An Automatic Method for Summary Evaluation Using Multiple Evaluation Results by a Manual Method

Hidetsugu Nanba
Faculty of Information Sciences,
Hiroshima City University
3-4-1 Ozuka, Hiroshima, 731-3194 Japan
nanba@its.hiroshima-cu.ac.jp

Manabu Okumura
Precision and Intelligence Laboratory,
Tokyo Institute of Technology
4259 Nagatsuta, Yokohama, 226-8503 Japan
oku@pi.titech.ac.jp

Abstract

To solve a problem of how to evaluate computer-produced summaries, a number of automatic and manual methods have been proposed. Manual methods evaluate summaries correctly, because humans evaluate them, but are costly. On the other hand, automatic methods, which use evaluation tools or programs, are low cost, although these methods cannot evaluate summaries as accurately as manual methods. In this paper, we investigate an automatic evaluation method that can reduce the errors of traditional automatic methods by using several evaluation results obtained manually. We conducted some experiments using the data of the Text Summarization Challenge 2 (TSC-2). A comparison with conventional automatic methods shows that our method outperforms other methods usually used.

1 Introduction

Recently, the evaluation of computer-produced summaries has become recognized as one of the problem areas that must be addressed in the field of automatic summarization. To solve this problem, a number of automatic (Donaway et al., 2000, Hirao et al., 2005, Lin et al., 2003, Lin, 2004, Hori et al., 2003) and manual methods (Nenkova et al., 2004, Teufel et al., 2004) have been proposed. Manual methods evaluate summaries correctly, because humans evaluate them, but are costly. On the other hand, automatic methods, which use evaluation tools or programs, are low cost, although these methods cannot evaluate summaries as accurately as manual methods. In this paper, we investigate an automatic method that can reduce the errors of traditional automatic methods by using several evaluation results obtained manually. Unlike other automatic methods, our method estimates manual evaluation scores. Therefore, our method makes it possible to compare a new system with other systems that have been evaluated manually.

There are two research studies related to our work (Kazawa et al., 2003, Yasuda et al., 2003). Kazawa et al. (2003) proposed an automatic evaluation method using multiple evaluation results from a manual method. In the field of machine translation, Yasuda et al. (2003) proposed an automatic method that gives an evaluation result of a translation system as a score for the Test of English for International Communication (TOEIC). Although the effectiveness of both methods was confirmed experimentally, further discussion of four points, which we describe in Section 3, is necessary for a more accurate summary evaluation. In this paper, we address three of these points based on Kazawa's and Yasuda's methods. We also investigate whether these methods can outperform other automatic methods.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 describes our method. To investigate the effectiveness of our method, we conducted some examinations and Section 4 reports on these. We present some conclusions in Section 5.

2 Related Work

Generally, similar summaries are considered to obtain similar evaluation results. If there is a set of summaries (pooled summaries) produced from a document (or multiple documents) and if these are evaluated manually, then we can estimate a manual evaluation score for any summary to be evaluated with the evaluation results for those pooled summaries. Based on this idea, Kazawa et

al. (2003) proposed an automatic method using multiple evaluation results from a manual method. First, n summaries for each document, m, were prepared. A summarization system generated summaries from m documents. Here, we represent the i^{th} summary for the j^{th} document and its evaluation score as x_{ij} and y_{ij}, respectively. The system was evaluated using Equation 1.

$$scr(x) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_j y_{ij} Sim(x, x_{ij}) + b \quad (1)$$

The evaluation score of summary x was obtained by summing parameter b for all the subscores calculated for each pooled summary, x_{ij}. A subscore was obtained by multiplying a parameter w_j, by the evaluation score y_{ij}, and the similarity between x and x_{ij}.

