User Interaction for reflecting User's Summarization need

Hiroyuki SAKAI   Shigeru MASUYAMA
Department of Knowledge-based Information Engineering
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
sakai@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp

Abstract

We propose a multiple-document summarization system with user interaction that summarizes more than one document to a document. Our system extracts keywords from sets of documents to be summarized and shows k best keywords with respect to scoring by our system to a user on the screen. From the shown keywords, the user selects those reflecting the user's summarization need. Our system controls produced summary by using these selected keywords. For evaluation of our method, we participated in TSC3 of NTCIR4 workshop by letting our system select all k keywords supposed to display for the user. Our participated system exhibited the best performance in content evaluation among systems not using sets of questions. Moreover, we evaluated effectiveness of user interaction in our system. With user interaction, our system attained both higher coverage and precision than that without user interaction.

Keywords: *multiple-document summarization, user interaction, keywords selection by a user, unimportant adnominal verb phrases deletion.*

1 Introduction

Recent rapid progress of computer and communication technologies enabled us to access enormous amount of machine-readable information easily. However, this has caused so called the information overload problem. In order to solve this problem, automatic summarization methods have been studied (see e.g., [13]). In particular, the necessity for a multiple-document summarization, which summarizes more than one document and produces a summary, has been increasing and has been intensively studied recently (see e.g., [10]).

In this paper, we define a multiple-document summarization as a technique for producing a summary from a relevant documents set. Such a documents set may be very large and may contain a number of topics. It is preferable that a summary produced by a multiple-document summarization system from the documents set covers all the topics. However, it is difficult to produce a summary that covers all topics with a small number of characters. For example, a documents set relevant to "releasing AIBO" contains some topics, e.g., what is AIBO?, how to sell AIBO?, etc. Moreover, important sentences extracted by a person considerably differ with the person [15]. We consider that the reason is as follows: "Summarization need", i.e., topics a person wants to read, differs with the person. Hence, we propose a multiple-document summarization system with user interaction for coping appropriately with user's summarization need.

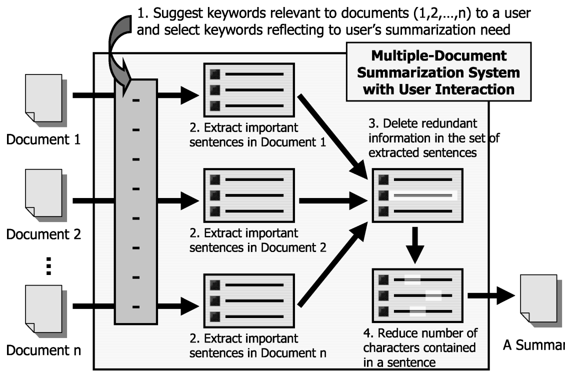

Figure 1. Outline of our multiple summarization system

Our system extracts keywords from a documents set to be summarized and shows k best keywords with respect to scoring by our system to a user on the screen. From the shown keywords, the user selects those reflecting user's summarization need. Our system controls a produced summary by using the keywords selected by the user. The outline of our multiple-document summarization system is illustrated in figure 1.

As a related work, Mani et al.[11] proposed a method for a user-focused summary. In their paper, a user selects 10 documents as training corpus to be used for producing a summary in order to produce a user-focused summary. In contrast, our multiple-document summarization system attempts to produce a summary reflecting a user's summarization need by choosing topics contained in a documents set to be summarized using keywords selected by the user. Since even the same user may have a different summarization need at different occasion, the learning method using training corpus for producing a summary may not be always applied. Hence, our system produces a summary reflecting user's summarization need by asking a user to select keywords reflecting the user's summarization need among keywords extracted from a documents set to be summarized. Moreover, a user's load caused by selecting keywords is much lighter than that by selecting documents.

We participated in TSC3 (Text Summarization Challenge - 3) of NTCIR4 workshop [1] and attained the best performance in content evaluation among systems not using sets of questions. Note that our system participated in TSC3 is an automatic summarization system without user interaction by letting our system select all k keywords supposed to display for the user. Moreover, we evaluated effectiveness of user interaction and that with user interaction attained both higher coverage and precision than that without user interaction.

