A Comparison of Summarization Methods Based on Task-based Evaluation

MOCHIZUKI Hajime and OKUMURA Manabu

School of Information Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292, Japan
E-mail:{motizuki,oku}@jaist.ac.jp

Abstract

A task-based evaluation scheme has been adopted as a new method of evaluation for automatic text summarization systems. It evaluates the performance of a summarization system in a given task, such as information retrieval and text categorization. This paper compares ten different summarization methods based on information retrieval tasks. In order to evaluate the system performance, the subjects' speed and accuracy are measured in judging the relevance of texts using summaries. We also analyze the similarity of summaries in order to investigate the similarity of the methods. Furthermore, we analyze what factors can affect evaluation results, and describe the problems that arose from our experimental design, in order to establish a better evaluation scheme.

1. Introduction

The importance of automatic text summarization research has been now increasing with the growing availability of on-line documents(Mani et al., 1998). Especially, the recent prevalence of information retrieval engines has created an important application for displaying retrieval results by using of the automatic summarization, whereby the user can quickly and accurately judge the relevance of texts returned as the result of a query. Here, rather than producing a generic summary, the summary that reflects the user's information need expressed in the query, 'query-biased summary,' would be considered as more suitable.

The traditional evaluation method in summarization research has been to measure the similarity between summaries that are produced automatically and by hand. However, this evaluation method has been criticized because it assumes that there is only one correct summary. A task-based evaluation scheme has been recently adopted as new way of evaluating summaries(Jing et al., 1998; Mani et al., 1998; Tombros and Sanderson, 1998). It evaluates the performance of a summarization system in a given task, such as information retrieval and text categorization.

This paper compares ten different summarization methods based on information retrieval tasks. To evaluate the system performance, subjects' speed and accuracy are measured when they judge the relevance of texts using summaries. This evaluation method has the advantage that it can evaluate the utility of a summarization system in the environment in which it is actually used, and for the purpose for which it is built.

Previously, the TIPSTER Text Summarization Evaluation (SUMMAC)(Mani et al., 1998) adopted the task-based evaluation scheme and compared the performance of multiple systems. However, what features in a system contribute to producing a good (or bad) summary has not yet been clarified. Therefore, we implement ten different summarization methods and compare them in the context of information retrieval tasks. From the results, we try to clarify what kinds of characteristics in summaries are good for information retrieval tasks.

The summarization methods we use for the comparison are:

(1) use a full text itself;

(2) use a document title;

(3) extract a few leading sentences;

(4) extract paragraphs that are related to a query;

(5,6) produce a generic/query-biased summary by extracting sentences based on term frequency;

(7,8) produce a generic/query-biased summary by extracting sentences based on lexical chains;

(9) use a summary produced by a commercial summarizer; and

(10) extract a passage using lexical chains.

We fix the length of summaries as 20% sentences of the full texts. The methods can be classified along two dimensions: First, methods can be grouped based on the degree of continuity of their summaries. And second, methods can be grouped based on whether they reflect the user's topic of interest (query-biased) or not (generic). We also analyze the similarity of the summaries in order to investigate the similarity of the methods.

In the experiment, 30 subjects are shown 10 queries and a list of 20 texts (summaries) per query, from BMIR-J2(Kitani et al., 1998), the test collection for Japanese IR systems. We ask them to judge the relevance of texts to a query and write down the total time they spent on 20 texts for each query. They can access the full text if they need it during the judgment. And the number of times that they access the full text is also recorded. Furthermore, we also evaluate the readability of the summaries as Japanese texts.

We use the following four criteria for comparison:

(1) accuracy of the judgments;

(2) time required for the task;

(3) the number of times when subjects need the help of full texts; and

(4) readability of summaries as texts.

Accuracy is measured by recall, precision, and F-measure.

Finally, in order to establish a better evaluation scheme, we analyze what factors can affect evaluation results, and describe the problems that arose from our experimental design, such as the selection of appropriate queries and documents.

In the next section, we explain ten summarization methods which we implemented. In section 3, we present the experimental procedure to compare their performance. In section 4, we analyze the results obtained from the experiments. In section 5, we describe the problems that arose from our experimental design.