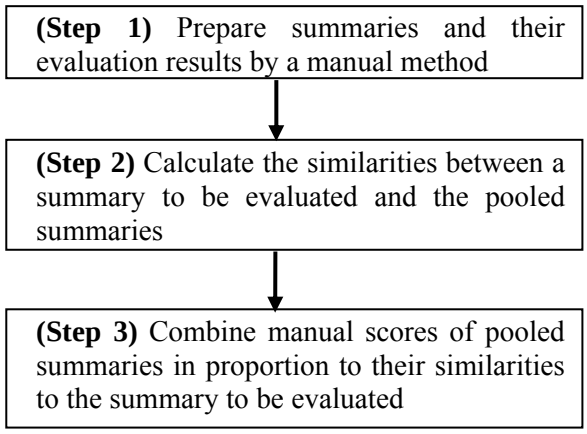
In the field of machine translation, there is another related study. Yasuda et al. (2003) proposed an automatic method that gives an evaluation result of a translation system as a score for TOEIC. They prepared 29 human subjects, whose TOEIC scores were from 300s to 800s, and asked them to translate 23 Japanese conversations into English. They also generated translations using a system for each conversation. Then, they evaluated both translations using an automatic method, and obtained W_H, which indicated the ratio of system translations that were superior to human translations. Yasuda et al. calculated W_H for each subject and plotted the values along with their corresponding TOEIC scores to produce a regression line. Finally, they defined a point where the regression line crossed W_H = 0.5 to provide the TOEIC score for the system.

Though, the effectiveness of Kazawa's and Yasuda's methods were confirmed experimentally, further discussions of four points, which we describe in the next section, are necessary for a more accurate summary evaluation.

3 Investigation of an Automatic Method using Multiple Manual Evaluation Results

3.1 Overview of Our Evaluation Method and Essential Points to be Discussed

We investigate an automatic method using multiple evaluation results by a manual method based on Kazawa's and Yasuda's method. The procedure of our evaluation method is shown as follows;



For each step, we need to discuss the following points.

(Step 1)

1. How many summaries, and what type (variety) of summaries should be prepared? Kazawa et al. prepared 6 summaries for each document, and Yasuda et al. prepared 29 translations for each conversation. However, they did not examine about the number and the type of pooled summaries required to the evaluation.

(Step 2)

2. Which measure is better for calculating the similarities between a summary to be evaluated and the pooled summaries? Kazawa et al. used Equation 2 to calculate similarities.

$$Sim(x, x_{ij}) = \frac{|\, x_{ij} \cap x \,|}{\min(|\, x_{ij} \,|, |\, x \,|)} \quad (2)$$

where $x_{ij} \cap x$ indicates the number of discourse units[1] that appear in both x_{ij} and x, and $|\, x \,|$ represents the number of words in x. However, there are many other measures that can be used to calculate the topical similarities between two documents (or passages).

As well as Yasuda's method does, using W_H is another way to calculate similarities between a summary to be evaluated and pooled summaries indirectly. Yasuda et al. (2003) tested DP matching (Su et al., 1992), BLEU (Papineni et al., 2002), and NIST[2], for the calculation of W_H. However there are many other measures for summary evaluation.

[1] Rhetorical Structure Theory Discourse Treebank. www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07 Linguistic Data Consortium.

[2] http://www.nist.gov/speech/tests/mt/mt2001/resource/

3. How many summaries should be used to calculate the score of a summary to be evaluated? Kazawa et al. used all the pooled summaries but this does not ensure the best performance of their evaluation method.

(Step 3)

4. How to combine the manual scores of the pooled summaries? Kazawa et al. calculated the score of a summary as a weighted linear sum of the manual scores. Applying regression analysis (Yasuda et al., 2003) is another method of combining several manual scores.

3.2 Three Points Addressed in Our Study

We address the second, third and fourth points in Section 3.1.

(Point 2) A measure for calculating similarities between a summary to be evaluated and pooled summaries:
There are many measures that can calculate the topical similarities between two documents (or passages). We tested several measures, such as ROUGE (Lin, 2004) and the cosine distance. We describe these measures in detail in Section 4.2.

(Point 3) The number of summaries used to calculate the score of a summary to be evaluated:
We use summaries whose similarities to a summary to be evaluated are higher than a threshold value.

(Point 4) Combination of manual scores:
We used both Kazawa's and Yasuda's methods.

4 Experiments

4.1 Experimental Methods

To investigate the three points described in Section 3.2, we conducted the following four experiments.

- **Exp-1:** We examined Points 2 and 3 based on Kazawa's method. We tested threshold values from 0 to 1 at 0.005 intervals. We also tested several similarity measures, such as cosine distance and 11 kinds of ROUGE.
- **Exp-2:** In order to investigate whether the evaluation based on Kazawa's method can outperform other automatic methods, we compared the evaluation with other automatic methods. In this experiment, we used the similarity measure, which obtain the best performance in Exp-1.
- **Exp-3:** We also examined Point 2 based on Yasuda's method. As a similarity measure, we tested cosine distance and 11 kinds of ROUGE. Then, we examined Point 4 by comparing the result of Yasuda's method with that of Kazawa's.
- **Exp-4:** In the same way as Exp-2, we compared the evaluation with other automatic methods, which we describe in the next section, to investigate whether the evaluation based on Yasuda's method can outperform other automatic methods.