2 Feature of our multiple-document summarization system

Our multiple-document summarization system proposed in this paper is different from previously proposed multiple-document summarization methods (see, e.g.,[3], [12], [5], [1], [9], [14], [7], [20]) in that:

1. Our system can produce a summary coping appropriately with each user's summarization need by asking a user select keywords reflecting the user's summarization need.

2. The keywords are extracted automatically from a documents set to be summarized by calculating a score to each noun contained in the documents set. The formula to calculate a score for a noun is customized for extracting important keywords in multiple-document summarization. The formula consists of not only frequency of nouns and document frequency used in $tf \cdot idf$[2] but also distribution of nouns in the documents set as well as location of nouns in documents or the documents set. The reason why such factors are used will be explained in the next section.

3. Our system deletes redundant adnominal verb phrases in sentences to reduce the number of characters in a sentence. The deletable adnominal verb phrases are decided by a statistical method using entropy based on a probability that verbs modify a noun, etc. Our previous method [17] improved for multiple-document summarization so that more deletable adnominal verb phrases are recognized, is used.

[1] http://www.lr.pi.titech.ac.jp/tsc/index-en.html

3 The method to extract relevant keywords

A relevant documents set S to be summarized may be regarded as a documents set obtained by a hypothetical query from the entire documents set N to be considered. In the case of TSC3, N is the set of newspaper articles, Mainichi and Yomiuri newspapers published in 1998 and 1999.

We explain a method to extract keywords relevant to such a hypothetical query from documents set S. Here, we define such keywords as relevant keywords t_i, $i = 1, 2, \ldots, k$. We assign scores to nouns contained in documents set S and nouns having a large score are extracted as relevant keywords. A large score is assigned if a noun fulfills the following four conditions.

1. The noun that appears frequently in the documents set S to be summarized.

2. The noun that appears uniformly in each document $d \in S$.

3. The noun that appears in the beginning of a document (i.e. the 1st sentence) and in the beginning of the documents set in the order of the time (i.e., the 1st document).

4. The noun that does not appear frequently in entire documents set N.

Our method for extracting relevant keywords consists of the following two steps.

Step 1: Calculate score $W(t_i, S)$ to noun t_i contained in documents set S.

Step 2: Extract k largest nouns with respect to score $W(t_i, S)$ as relevant keywords. □

The score $W(t_i, S)$ is calculated by the following formula 1.

$$\begin{aligned} W(t_i, S) &= (0.5 + \frac{Tf(t_i, S)}{max_{i=1,\ldots,n} Tf(t_i, S)}) \\ &\quad \times (0.5 + \frac{En(t_i, S)}{max_{i=1,\ldots,n} En(t_i, S)}) \\ &\quad \times max_{d \in S} \frac{1 + nl(d) - nlf(t_i, d)}{nl(d)} \\ &\quad \times max_{d \in S} \frac{1 + |S| - rt(t_i, d)}{|S|} \\ &\quad \times idf(n, N) \qquad (1) \end{aligned}$$

where,

$Tf(t_i, S)$: frequency of noun t_i contained in documents set S. This is calculated by the following formula 2.

$$Tf(t_i, S) = \sum_{d \in S} tf(t_i, d) \quad (2)$$

where, $tf(t_i, d)$ is a frequency of noun t_i in document d.

$En(t_i, S)$: entropy based on the probability that noun t_i appears in document $d \in S$. This is calculated by formula 3 to be introduced later.

$nl(d)$: the number of sentences in document $d \in S$.

$nlf(t_i, d)$: the line number of a sentence containing noun t_i for the first time in document $d \in S$.

$rt(t_i, d)$: the number of documents from 1st document to document d containing noun t_i for the first time in documents set S in the order of the time.

$idf(n, N)$**:** idf[2] value assigned to noun n in entire documents set N.