2. Summarization Methods

We use the following ten summarization methods for comparison.

(1) use a full text (**full**)
The original text is presented as it is.

(2) use a document title (**title**)
The title of a text is presented as a summary.

(3) extract a few leading sentences (**lead**)
It has been said that the leading few sentences of an article are important and provide a good summary(Brandow et al., 1995). This method extracts the first 20% sentences of the article which included the title as a generic summary.

(4) extract paragraphs that are related to a query (**f-seg**)
The similarity between a query and each paragraph in a document is calculated and then the paragraph with the biggest similarity is extracted as a 'query-biased' summary. If the length of the extracted paragraphs exceeds 20% of the full text, we select the first 20% of sentences from the paragraphs.

The following formula (1) is used to calculate the similarity between a query vector Q and each paragraph vector D_j:

$$sim(Q, D_j) = \sum_i (tf_{q_i} \times log\frac{N}{df_i})^2 \times w_i, \qquad (1)$$

where tf_{q_i} is the frequency of word i in the query, N is the total number of paragraphs in the document set, and df_i is the number of paragraphs in which word i occurs. The importance score, w_i, of term i can be calculated by the standard tf.idf method(Salton and Buckley, 1988), as follows:

$$w_i = tf_i \times log\frac{N}{df_i}, \qquad (2)$$

where tf_i is the term frequency of word i in the paragraph.

The document title is also treated as a paragraph.

(5) produce a generic summary by extracting sentences based on term frequency (**tf.idf**)
The frequency of term occurrences within a document has often been used for calculating the importance of sentences(Luhn, 1958; Zechner, 1996). In this method, sentences are scored as the sum of the scores of the words in the sentence. The score, S_j, of sentence j is calculated by the following formula:

$$S_j = \sum_i w_i. \qquad (3)$$

Similar to the method that extracts paragraphs, the importance score, w_i, of word i is calculated by formula (2), except that tf_i means the term frequency of word i in the document, N is the total number of documents, and df_i is the document frequency of word i in the whole set of documents.

The top 20% of sentences are extracted as a generic summary by the term frequency method.

(6) produce a query-biased summary by extracting sentences based on term frequency (**q-tf.idf**)
Sentences are scored by the same formula (3) as in the generic term frequency method. However, words are scored to bias toward the query terms by the following formula:

$$w_i = \begin{cases} tf_i \times log\frac{N}{df_i} & \text{for non-query terms,} \\ \alpha \times tf_i \times log\frac{N}{df_i} & \text{for query terms,} \end{cases} \qquad (4)$$

where α is a constant set to 3, based on the results of preliminary experiments, to produce summaries that are different from the ones produced by the generic term frequency method.

This method is an earlier method for producing query-biased summaries(Tombros and Sanderson, 1998).

(7) produce a generic summary by extracting sentences based on lexical chains (**cf.idf**)
In this method, the importance of a sentence is calculated based on the importance of lexical chains in the sentence.

Lexical chains(Morris and Hirst, 1991) are sequences of words that are in a lexical cohesion relation(Halliday and Hasan, 1976) with each other, and tend to indicate the topics which exist in the document. There are several methods to calculate lexical cohesion between words: using a thesaurus, such as WordNet(Miller, 1990), that records the synonymy and hyponymy relationships between words, or estimating the degree of semantic similarity between words using co-occurrence information between words in a corpus. We use the latter method in this paper.

The semantic similarity score between words X and Y is calculated by the cosine distance, as in the following formula (5):

$$sim(X,Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}, \quad (5)$$

where x_i and y_i are the term frequency of words X and Y in document i, and n is the total number of documents in the corpus. This method calculates the similarity score between words based on the degree of their co-occurrence in the same documents. We used a year's worth of newspaper articles as the corpus for calculating the similarity score. The number of articles is 101,058 and the total number of words in the corpus is 11,429,112.