4.2 Automatic Evaluation Methods Used in the Experiments

In the following, we show the automatic evaluation methods used in our experiments.

Content-based evaluation (Donaway et al., 2000)
This measure evaluates summaries by comparing their content words with those of the human-produced extracts. The score of the content-based measure is obtained by computing the similarity between the term vector using tf*idf weighting of a computer-produced summary and the term vector of a human-produced summary by cosine distance.

ROUGE-N (Lin, 2004)
This measure compares n-grams of two summaries, and counts the number of matches. The measure is defined by Equation 3.

$$ROUGE-N = \frac{\sum_{S \in R} \sum_{gram_N \in S} Count_{match}(gram_N)}{\sum_{S \in R} \sum_{gram_N \in S} Count(gram_N)} \quad (3)$$

where $\mathrm{Count}(\mathrm{gram}_N)$ is the number of an N-gram and $\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_N)$ denotes the number of n-gram co-occurrences in two summaries.

ROUGE-L (Lin, 2004)
This measure evaluates summaries by longest common subsequence (LCS) defined by Equation 4.

$$ROUGE-L = \frac{\sum_{i=i}^{u} LCS_{\cup}(r_i, C)}{m} \quad (4)$$

where $LCS_U(r_i,C)$ is the LCS score of the union's longest common subsequence between reference sentences r_i and the summary to be evaluated, and m is the number of words contained in a reference summary.

ROUGE-S (Lin, 2004)
Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. ROUGE-S measures the overlap of skip-bigrams in a candidate summary and a reference summary. Several variations of ROUGE-S are possible by limiting the maximum skip distance between the two in-order words that are allowed to form a skip-bigram. In the following, ROUGE-SN denotes ROUGE-S with maximum skip distance N.

ROUGE-SU (Lin, 2004)
This measure is an extension of ROUGE-S; it adds a unigram as a counting unit. In the following, ROUGE-SUN denotes ROUGE-SU with maximum skip distance N.

4.3 Evaluation Methods

In the following, we elaborate on the evaluation methods for each experiment.

Exp-1: An experiment for Points 2 and 3 based on Kazawa's method
We evaluated Kazawa's method from the viewpoint of "Gap". Differing from other automatic methods, the method uses multiple manual evaluation results and estimates the manual scores of the summaries to be evaluated or the summarization systems. We therefore evaluated the automatic methods using Gap, which manually indicates the difference between the scores from a manual method and each automatic method that estimates the scores. First, an arbitrary summary is selected from the 10 summaries in a dataset, which we describe in Section 4.4, and an evaluation score is calculated by Kazawa's method using the other nine summaries. The score is compared with a manual score of the summary by Gap, which is defined by Equation 5.

$$Gap = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} | scr'(x_{kl}) - y_{kl} |}{m \times n} \quad (5)$$

where x_{kl} is the k^{th} system's l^{th} summary, and y_{kl} is the score from a manual evaluation method for the k^{th} system's l^{th} summary. To distinguish our evaluation function from Kazawa's, we denote it as *scr'*(*x*). As a similarity measure in *scr'*(*x*), we tested ROUGE and the cosine distance.

We also tested the coverage of the automatic method. The method cannot calculate scores if there are no similar summaries above a given threshold value. Therefore, we checked the coverage of the method, which is defined by Equation 6.

$$Coverage = \frac{\textit{The number of summaries evaluated by the method}}{\textit{The number of given summaries}} \quad (6)$$

Exp-2: Comparison of Kazawa's method with other automatic methods
Traditionally, automatic methods have been evaluated by "Ranking". This means that summarization systems are ranked based on the results of the automatic and manual methods. Then, the effectiveness of the automatic method is evaluated by the number of matches between both rankings using Spearman's rank correlation coefficient and Pearson's rank correlation coefficient (Lin et al., 2003, Lin, 2004, Hirao et al., 2005). However, we did not use both correlation coefficients, because evaluation scores are not always calculated by a Kazawa-based method, which we described in Exp-1. Therefore, we ranked the summaries instead of the summarization systems. Two arbitrary summaries from the 10 summaries in a dataset were selected and ranked by Kazawa's method. Then, Kazawa's method was evaluated using "Precision," which calculates the percentage of cases where the order of the manual method of the two summaries matches the order of their ranks calculated by Kazawa's method. The two summaries were also ranked by ROUGE and by cosine distance, and both Precision values were calculated. Finally, the Precision value of Kazawa's method was compared with those of ROUGE and cosine distance.