$En(t_i, S)$ is an entropy based on a probability that noun t_i appears in document $d \in S$. For example, $En(t_i, S)$ assigned to noun t_i contained only in one document $d \in S$ is 0. Though such noun t_i may be an important noun for document d, it may be an irrelevant noun for documents set S. Hence, noun t_i with small entropy value should not be extracted as a relevant keyword. However, a noun that appears uniformly in each document contained in documents set S has a large entropy value. $En(t_i, S)$ is calculated by the following formula 3.

$$En(t_i, S) = -\sum_{d \in S} P(t_i, d) \log_2(P(t_i, d)) \quad (3)$$

$$where,\ P(t_i, d) = \frac{tf(t_i, d)}{Tf(t_i, S)} \quad (4)$$

The 3rd term in formula 1 is to assign a large value to a noun appearing in the beginning of a document. The 4th term in formula 1 is to assign a large value to a noun appearing in the beginning of a documents set in the order of time. The reason why these members are included is that the 1st sentence in the 1st document frequently contains important information (see, e.g.,[4],[16]).

4 The method to extract important sentences

The method to extract important sentences measures similarity between a sentence and the set of keywords selected by a user, and extracts sentences having large similarity with the set of selected keywords as important sentences. The similarity is calculated as cosine metric between a vector of a sentence and a vector of the set of relevant keywords. If the same noun as selected relevant keywords is contained frequently in a sentence, the cosine metric assigned to the sentence has a large value. The method to extract important sentences is summarized as follows: Here, we define k relevant keywords shown to a user as keywords set K and define relevant keywords selected by a user as keywords set U (we define the number of keywords shown to a user to be k).

Step 1: Recalculate score $W(t_i, S)$ assigned to relevant keywords t_i's by the following formula 5.

$$W'(t_i, S) = \begin{cases} (1 + 0.5k) W(t_i, S), & t_i \in U \\ W(t_i, S), & \text{otherwise} \end{cases} \quad (5)$$

Step 2: Generate relevant keyword vector $\mathbf{V_K}$ consisting of score $W'(t_i, S)$ $(i = 1, 2, \ldots, k)$ assigned to each relevant keyword ($t_i \in K$).

$$\mathbf{V_K} = (W'(t_1, S),\ W'(t_2, S), \ldots,\ W'(t_k, S))$$

Step 3: Generate sentence vector $\mathbf{V_s}$ consisting of score $W'(t_j, S)$ $(j = 1, 2, \ldots, m)$ assigned to each noun contained in sentence s ($t_j \in s$. Note that, in this paper, we use this notation by regarding sentence s as a set of words).

$$\mathbf{V_s} = (W'(t_1, S),\ W'(t_2, S), \ldots,\ W'(t_m, S))$$

Step 4: Calculate a cosine metric between vector $\mathbf{V_K}$ and vector $\mathbf{V_s}$ as similarity $sim(s, K)$ by the following formula 6.

$$sim(s, K) = \frac{\mathbf{V_K} \cdot \mathbf{V_s}}{|\mathbf{V_K}||\mathbf{V_s}|} \quad (6)$$

Step 5: Extract m largest sentences with respect to similarity $sim(s, K)$ and output these m sentences in the order of the time. □

In documents set S, in the order of the time, the 1st sentence contained in the 1st document containing extracted important sentences is always extracted as an important sentence in order to improve the readability.

5 The method to delete redundant information

In the multiple-document summarization, it is necessary to measure the degree of similarity of contents in extracted sentences (or documents) and to delete redundant information. This is because, the documents including the same contents may exist in a documents set to be summarized. Our multiple-document summarization system identifies similar sentences in extracted important sentences set and similar documents in the documents set, and deletes redundant information contained therein.

First, redundant information contained in the sentences set is deleted as follows.

Step 1: Measure the difference $d(s_1, s_2)$ between cosine metric $sim(s_1, K)$ assigned to sentence s_1 and $sim(s_2, K)$ assigned to sentence s_2.

$$d(s_1, s_2) = |sim(s_1, K) - sim(s_2, K)| \quad (7)$$

Step 2: If $d(s_1, s_2)$ has a value smaller than a threshold value, delete sentence s_i $(i = 1 \; or \; 2)$ having a smaller cosine metric $sim(s_i, K)$. □

We determined the threshold value to be 0.0001 in Step 2. This is a sufficiently small value to regard contents of s_1 identical to contents of s_2.

Next, redundant information contained in the documents set is deleted as follows. Here, we define a set of important sentences contained in document d_i as sd_i. The method is as follows.