Using this similarity score, we construct clusters of words. In calculating the similarity score between clusters C_i and C_j, we use the shortest distance method:

$$SIM(C_i, C_j) = \max_{X \in C_i, Y \in C_j} sim(X,Y). \quad (6)$$

Sequences of words in the same clusters are then regarded as lexical chains. In the 200 documents that we will use in the experiments in the next section, 56,544 clusters of words (lexical chains) are constructed.

Using the lexical chains, sentences are scored as the sum of the scores of the lexical chains in the sentence. Similar to the case of the term frequency method, the score, S_j, of sentence j is calculated by the formula (3), except that w_i indicates the importance of lexical chain i.

To calculate the importance score, w_i, of a lexical chain i in a document, we define the following formula based on the standard $tf.idf$ measure:

$$w_i = \mid i \mid \times log \frac{N}{df_i}, \quad (7)$$

where $\mid i \mid$ is the number of terms in chain i, N is the total number of documents in the document set, and df_i is the document frequency of chain i in the whole set of documents. The top 20% of sentences are extracted as a generic summary.

(8) produce a query-biased summary by extracting sentences based on lexical chains (**q-cf.idf**)
Sentences are scored by the same formula (3) as in the generic sentence extraction method based on lexical chains. However, lexical chains are scored to bias toward the query terms by the following formula:

$$w_i = \begin{cases} \mid i \mid \times log \frac{N}{df_i} & \text{for chains that do not include query terms,} \\ \alpha \times \mid i \mid \times log \frac{N}{df_i} & \text{for chains that include query terms,} \end{cases} \quad (8)$$

where α is a constant set to 3, similar to **q-tf.idf**, to produce summaries that are different from the ones produced by the generic summarization method. The top 20% of sentences are extracted as a query-biased summary based on lexical chains.

(9) use a summary by one of commercial summarizers (**J**)
The top 20% sentences are selected as a summary by one of the commercial summarizers. We select the best summarizer based on the result of a preliminary experiment where we compare three Japanese word processing softwares with summarization function.

(10) extract a passage using lexical chains (**lex**)
In this method, passages are extracted as a summary by calculating the similarity between a query and a document by our passage retrieval method based on lexical chains(Mochizuki et al., 2000).

Passages that are related to the query can be extracted by first searching the lexical chains that include query terms. Consider Figure 1. Three query terms match lexical chains A1, A2, and B1, B2, and C1, C2, C3, respectively.

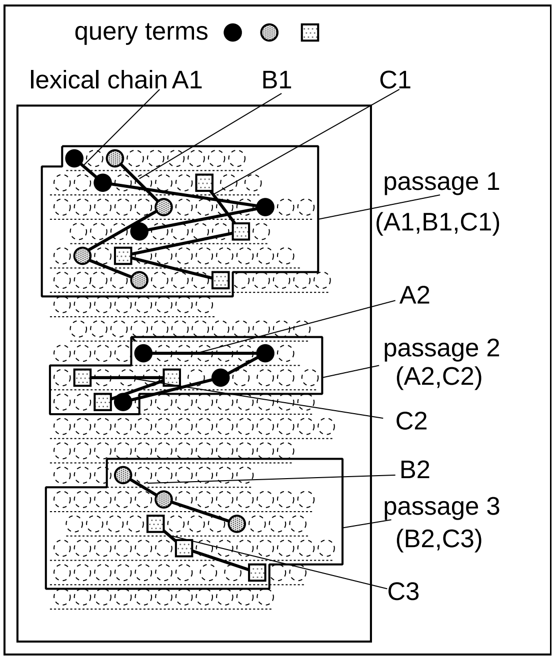

Figure 1: An example of a passage

Passages are then fixed as the maximal fragments that are covered by overlapped lexical chains for query terms. Therefore, in Figure 1, three passages are extracted based on the three query terms.

The passage is then scored and used to determine the best passage for a document. The score is based on two factors:

- the scores of lexical chains that are included in the passage,
- the degree of the overlap of the lexical chains in the passage.