Exp-3: An experiment for Point 2 based on Yasuda's method
An arbitrary system was selected from the 10 systems, and Yasuda's method estimated its manual score from the other nine systems. Yasuda's method was evaluated by Gap, which is defined by Equation 7.

$$Gap = \frac{\sum_{k=1}^{m} | s(x_k) - y_k |}{m} \quad (7)$$

where x_k is the k^{th} system, $s(x_k)$ is a score of x_k by Yasuda's method, and y_k is the manual score for the k^{th} system. Yasuda et al. (2003) tested DP matching (Su et al., 1992), BLEU (Papineni et al., 2002), and NIST[3], as automatic methods used in their evaluation. Instead of those methods, we

[3] http://www.nist.gov/speech/tests/mt/mt2001/resource/

tested ROUGE and cosine distance, both of which have been used for summary evaluation.

If a score by Yasuda's method exceeds the range of the manual score, the score is modified to be within the range. In our experiments, we used evaluation by revision (Fukushima et al., 2002) as the manual evaluation method. The range of the score of this method is between zero and 0.5. If the score is less than zero, it is changed to zero and if greater than 0.5 it is changed to 0.5.

Exp-4: Comparison of Yasuda's method and other automatic methods

In the same way as for the evaluation of Kazawa's method in Exp-2, we evaluated Yasuda's method by Precision. Two arbitrary summaries from the 10 summaries in a dataset were selected, and ranked by Yasuda's method. Then, Yasuda's method was evaluated using Precision. Two summaries were also ranked by ROUGE and by cosine distance and both Precision values were calculated. Finally, the Precision value of Yasuda's method was compared with those of ROUGE and cosine distance.

4.4 The Data Used in Our Experiments

We used the TSC-2 data (Fukushima et al., 2002) in our examinations. The data consisted of human-produced extracts (denoted as "PART"), human-produced abstracts (denoted as "FREE"), computer-produced summaries (eight systems and a baseline system using the lead method (denoted as "LEAD"))[4], and their evaluation results by two manual methods. All the summaries were derived from 30 newspaper articles, written in Japanese, and were extracted from the Mainichi newspaper database for the years 1998 and 1999. Two tasks were conducted in TSC-2, and we used the data from a single document summarization task. In this task, participants were asked to produce summaries in plain text in the ratios of 20% and 40%.

Summaries were evaluated using a ranking evaluation method and the revision method evaluation. In our experiments, we used the results of evaluation from the revision method. This method evaluates summaries by measuring the degree to which computer-produced summaries are revised. The judges read the original texts and revised the computer-produced summaries in terms of their content and readability. The human revisions were made with only three editing operations (insertion, deletion, replacement). The degree of the human revision, called the "edit distance," is computed from the number of revised characters divided by the number of characters in the original summary. If the summary's quality was so low that a revision of more than half of the original summary was required, the judges stopped the revision and a score of 0.5 was given.

The effectiveness of evaluation by the revision method was confirmed in our previous work (Nanba et al., 2004). We compared evaluation by revision with ranking evaluation. We also tested other automatic methods: content-based evaluation, BLEU (Papineni et al., 2001) and ROUGE-1 (Lin, 2004), and compared their results with that of evaluation by revision as reference. As a result, we found that evaluation by revision is effective for recognizing slight differences between computer-produced summaries.

4.5 Experimental Results and Discussion

Exp-1: An experiment for Points 2 and 3 based on Kazawa's method

To address Points 2 and 3, we evaluated summaries by the method based on Kazawa's method using 12 measures, described in Section 4.4, as measures to calculate topical similarities between summaries, and compared these measures by Gap. The experimental results for summarization ratios of 40% and 20% are shown in Tables 1 and 2, respectively. Tables show the Gap values of 12 measures for each Coverage value from 0.2 to 1.0 at 0.1 intervals. Average values of Gap for each measure are also shown in these tables. As can be seen from Tables 1 and 2, the larger the threshold value, the smaller the value of Gap. From the result, we can conclude for Point 3 that more accurate evaluation is possible when we use similar pooled summaries (Point 2). However, the number of summaries that can be evaluated by this method was limited when the threshold value was large.