Step 1: Generate vector $\mathbf{V_{sd_1}}$, consisting of score $W'(t_i, S)$ $(i = 1, 2, \ldots, n)$ assigned to nouns contained in sd_1.

$$\mathbf{V_{sd_1}} = (W'(t_1, S), W'(t_2, S), \ldots, W'(t_n, S))$$

Step 2: Generate vector $\mathbf{V_{sd_2}}$, consisting of score $W'(t_j, S)$ $(j = 1, 2, \ldots, m)$ assigned to nouns contained in sd_2.

$$\mathbf{V_{sd_2}} = (W'(t_1, S), W'(t_2, S), \ldots, W'(t_m, S))$$

Step 3: Calculate a cosine metric between vector $\mathbf{V_{sd_1}}$ and vector $\mathbf{V_{sd_2}}$ as similarity $sim(sd_1, sd_2)$.

$$sim(sd_1, sd_2) = \frac{\mathbf{V_{sd_1}} \cdot \mathbf{V_{sd_2}}}{|\mathbf{V_{sd_1}}||\mathbf{V_{sd_2}}|} \quad (8)$$

Step 4: If $sim(sd_1, sd_2)$ has a value larger than a threshold value, delete document d_i $(i = 1 \; or \; 2)$ having a smaller score $W(sd_i)$ calculated by formula 9 ($sd_i \subset d_i$). □

$$W(sd_i) = \sum_{s \in sd_i} sim(s, K) \quad (9)$$

Documents d_1 and d_2 are newspaper articles issued on the same day. We determined the threshold value to be 0.85 in Step 4 by trial and error using sample data provided by the organizer of TSC3. Note that this sample data was not used in the formal run as a documents set to be summarized. This is a sufficiently large value to regard contents of sd_1 identical to contents of sd_2.

Note that if d_i is deleted, sentences contained in document d_i are not extracted and important sentences extracted by our system are changed. Hence, our system executes this algorithm to delete documents and the algorithm to extract important sentences iteratively until no document is deleted by this algorithm.

6 The method to reduce the number of characters in a sentence

Our system deletes redundant adnominal verb phrases in sentences to reduce the number of characters in a sentence. We define adnominal verb phrases as phrases that modify a noun and include a verb modifying the noun. For example, in the case of "*SONY ga kaihatsu shita aibo*(ソニーが開発したアイボ: the AIBO developed by SONY", "*SONY ga kaihatsu shita*(ソニーが開発した: developed by SONY)" is an adnominal verb phrase, which modifies noun "*aibo*(アイボ: AIBO)". Here, the adnominal verb phrase "*SONY ga kaihatsu shita*(ソニーが開発した: developed by SONY)" may be deleted if a user has known that AIBO had been developed by SONY. We define an adnominal verb phrase modifying a noun n as $VP(n)$. Redundant adnominal verb phrases are deleted by an improved method of [17] proposed by us in order to apply to multiple documents summarization. We introduce $CV(n, s)$ in formula 13 and $CT(c, s)$ in formula 16 in order to recognize more deletable adnominal verb phrases than our previous method. The method is as follows.

Step 1: Calculate score $endf(n)$ to assign to noun n modified by adnominal verb phrase $VP(n)$ by formula 10.

Step 2: Calculate score $W(VP(n), s)$ for adnominal verb phrase $VP(n)$ by formula 13.

Step 3: Delete adnominal verb phrase $VP(n)$ if the score $endf(n)$ has a value smaller than threshold value $\theta(endf(n))$ and the score $W(VP(n), s)$ has a value smaller than threshold value $\theta(W(VP(n), s))$. □

We decided threshold value $\theta(endf(n))$ as 0.7 and threshold value $\theta(W(VP(n), S))$ as 8.7 in Step 3. These threshold values are decided by preliminary experiments with training corpus not to be summarized in the experiments. Score $endf(n)$ expresses the modifier necessity of noun n and is calculated by the following formula 10.

$$endf(n) = \frac{1 + H(n)}{idf(n, N)} \quad (10)$$

Here, $H(n)$ is an entropy based on a probability that verbs modify noun n. It reflects "frequency of modification of noun n by adnominal verb phrases", "variety of adnominal verb phrases modifying noun n". $H(n)$ is calculated by the following formula 11:

$$H(n) = -\sum_{v \in V(n)} P(v, n) \log_2(P(v, n)) \quad (11)$$

$$P(v, n) = \frac{f(v, n)}{\sum_{v \in V(n)} f(v, n)} \quad (12)$$

where,

$V(n)$**:** set of verbs contained in adnominal verb phrases modifying noun n in entire documents set N,

$P(v, n)$**:** probability that verb $v \in V(n)$ modifies noun n.