Lexical chains are scored based on the number of words in the chain. The passage is then scored as the sum of the scores of the lexical chains in the passage, weighted by the degree of overlap of the chains. The weight is set to the square of the number of the overlapped chains. Therefore, a passage in which many query terms densely co-occur will be extracted for the query. In Figure 1, passage 1 will be selected as the passage for the document, since it contains three long chains overlapping each other.

3. Experiments

To compare the summaries made by the methods explained in the last section, we performed a task-based evaluation experiment. It assumes that summaries can be evaluated indirectly by the accuracy and speed of a given task, such as information retrieval and text categorization, where human subjects use the summaries for the task. We select information retrieval as the task for evaluation.

It measure the effectiveness of how the summaries help the users to judge the relevance of documents. Our task-based evaluation follows the summarization evaluation proposed for TIPSTER III(Hand, 1997; Mani et al., 1998).

We evaluated titles and summaries from 8 summarization methods at 20% length[1] as well as full texts in an information retrieval task. We measure subjects' accuracy and speed in judging relevance.

3.1. Material and methods

The documents that are used for evaluation are articles from the BMIR-J2 test collection(Kitani et al., 1998). The BMIR-J2 consists of 5,080 articles from the Mainichi newspaper and 50 queries[2]. Ten queries are selected and 20 documents for each query are semi-automatically selected from the collection using an IR system. Manual intervention was needed to keep the percentage of documents relevant to the query higher than 50% because a smaller number of relevant documents would make the results of the experiments less reliable. The average length of the queries is 3.2 words, and the average length of the documents is 1,323.3 bytes. The average number of relevant documents is 12.8 in 20 documents.

We used 30 subjects (all are postgraduate students in the school of information science) and asked them to judge the relevance of documents to a query and write down the total time they spent on the 20 documents for each query. The subjects were also required to evaluate readability of the summaries as Japanese texts according to the four-grade evaluation: (1) the summary is readable; (2) a little readable; (3) a little unreadable; and (4) unreadable.

Since the human subjects can read each query-document pair only once, the 30 subjects are divided into 10 groups (each group consists of 3 subjects) and the 10 groups rotate through the 10 forms of documents (including full texts) and the 10 queries.

The subjects are also instructed that they can access the full texts as well as the documents that are presented. The number of times that the full text is accessed is also recorded. Since the subjects are not told which one of the presented documents is the full text, they may try to access the full text even if they have already read it, if they think they need it.

3.2. Evaluation criteria

We use the following four criteria for comparison:

(1) accuracy of the judgments
Accuracy is measured by recall, precision, and F-measure. Recall (R) is the ratio of the number of relevant documents correctly judged by subjects to the total number of relevant documents. Precision (P) is the ratio of the number of relevant documents correctly judged by subjects to the total number of documents that subjects judge as relevant. F-measure is calculated by the following formula:

$$F - measure = \frac{2 \times P \times R}{P + R}. \quad (9)$$

(2) time required for the task
The average times which the subjects spent in the task at each query are calculated and are compared.

(3) the number of times the subjects need the help of full texts

(4) readability of summaries as texts
Four-grade evaluation is scored as follows: 'readable' as 10, 'a little readable' as 5, 'a little unreadable' as -5, and 'unreadable' as -10.

4. Results and Comparisons

4.1. Results

Although we assumed that performance would be almost uniform among subjects, upon examining the accuracy data we found some differences in their F-measures.

[1]We admit that we should make more thorough experiments with multiple summary lengths, since different summary lengths will yield different results(Jing et al., 1998; Mittal et al., 1999).

[2]BMIR-J2 was constructed by the SIG Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

The distribution of subjects' average F-measures is shown in Figure 2. Since, the performance of 21 out of 30 subjects is rather close (enclosed in the box), we used the results of those 21 subjects for evaluation.