Of the 12 measures, unigram-based methods, such as cosine distance and ROUGE-1, produced good results. However, there were no significant differences between measures except for when ROUGE-L was used.

[4] In Exp-2 and 4, we evaluated "PART", "LEAD", and eight systems (candidate summaries) by automatic methods using "FREE" as the reference summaries.

Table 1 Comparison of Gap values for several measures (ratio: 40%)

Measure \ Coverage	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	Average
R-1	0.080	0.070	0.067	0.057	0.064	0.062	0.058	0.045	0.041	0.062
R-2	0.082	0.074	0.070	0.070	0.069	0.063	0.059	0.051	0.042	0.065
R-3	0.083	0.074	0.075	0.071	0.069	0.063	0.059	0.051	0.045	0.066
R-4	0.085	0.078	0.076	0.073	0.069	0.064	0.060	0.051	0.043	0.067
R-L	0.102	0.100	0.097	0.094	0.091	0.090	0.089	0.082	0.078	0.091
R-S	0.083	0.077	0.073	0.073	0.069	0.067	0.064	0.060	0.045	0.068
R-S4	0.083	0.072	0.071	0.069	0.066	0.066	0.060	0.054	0.044	0.065
R-S9	0.083	0.075	0.069	0.070	0.067	0.066	0.066	0.057	0.046	0.067
R-SU	0.083	0.077	0.070	0.071	0.069	0.068	0.064	0.057	0.043	0.067
R-SU4	0.082	0.073	0.069	0.069	0.065	0.068	0.063	0.051	0.043	0.065
R-SU9	0.083	0.074	0.070	0.068	0.066	0.067	0.066	0.054	0.046	0.066
Cosine	0.081	0.074	0.065	0.062	0.059	0.056	0.057	0.039	0.043	**0.059**
Threshold	Small ◄								► Large	

Table 2 Comparison of Gap values for several measures (ratio: 20%)

Measure \ Coverage	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	Average
R-1	0.129	0.104	0.102	0.976	0.090	0.089	0.089	0.083	0.082	0.096
R-2	0.132	0.115	0.107	0.109	0.096	0.093	0.079	0.081	0.082	0.099
R-3	0.132	0.115	0.116	0.111	0.102	0.092	0.080	0.078	0.079	0.101
R-4	0.134	0.121	0.121	0.112	0.103	0.090	0.080	0.080	0.078	0.102
R-L	0.140	0.135	0.134	0.125	0.117	0.110	0.105	0.769	0.060	0.111
R-S	0.130	0.119	0.113	0.106	0.098	0.099	0.089	0.089	0.087	0.103
R-S4	0.130	0.114	0.109	0.105	0.102	0.092	0.085	0.088	0.085	0.101
R-S9	0.130	0.119	0.113	0.105	0.095	0.097	0.095	0.085	0.084	0.103
R-SU	0.130	0.118	0.109	0.109	0.097	0.098	0.088	0.089	0.079	0.102
R-SU4	0.130	0.111	0.107	0.106	0.100	0.090	0.086	0.084	0.087	0.100
R-SU9	0.130	0.116	0.108	0.105	0.096	0.090	0.085	0.085	0.082	0.099
Cosine	0.128	0.106	0.102	0.094	0.091	0.090	0.079	0.080	0.057	**0.092**
Threshold	Small ◄								► Large	

Exp-2: Comparison of Kazawa's method with other automatic methods (Point 2)

In Exp-1, cosine distance outperformed the other 11 measures. We therefore used cosine distance in Kazawa's method in Exp-2. We ranked summaries by Kazawa's method, ROUGE and cosine distance, calculated using Precision.

The results of the evaluation by Precision for summarization ratios of 40% and 20% are shown in Figures 1 and 2, respectively. We plotted the Precision value of Kazawa's method by changing the threshold value from 0 to 1 at 0.05 intervals. We also plotted the Precision values of ROUGE-2 as dotted lines. ROUGE-2 was superior to the other 11 measures in terms of Ranking. The X and Y axes in Figures 1 and 2 show the threshold value of Kazawa's method and the Precision values, respectively. From the result shown in Figure 1, we found that Kazawa's method outperformed ROUGE-2, when the threshold value was greater than 0.968. The Coverage value of this point was 0.203. In Figure 2, the Precision curve of Kazawa's method crossed the dotted line at a threshold value of 0.890. The Coverage value of this point was 0.405.