$f(v, n)$**:** frequency of verb v modifying noun n in entire documents set N.

Next, $W(VP(n), s)$ is calculated by the following formula 13.

$$W(VP(n), s) = \frac{NM(n)IM(VP(n), s)}{0.5 + 0.5CV(n, s)} \tag{13}$$

$$NM(n) = 0.5 + \frac{endf(n)}{J(n)} \tag{14}$$

where,

$IM(VP(n), s)$**:** a factor to reflect rating of context in adnominal verb phrase $VP(n)$ contained in sentence s.

$CV(n, s)$**:** the number of occurrences of noun n modified by adnominal verb phrases from the 1st sentence in the 1st document to sentence s in document $d \in S$ in documents set S in the order of the time.

$J(n)$**:** the number of common nouns contained in noun n if noun n is a compound noun.

The $IM(VP(n), s)$ is calculated by the following formula 15.

$$IM(VP(n), s) = 0.5 + R \sum_{c \in VP(n)} I(c, s) \tag{15}$$

$$I(c, s) = \frac{W(c, S)}{0.5 + 0.5CT(c, s)} \tag{16}$$

where,

R**:** the number of segments composing adnominal verb phrase $VP(n)$,

$W(c, S)$**:** the score calculated by formula 1 to noun c contained in adnominal verb phrase $VP(n)$.

$CT(c, s)$**:** the number of occurrences of noun c contained in adnominal verb phrases from the 1st sentence in the 1st document to sentence s in document $d \in S$ in documents set S in the order of the time.

7 Implementation

We implemented our method and developed a multiple summarization system. We used JUMAN [2] as a morphological analyzer, and KNP[3] as a parser. Figure 2 exemplifies a screen shot of our multiple-document summarization system and figure 3 exemplifies a summary produced by our system. The documents set to be summarized contains 9 documents relevant to "releasing AIBO" and the summary consists of less than 236 characters. Moreover, figure 4 exemplifies a summary when a user selects keywords relevant to the movement and performance of AIBO (e.g., "人工知能 (artificial intelligence)") and deletes keywords relevant to the way to sell (e.g., "予約 (Reservation)"). Comparing Figure 3 with Figure 4, we can make sure that summaries have been changed by keywords selected by a user.

Figure 2. Screen shot of our multiple-document summarization system

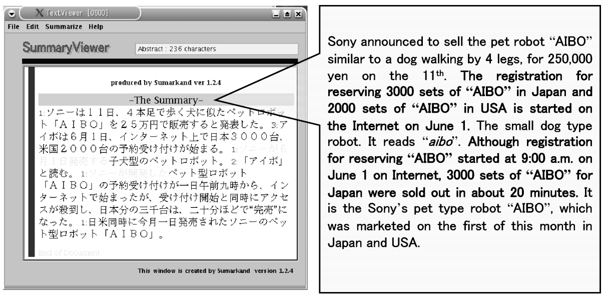

Figure 3. A summary produced by our system

8 Evaluations of our system in TSC3

We participate in TSC3 (Text Summarization Challenge - 3) of NTCIR4 workshop for evaluation of information access techniques. The purpose of TSC3 is to evaluate performance of automatic multiple-document summarization that summarizes newspaper articles from two sources (Mainichi and Yomiuri newspapers published between 1998 and 1999). The

[2]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[3]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