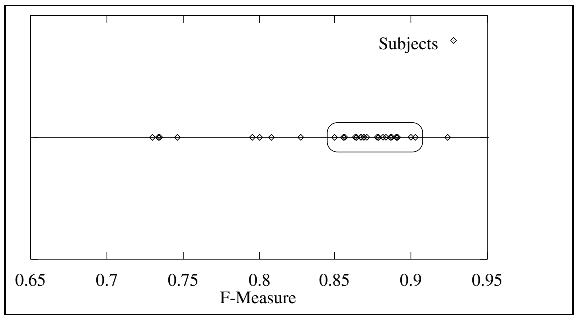

Figure 2: Distribution of F-measure values of subjects

The results of the experiments are summarized in Table 1. The results are shown as averages per query. 'Readability' is the average score which is calculated from the results of the four-grade judgment by subjects. 'No. of references' indicates the number of times when the full text is accessed. 'Discontinuity' indicates the average number of sentences in the full text between the adjoining two sentences in the summary. The summary is said to be continuous if its discontinuity is close to 0. 'Sum. ratio in word number' indicates the average length of summaries when it is calculated in the number of words.

We tested the statistical significance of the agreement among subjects. Using the same methodology as in (Jing et al., 1998; Passonneau and Litman, 1993), we performed Cochran's Q-test(Degroot et al., 1981) on the data from the subjects. For our task, Cochran's Q-test evaluates the null hypothesis that the total number of human subjects judging the same document as relevant is randomly distributed. The results show that this hypothesis is false and that the agreement among subjects is significant. For all ten queries, the probability that subjects judge a document as relevant is much higher than would be expected by chance ($p < 10^{-5}$).

4.2. Comparisons

In this subsection, we analyze the results from the experiments. First, we compare the results among all methods in the four evaluation criteria. Next, we classify the methods into groups based on the similarity of their summaries and compare the groups in the same criteria.

4.2.1. Comparison among all methods

When all the summarization methods are compared by the mean F-measure value, our method (**lex**) and **J** outperform the full text (**full**), and **lex** is the best. The accuracy of others are almost same as or slightly lower than **full**. However, when we perform the one-way ANOVA using their F-measures, the statistical significance is not reached in the result ($p < 0.9725$). A major reason of the small difference of accuracy can be considered that the subjects can access the full text if they need it during the judgment in the experiment. Accessing the full text might make the performance of a summary better than the real one with only the summary. Therefore, we will have to adjust the accuracy score according to the number of times the full text is accessed. However, since it is difficult to guess the influence of the full text access, we have not adjusted it.

As for the times required for the task, it can be said that all summarization methods save time, since **full** takes the longest. The time is remarkably short with the **title** method while the other methods show similar times.

It can be considered that **title** shows the best overall performance, since it needs a considerably short amount of time for the task, although it is slightly inferior to **lex** in terms of accuracy. Therefore, as one of better design for the information retrieval system, we can adopt a system that displays the title as a summary and can access the full text if it is necessary. However, the title is not always attached to a text. Therefore, the **title** method can not always be used. On the other hand, it can be said that **lex** is the best overall in the automatic summarization methods which can be used for the text without the title.

As for readability, **J, lead**, and **cf.idf** can yield higher evaluation scores than **full**. The score of **lex** is same as **full**. The scores of other methods are lower than **full**. The score of **title** is the lowest. The methods which produce summaries with high readability tend to extract sentences from the leading part of the original text, as will be described later.

4.2.2. Comparison based on the similarity of summaries

Next, we classify the summarization methods into groups based on the similarity of their produced summaries, and compare the groups in the four criteria.

The summaries are classified by the following steps. First, each summary is expressed as a binary vector of an original document. Each element of the vector corresponds to each sentence in the document, and its value is 1 if the sentence is selected in the summary. Second, the similarity between each pair of vectors is calculated by the cosine distance. Finally, we construct clusters of the summarization methods using this similarity. In calculating the similarity between two clusters, we use two different measurement methods, the shortest distance method and the mean distance method. Figure 3 shows the resulting clusters.

In this figure, the summarization methods are classified into three groups with both measurement methods: **Group 1** includes **J, lead** and **lex**; **Group 2** includes **tf.idf, cf.idff, q-tf.idf** and **q-cf.idf**; and **Group 3** includes **f-seg**. We don't consider **full** and **title**, since the comparison of the similarity with them does not make sense.

