Information Gain Ratio meets Maximal Marginal Relevance

— A method of Summarization for Multiple Documents —

Tatsunori MORI and Takuro SASAKI
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{mori,takuro}@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method to make a summary from multiple documents with taking account of comprehensibility and readability. As for comprehensibility, we show an integration of MMR into the term-weighting method based on IGR. As for readability, we propose a method to generate a summary based on clustering important sentences according to sub-topics and making a keyword list as a very brief summary for each cluster.

By the evaluation in NTCIR3 TSC2, we show that the proposed method works well to generate comprehensive summaries when the length of summary is short and the target is a small (7 or less) number of documents.

1 Introduction

Since a huge amount of documents are available in digital form, it is one of difficult problems for users to obtain really necessary information efficiently. In order to deal with such problems, many researchers are engaged in studies on Information Retrieval (IR) and Text Clustering. Automated text summarization is also one of the important topics as the method to reduce the amount of document to be read. Especially, summarization of multiple documents is widely noticed as one of ways to achieve more efficient and effective presentation of documents.

In multi-document summarization, it should be more realistic to summarize not a large set of heterogeneous documents directly but a set of loosely related documents. For example, let us consider the situation that a user uses a summarization system with a navigation interface for IR results like Scatter/Gatter[3]. The navigation system classify the retrieved documents into some adequate number of clusters. Then, the user select some topic-related clusters to narrow down the candidate documents. After some iterations of clustering and user's selection, the user may obtain a small number of documents to be read. If the summarization system can make a effective summary, which can be a substitute of the original documents, the user can save the time to read. In this paper, therefore, we will discuss a summarization system under the following conditions:

- Target is a set of documents that are originally retrieved by an IR system on a certain topic and narrowed down to some small number of related documents.
- The output is an informative summary, which can be a substitute of the original (target) documents.

In order to satisfy the second condition, the following points are important.

Comprehensibility The summary should include main content of target documents exhaustively.

Readability The summary should be a self-contained document and should be readable.

Many of recent researches put emphasis on the comprehensibility while keeping readability of summaries to some extent. For example, Radev et al.[12, 11] proposed a method that classifies given documents into some clusters and makes one sub-summary for each cluster, then places them in an order.

In the method of Stein et al.[13], first of all, one summary is made from each document. The set of such sub-summaries are classified into several clusters, and one typical summary is selected for each cluster as its summary. Then, the selected summaries are lined up.

Goldstein et al.[4] proposed the method called MMR-MD (Maximal Marginal Relevance – Multi-Document), which collects passages related to the query from newspaper articles retrieved by an IR system and arranges them into one summary. The main contribution of MMR-MD is a ranking mechanism for passages. The method considers not only the similarity between a passage and a query, but also the similarity between the passage and each of higher-ranked passages. If a passage has high similarity with some of passages already selected, it should be redundant. In such a case, some penalty is given to the passage and re-ranked. The method also makes use of clustering to make some clusters of related documents. The clusters of documents are used in ranking passages.

All of these methods described above utilize some (non-hierarchical) clustering mechanisms to find groups of similar documents. Since the main matter of concern is to extract common information in a cluster of documents and they treat each cluster separately, they seem to have the following problems:

- They cannot explicitly treat the important parts that come from difference among clusters.
- Their result of summary heavily depend on size of each cluster, because common parts of documents may vary according to cluster size.

We think that much contrivance is needed to make good summaries, which have not only the common information of all clusters but also the important information specific to each clusters. If we can obtain the detail of similarity structure of target documents, we may take account of not only the common part of clusters but also the difference among clusters. Such information would be a good clue to make a summary and improves the quality of extracting important parts of documents.

Besed on the consideration described above, in this paper, we will approach to the multi-document summarization in the following way. In the viewpoint of comprehensibility, we propose an integration of MMR into a term-weighting method proposed by Mori[7, 6]. Since the term-weighting method gives a weight to each word according to the word's contribution to determining the hierarchical structure of document clusters, it can reflect both commonality and difference among documents into weights of terms. In the case of summarizing *each* document in IR result *separately*, we need a term-weighting method and an improtant sentence selection only, because we do not have to care about controlling redundancy. Each document can be supposed to have no redundancy and to be a coherent text. Actually, Mori[7, 6] reported that the sentence extraction based on their term-weighing method works very well in such a situation.

We, however, have to consider the redundancy control in the case of multiple documents summarization, because contents of one document may overlap with other documents. Therefore, we propose an introduction of MMR into the framework by Mori[7, 6]. While MMR is originally designed to treat the relevance of passages to a query and redundancy among passages simultaneously, we modify it in order to deal with both *importance* and redundancy of *sentences*.

In the viewpoint of readability, we propose a method to generate a summary by clustering extracted sentences according to sub-topics and making a keyword list as a very brief summary for each cluster. Note that we aim at making not one very cohesive and readable text of summary but an informative text, from which users can obtain necessary information appropriately.

2 Issues on Multiple Document Summarization

In summarization of multiple documents, firstly, we have to take account of the following some extra points, which we do not consider in single document summarization.

Viewpoint 1 (Multi-document summarization)

1. *Finding important parts in each document.*
2. *Elimination of redundant parts of documents by detecting common parts.*
3. *Finding differences among documents and arranging them.*

Here, we suppose that target documents are adequately gathered through information retrieval process for a certain topic and user's selection in a navigation process.