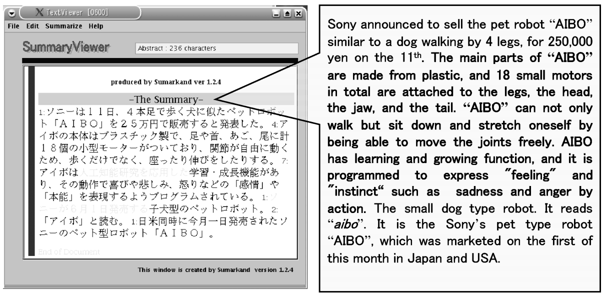

Figure 4. A summary produced by changing relevant keywords

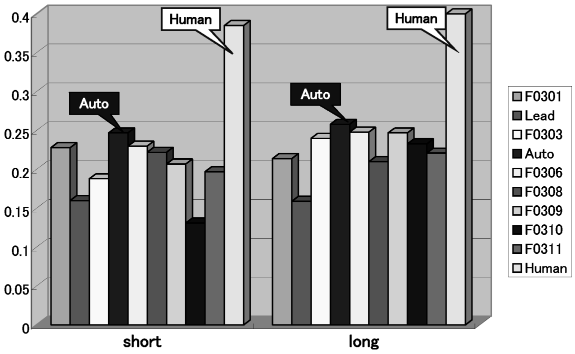

Figure 5. Results on Content evaluation

tasks of TSC3 are "abstraction" and "extraction". The evaluation methods of "abstraction" are "content evaluation", "pseudo question-answering" and "readability evaluation" (see task overview of TSC3[6]).

For participating in TSC3, we denote the following execution of our system by "Auto" for realizing an automatic multiple-document summarization system without user interaction and our system without user interaction participated in TSC3 by "Auto".

Auto: the execution of our system where 12 best keywords with respect to scoring by the system are selected (i.e., our system shows 12 keywords on the screen and all these keywords are selected).

The number of keywords selected by our system is determined by trial and error using sample data provided by the organizer of TSC3.

8.1 Evaluation results on abstraction

The result of content evaluation is shown in figure 5. The evaluation result of pseudo question-answering[6] is shown in Table 1. Here, "Auto" denotes our system that participated in TSC3. "Lead" is the lead method, a baseline method. In TSC3, we are given the sets of questions about important information of the document sets by the organizer of TSC3. Note that these sets of questions are produced manually from summaries made by human as correct data. (For example: How much is AIBO ? etc.) Here, we

Table 1. Results for Pseudo Question-Answering

System-ID	Short exact	Short edit	Long exact	Long edit
F0301	0.394	0.677	0.399	0.706
F0303	0.257	0.556	0.266	0.602
Auto	0.367	0.653	0.356	0.677
F0306	0.342	0.614	0.327	0.630
F0307	0.439	0.710	0.442	0.751
F0308	0.321	0.601	0.313	0.611
F0309	0.390	0.684	0.356	0.633
F0310	0.133	0.427	0.201	0.549
F0311	0.304	0.579	0.308	0.628
Lead	0.300	0.589	0.275	0.602
Human	0.461	0.716	0.426	0.721

exclude evaluation results of a system that uses the sets of questions for producing summaries of multiple documents. The reason is as follows. As mentioned above, the sets of questions are produced from summaries made by human as correct data. Hence, we consider that using the sets of questions as machine-readable information for producing summaries is not realistic. Moreover, we consider that comparing systems using the sets of questions with systems not using them by ranking is unfair.

By the result shown in figure 5, our system that implemented "Auto" has attained the best performance among the systems not using the sets of questions. However, since we did not introduce any method for improving readability, our system had not attained a good performance in readability evaluation. Hence, introducing some methods may improve readability (e.g., a method to delete unnecessary adnominal phrases[19]).

8.2 Evaluation results on extraction

The purpose of this subtask is to evaluate extracting important sentences and deleting redundant information from the important sentences set. A scoring tool and human-produced extracts are provided and each system is evaluated by coverage and precision scores where redundancy is taken into account by this scoring tool. The coverage and precision are shown in figures 6 and 7 (the coverage and precision are described in [6]).