In **Group 1**, all three summarization methods have the common feature that the summaries with high continuity are extracted. Both the similarity between **lex** and **lead**, and between **J** and **lead** are high, while the similarity be-

	full	title	lead	f-seg	tf.idf	q-tf.idf	cf.idf	q-cf.idf	lex	J
Recall	87.1%	86.7	85.9	87.2	86.3	89.6	87.0	85.3	90.5	86.5
Precision	89.0%	89.1	88.9	88.8	89.7	85.3	85.3	87.0	88.5	91.3
F-measure	87.2%	87.0	86.6	87.1	87.2	86.6	84.9	84.9	89.1	87.6
Readability	4.1	1.8	4.6	3.7	3.8	4.0	4.3	3.9	4.1	5.5
Time (min:sec)	15:38	7:54	9:47	10:55	10:37	9:54	10:54	10:29	10:41	10:52
No. of references	0.6	4.8	3.8	2.8	2.6	1.7	2.0	1.5	1.9	2.0
Discontinuity	0.0	0.0	0.0	0.0	3.6	3.5	3.8	3.6	0.0	1.4
Sum. ratio in word number	100.0%	5.3	19.1	21.7	32.1	31.5	30.5	30.2	23.6	27.7

Table 1: Experimental results

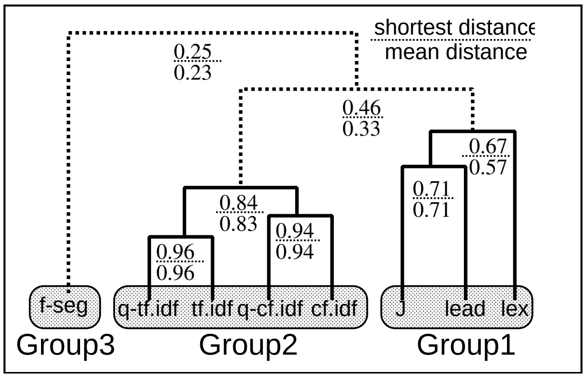

Figure 3: Similarity among the summarization methods

tween **lex** and **J** is not so high. Therefore, it can be inferred that both summaries of **lex** and **J** include the leading part of an original text to a certain extent, and **J** also includes certain sentences which are distant from the leading part of the original document, although the detailed method of **J** is unknown.

In **Group 2**, all four summarization methods have the common feature that the summaries with low continuity are extracted. Although those methods can be classified in two ways: whether the score is calculated by taking into account a query, and which of the score of term frequency or lexical chains (**tf** or **cf**) is used for calculating a score, it can be said that these differences do not affect in this group of the summaries, since the similarity between summaries of any two methods are very high.

Group 3 consists only of **f-seg**. Similar to the summarization methods in **Group 1**, **f-seg** extracts summaries with high continuity, though selected sentences are greatly different from those of any methods in **Group 1**.

Lex and **J**, which yield higher accuracy both belong to **Group 1**. However, it can be said that the summary that is simply extracted from the leading part is not so effective, since the accuracy of **lead** in **Group 1** is not so good.

We can also compare the summarization methods of each group from the point of view of query biasing. In both **Group 1** and **Group 3**, the relevance judgment can be done more easily with query-biased summaries (less time required, and less full-text access), and the accuracy of the judgment is also improved. In case of the methods which produce a summary with higher continuity, the summaries are effective in improving the accuracy if they are produced by taking into a query account. The methods in **Group 1** also obtain relatively higher score in the evaluation of readability. In case of **Group 2**, the relevance judgment can be done more easily with query-biased summaries (less time required, and less full-text access), but the accuracy of the judgment is not necessarily improved.

5. Conclusion and Future work

In this paper, we compared ten different summarization methods based on the task of information retrieval. From the experimental results, we got the following major findings:

- The methods which produce the summaries of high readability tend to extract sentences from the leading part of the original text.
- The summaries with high continuity are more readable and yield higher accuracy in general.
- Relevance judgment can be done more easily with query-biased summaries (less time required, and less full text access), but the accuracy of the judgment is not always improved. In the case of methods which produce a summary with high continuity, the summaries are effective in improving the accuracy if they are produced by taking into account a query.
- Any summarization method saves time.