To improve these Coverage values, we need to prepare more summaries and their manual evaluation results, because the Coverage is critically dependent on the number and variety of pooled summaries. This is exactly the first point in Section 3.1, which we do not address in this paper. We will investigate this point as the next step in our future work.

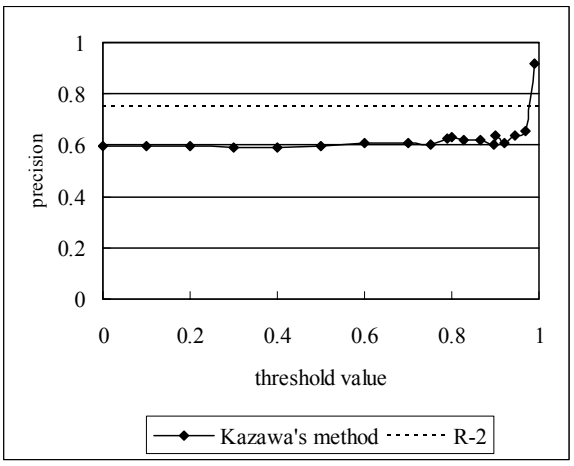


Figure 1 Comparison of Kazawa's method and ROUGE-2 (ratio: 40%)

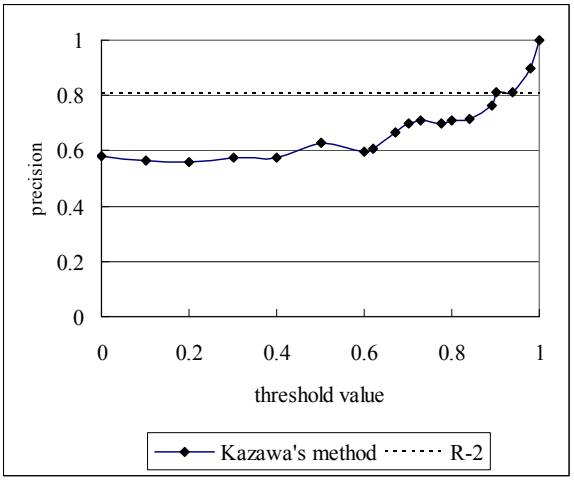


Figure 2 Comparison of Kazawa's method and ROUGE-2 (ratio: 20%)

Exp-3: An experiment for Point 3 based on Yasuda's method

For Point 2 in Section 3.2, we also examined Yasuda's method. The experimental result by Gap is shown in Table 3. When the ratio is 20%, ROUGE-SU4 is the best. The N-gram and the skip-bigram are both useful when the summarization ratio is low.

For Point 4, we compared the result by Yasuda's method (Table 3) with that of Kazawa's method (in Tables 1 and 2). Yasuda's method could accurately estimate manual scores. In particular, the Gap values of 0.023 by ROUGE-2 and by ROUGE-3 are smaller than those produced by Kazawa's method with a threshold value of 0.9 (Tables 1 and 2). This indicates that regression analysis used in Yasuda's method is superior to that used in Kazawa's method.

Table 3 Gap between the manual method and Yasuda's method

	Ratio		Average
	20%	40%	
Cosine	0.037	0.031	0.035
R-1	0.033	**0.022**	0.028
R-2	0.028	0.023	**0.025**
R-3	0.028	0.023	**0.025**
R-4	0.036	0.024	0.030
R-L	0.040	0.038	0.039
R-S(∞)	0.051	0.060	0.055
R-S4	0.025	0.040	0.033
R-S9	0.042	0.052	0.047
R-SU(∞)	0.027	0.055	0.041
R-SU4	**0.022**	0.037	0.029
R-SU9	0.023	0.048	0.036

Exp-4: Comparison of Yasuda's method with other automatic methods

We also evaluated Yasuda's method by comparison with other automatic methods in terms of Ranking. We evaluated 10 systems by Yasuda's method with ROUGE-3, which produced the best results in Exp-3. We also evaluated the systems by ROUGE and cosine distance, and compared the results. The results are shown in Table 4.