8.3 Evaluation of user interaction

Our system is essentially a multiple-document summarization system with user interaction. Hence, we evaluate effectiveness of user interaction of our system in this subsection by using the set of questions

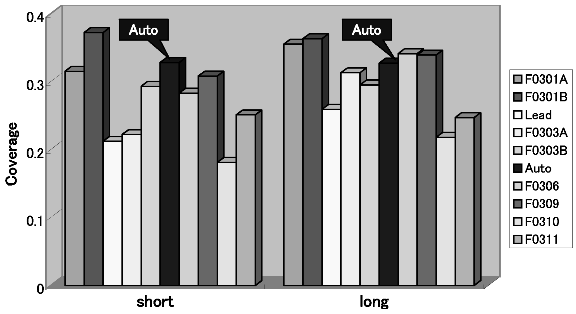

Figure 6. Coverage in extraction

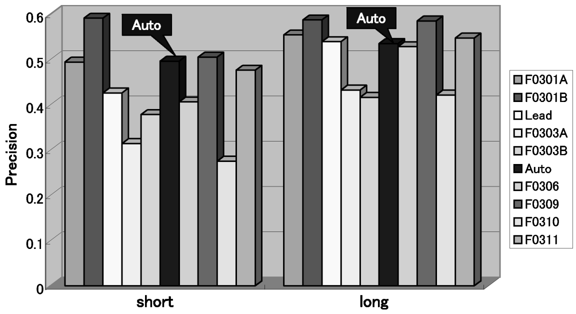

Figure 7. Precision in extraction

provided for pseudo question-answering in TSC3. For evaluating it, we consider the following execution of our system:

Interaction: Execution of our system where relevant keywords contained in the set of questions provided by TSC3 are selected, and relevant keywords not contained in the set of questions are deleted.

The "Interaction" simulates user interaction on our system. (i.e., we regard the set of questions as user's summarization need. Since the set of questions produced from summaries by human (i.e., a user), we will be able to regard the questions as user's summarization need.) Note that the "Interaction" is an execution to be compared with the "Auto" and the system that implemented "Interaction" did not participate in TSC3. The coverage and precision of "Interaction" is shown in Figure 8. Moreover, the coverage and precision of "Auto" and "Lead" are shown for comparison. Here, the coverage and precision are obtained by using the scoring tool provided for "extraction" by the organizer of TSC3.

9 Discussion

From the result shown in figure 5, our system participating in TSC3 as "Auto" had attained the best performance among the systems not using the sets of questions in "abstraction". Hence, we conclude that our system is effective to attain a good performance in multiple-document summarization. Moreover, from

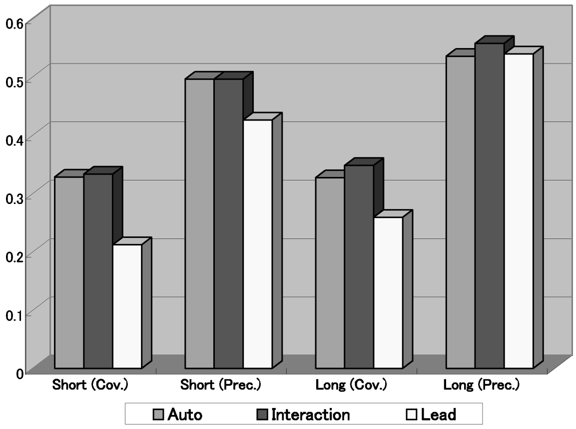

Figure 8. Evaluation of user interaction

the result shown in figures 6 and 7, our system had attained a good performance in "short" of "extraction" among the systems not using the sets of questions. In TSC3, the documents set to be summarized contains documents from two different sources, i.e., Mainichi and Yomiuri. Thus redundancy reduction is more important than the case where the documents set to be summarized consists of documents from a single source. Hence we consider that the reason why our system exhibited a good performance is that it has a good ability in reduction of redundant information. However, performance by our system in "long" of "extraction" is not so good. Hence, we consider that the capability to extract essential important sentences is satisfiable. However, the capability of our system to extract a wide range of topics is not high.

From the result shown in figure 8, we conclude that the "Interaction" is more effective than the "Auto". Moreover, the effectiveness of user interaction in the case of "long" is more remarkable than that of "short". The reason why the effectiveness of user interaction in the case of "long" is more remarkable is as follows. In the case of "short", our system has to extract sentences fewer than that of "long". Even if a user had changed relevant keywords to use for sentence extraction, the sentences extracted by our system are not necessarily changed in the case of "short". However, the extracted sentences are greatly changed in the case of "long" when a user has changed relevant keywords. Hence, we consider that sentences are extracted well by changing relevant keywords in the case of "long".