From the experience of this evaluation experiment, we enumerated the followings as points that should be considered for future evaluation experiments.

- Elimination of the influence of accessing the full text: The statistical significance could not be reached in the comparison among all summarization methods in the evaluation. A major reason of the small difference can be considered that the reference of the full text was allowed in the experiment, though the degree of the influence is difficult to estimate. Moreover, it can be considered that the time required for the task is also affected by the reference of the full text, though the degree of the influence is difficult to estimate. Therefore, we should not allow the reference of the full text in the experiment of relevance judgment.

- Separation of the evaluation of readability:
  In the experiment, the difference of times required for the task among the summarization methods was not remarkable except for **title**. A major reason of the small difference can be considered that the evaluation of readability was done together with the relevance judgment.

  Since the required time for evaluating readability is longer than the time for judging relevance, it can be inferred that the difference of the time for the relevance judgment has been narrowed by readability judgment. Therefore, we should separate two experiments completely.

- Selection of queries and documents:
  We think the following combinations of a query and documents are not suitable for the comparative evaluation among summarization methods:

  1. The combination of documents and a query where words in the query are scattered comparatively equally in the documents.
     In this case, summaries which are good for judging the relevance are easily extracted from any part in the document.
  2. The combination of documents and a query where words in the query occur in the documents frequently.
     In this case, similar sentences tend to be extracted as a summary in both query-biased and generic methods.

  To further clarify the difference among the summarization methods, we must choose a set of queries and documents by taking into consideration at least the above points. However, the criterion for the selection of the suitable evaluation set is a difficult problem and not clear at present.

6. Acknowledgments

The authors would like to express our gratitude to Dr. Akihiko Takano, Dr. Yoshiki Niwa, Dr. Shingo Nishioka and other members in the Central Research Laboratory, at Hitachi co., Ltd. for their helpful suggestions and for allowing us to use their system for word association calculation. We are also grateful to the SIG Database Systems of the Information Processing Society of Japan for allowing us to use their test collection BMIR-J2.

7. References

Brandow, R., K. Mitze, and L.F. Rau, 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing & Management*, 31(5):675–685.

Degroot, M. H. et al. (eds.), 1981. *Encyclopedia of Statistical Sciences*, volume 2. A Wiley-Interscience Publication, pages 24–26.

Halliday, H.A.K. and R. Hasan, 1976. *Cohesion in English*. Longman.

Hand, T.F., 1997. A proposal for task-based evaluation of text summarization systems. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*. pages 31–38.

Jing, H., R. Barzilay, K. McKeown, and M. Elhadad, 1998. Summarization Evaluation Methods: Experiments and Analysis. In *Intelligent Text Summarization*. AAAI Press, pages 51–59.

Kitani, T. et al., 1998. Japanese Test Collection for Evaluation of Information Retireval Systems, BMIR-J2. In *IPSJ SIG-DBS-144-3*. pages 15–22. In Japanese.

Luhn, H.P., 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Mani, I. et al., 1998. The tipster summac text summarization evaluation. Technical Report MTR 98W0000138, MITRE Technical Report.

Miller, G., 1990. Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235–312.

Mittal, V., M. Kantrowitz, J. Goldstein, and J. Carbonell, 1999. Selecting text spans for document summaries: Heuristics and metrics. In *Proc. of the 16th National Conference on Artificial Intelligence*. pages 467–473.

Mochizuki, H., M. Iwayama, and M. Okumura, 2000. Passage-Level Document Retrieval Using Lexical Chains. In *Proc. of the 6th RIAO Conference*. To appear.

Morris, J. and G. Hirst, 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

Passonneau, R.J. and D.J. Litman, 1993. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *31st Annual Meeting of the Association for Computational Linguistics*. pages 148–155.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.

Tombros, A. and M. Sanderson, 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. of 21st Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*. pages 2–10.

Zechner, K., 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proc. of 16th International Conference on Computational Linguistics*. pages 986–989.