Table 4 Comparison between Yasuda's method and automatic methods

	Ratio		Average
	20%	40%	
Yasuda	**0.867**	**0.844**	**0.856**
Cosine	0.844	0.800	0.822
R-1	0.822	0.778	0.800
R-2	0.844	0.800	0.822
R-3	0.822	0.800	0.811
R-4	0.822	**0.844**	0.833
R-L	0.822	0.800	0.811
R-S(∞)	0.667	0.689	0.678
R-S4	0.800	0.756	0.778
R-S9	0.733	0.689	0.711
R-SU(∞)	0.711	0.711	0.711
R-SU4	0.800	0.822	0.811
R-SU9	0.756	0.711	0.733

As can be seen from Table 4, Yasuda's method produced the best results for the ratios of 20% and 40%. Of the automatic methods compared, ROUGE-4 was the best.

As evaluation scores by Yasuda's method were calculated based on ROUGE-3, there were no striking differences between Yasuda's method and the others except for the integration process of evaluation scores for each summary. Yasuda's method uses a regression analysis, whereas the other methods average the scores for each summary. Yasuda's method using ROUGE-3 outperformed the original ROUGE-3 for both ratios, 20% and 40%.

5 Conclusions

We have investigated an automatic method that uses several evaluation results from a manual method based on Kazawa's and Yasuda's methods. From the experimental results based on Kazawa's method, we found that limiting the number of pooled summaries could produce better results than using all the pooled summaries. However, the number of summaries that can be evaluated by this method was limited. To improve the Coverage of Kazawa's method, more summaries and their evaluation results are required, because the Coverage is critically dependent on the number and variety of pooled summaries.

We also investigated an automatic method based on Yasuda's method and found that the method using ROUGE-2 and -3 could accurately estimate manual scores, and could outperform Kazawa's method and the other automatic methods tested. From these results, we can conclude that the automatic method performed the best when ROUGE-2 or 3 is used as a similarity measure, and a regression analysis is used for combining manual method.

References

Robert L. Donaway, Kevin W. Drummey and Laura A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization: 69–78.

Takahiro Fukushima and Manabu Okumura. 2001. Text Summarization Challenge/Text Summarization Evaluation at NTCIR Workshop2. *Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization*: 45–51.

Takahiro Fukushima, Manabu Okumura and Hidetsugu Nanba. 2002. Text Summarization Challenge 2/Text Summarization Evaluation at NTCIR Workshop3. *Working Notes of the 3rd NTCIR Workshop Meeting, PART V*: 1–7.

Tsutomu Hirao, Manabu Okumura, and Hideki Isozaki. 2005. Kernel-based Approach for Automatic Evaluation of Natural Language Generation Technologies: Application to Automatic Summarization. *Proceedings of HLT-EMNLP 2005*: 145–152.

Chiori Hori, Takaaki Hori, and Sadaoki Furui. 2003. Evaluation Methods for Automatic Speech Summarization. *Proceedings of Eurospeech 2003*: 2825–2828.

Hideto Kazawa, Thomas Arrigan, Tsutomu Hirao and Eisaku Maeda. 2003. An Automatic Evaluation Method of Machine-Generated Extracts. *IPSJ SIG Technical Reports, 2003-NL-158*: 25–30. (in Japanese).

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. *Proceedings of the Human Language Technology Conference 2003*: 71–78.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the ACL-04 Workshop "Text Summarization Branches Out"*: 74–81.

Hidetsugu Nanba and Manabu Okumura. 2004. Comparison of Some Automatic and Manual Methods for Summary Evaluation Based on the Text Summarization Challenge 2. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*: 1029–1032.

Ani Nenkova and Rebecca Passonneau, 2004. Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of HLT-NAACL 2004*: 145–152.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2001. BLEU: A Method for Automatic Evaluation of Machine Translation. *IBM Research Report,* RC22176 (W0109-022).

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A New Quantitative Quality Measure for Machine Translation Systems. *Proceedings of the 14th International Conference on Computational Linguistics*: 433–439.

Simone Teufel and Hans van Halteren. 2004. Evaluating Information Content by Factoid Analysis: Human Annotation and Stability. *Proceedings of EMNLP 2004*: 419–426.

Kenji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto and Masuzo Yanagida. 2003. Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System. *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*: 371–378.