Our system does not measure the degree of similarity of each relevant keyword shown to a user. Hence, similar relevant keywords may be shown to a user. For example, "アイボ" and "AIBO", which are the name of the same product, are recognized as keywords with different meaning. We consider that synonyms in a set of relevant keywords should be shown as one keyword in order to reduce user's load. For recognizing synonyms in high accuracy, including e.g., a method proposed in a paper[8], a method to extract synonyms from parenthetical expressions, and e.g., a method proposed in a paper[18], a method to extract abbreviations from the documents set may be useful.

10 Conclusion

We propose a multiple-document summarization system with user interaction. Our system extracts keywords from a documents set to be summarized and shows the extracted keywords to a user. The user selects keywords reflecting his summarization need, which controls output of summaries. We participated in TSC3 of NTCIR4 for evaluation and our system without user interaction achieved the best performance in content evaluation of “abstraction” among systems not using sets of questions. Moreover, we attained a good performance in “extraction” in the case of “short” among the systems not using sets of questions.

We simulate user interaction by using the sets of questions and make sure that the user interaction is effective.

Acknowledgment

This work was supported in part by The 21st Century COE Program “Intelligent Human Sensing”, from the Ministry of Education, Culture, Sports, Science and Technology and The Grant-in-Aid ((C)(2)13680444) from the Japan Society for the Promotion of Science of Japan.

References

[1] R. Ando, B. Boguraev, R. Byrd, and M. Neff. Multi-document summarization by visualizing topical content. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 79–88, 2000.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computatonal Linguistics*, pages 550–557, 1999.

[4] H. Edmundson. New methods in automatic abstracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, 1969.

[5] J. Goldstein, V. Mittal, J. Carbonel, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 40–48, 2000.

[6] T. Hirao and M. Okumura. Text summarization challenge 3 -text summarization evaluation at ntcir workshop 4-. In *Working Notes of the 4th NTCIR Workshop Meeting*.

[7] T. Hirao, K. Takeushi, H. Isozaki, Y. Sasaki, and E. Maeda. SVM-based multi-document summarization integrating sentence extraction with *bunsetsu* eliminate. *IEICE Trans. on Information and Systems*, E86-D(9):1702–1709, 2003.

[8] T. Hisamitsu and Y. Niwa. Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures: A comparative evaluation of bigram statistics. In *Recent Advances in Computational Terminology, Edited by D. Bourigault, C. Jacquemin, M. -C. L’Homme, John Benjamins Publishing Company*, pages 209–224, 2001.

[9] C.-Y. Lin and E. Hovy. From single to multi-document summarizaton: A prototype system and its evaluation. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, pages 457–464, 2002.

[10] I. Mani. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

[11] I. Mani and E. Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI’98)*, pages 821–826, 1998.

[12] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

[13] I. Mani and M. T.Maybury. *Advances in Automatic Text Summarization*. the MIT Press, 1999.

[14] C. Nobata and S. Sekine. A summarization system with categorization of document sets. In *Working Notes of the Third NTCIR Workshop Meeting*, pages 33–38, 2002.

[15] T. Nomoto and Y. Matsumoto. An experimental comparison of supervised and unsupervised approaches to text summarization. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 630–632, 2001.

[16] K. Ohtake, D. Okamoto, M. Kodama, and S. Masuyama. Yet another summarization system with two modules using empirical knowledge. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 225–234, 2001.

[17] H. Sakai and S. Masuyama. Unsupervised knowledge acquisition about the deletion possibility of adnominal verb phrases. In *Proceedings of Workshop on Multilingual Summarization and Question Answering 2002 (post-conference workshop to be held in conjunction with COLING-2002)*, pages 49–56, 2002.

[18] H. Sakai and S. Masuyama. Knowledge acquisition from a single corpus on correspondences between abbreviations and their original nouns for japanese nouns. In *Proceedings of The Nineth Annual Meeting of The Association for Natural Language Processing*, pages 226–229(in Japanese), 2003.

[19] H. Sakai and S. Masuyama. A statistical method for acquiring knowledge about the abbreviation possibility of some of multiple adnominal phrases. *IEICE Trans. on Information and Systems*, E86-D(9):1710–1718, 2003.

[20] T. Yoshida and S. Masuyama. Multiple document summarization system GOLD. *IEICE Trans. on Information and Systems*, E86-D(9):1719–1727, 2003.