Secondly, if the target documents are the result of information retrieval in a certain topic, we also have to consider a *query*, or information needs.

Viewpoint 2 (Query) *Query, which represents the user's information needs in the retrieval.*

Thirdly, we have to consider the characteristic of target documents. It may vary according to given query and document database in the IR system. If the targets are newspaper articles like NTCIR3 TSC2, there are, at least, two types of documents as follows:

Viewpoint 3 (Target set of documents)

1. *Set of documents that describes one specific topic in time series like articles about one crime.*
2. *Set of documents that describes one broad topic from several different views. Therefore, the set may consists of several sub-topics.*

If we adopt the important sentence extraction as summarization method, Viewpoint 1 is related to the selection process of important sentence, on the other hand, Viewpoint 3 has a relation to the arrangement process of selected important sentences in order to make one summary text. Viewpoint 2 may be in connection with both of those processes.

In this paper, as shown in Figure 1, we treat the above viewpoints with the following basis:

1. As for Viewpoint 1-1, 1-3, and Viewpoint 2, we treat them simultaneously with the method to map the similarity structure of documents into term weights, which is proposed by Mori[7, 6].
2. As for Viewpoint 1-2, we deal with it by a variant of MMR for single sentences.
3. As for Viewpoint 3, we realize it by classifying original documents into several clusters by single link clustering algorithm, and giving a keyword list to each cluster as a sub-title.

3 Important Sentence Extraction by Integration of MMR and term weighting based on Information Gain Ratio

In this section, we will propose an important sentence extraction method based on integration of the following schemes:

1. Calculation of sentence importance based on Information Gain Ratio
2. Control of redundancy in summaries by MMR.

This extraction method has the following features:

- Since similarity structure among documents is reflected into weights of terms, sentence extraction can be performed independently of cluster structure of documents. Therefore, the method does not have the restriction that the compression ratio must be specify for each cluster.
- Since a kind of bias by query is integrated into the process of mapping cluster structure to term weights,

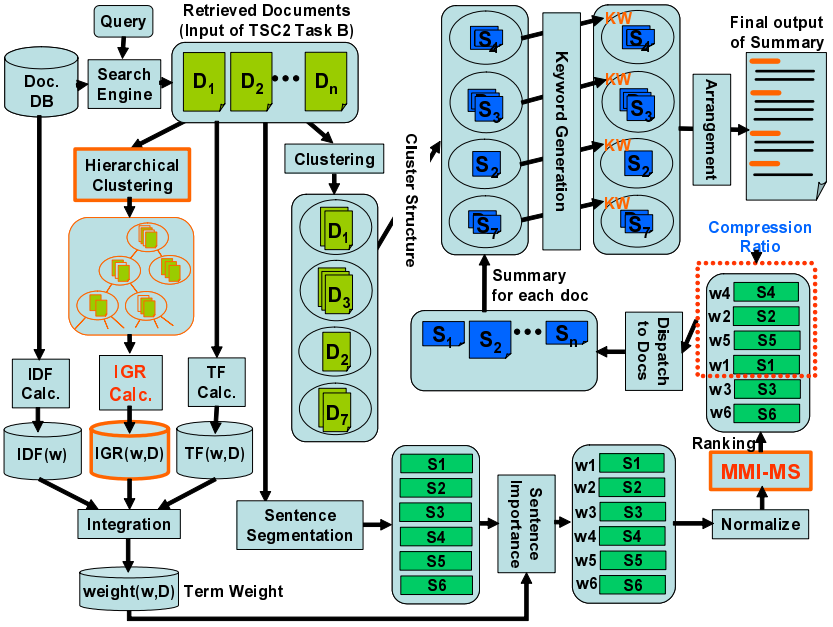

Figure 1. Overview of our proposed method for multi-document summarization

– the method can take account of topic of retrieval unlike Stein's method,

– and the method do not need any queries explicitly unlike MMR-MD. Moreover, the method has an effect of query expansion in term weighting, which MMR-MD does not have.

In order for MMR to be effective in redundancy elimination, in this section, especially we will discuss 1) the normalization of sentence weight and similarity score among sentences, and 2) choice of the value of the parameter λ, which controls the degree of redundance in MMR.

3.1 Term weighting method based on Information Gain Ratio

In order to summarize each of documents in an IR result, Mori[7, 6] proposed a method of important sentence extraction by term weighting based on Information Gain Ratio(IGR).

This method extracts the similarity structure among a set of documents through a hierarchical clustering, then gives higher weights to words that contribute to forming the structure. As an measure for the contribution, it utilizes IGR[10] of probabilistic distribution of words in a cluster. The IGR $gain\_r(w, C)$ of a word w in a cluster C is calculated as follows:

$$
\begin{aligned}
gain\_r(w,C) &= \frac{gain(w,C)}{split\_info(C)} \qquad (1)\\
gain(w,C) &= entropy(w,C) - entropy_p(w,C)\\
entropy(w,C) &= -p(w|C)\log_2 p(w|C)\\
&\quad -(1-p(w|C))\log_2(1-p(w|C))\\
p(w|C) &= freq(C,w)/|C|\\
entropy_p(w,C) &= \sum_i \frac{|C_i|}{|C|} entropy(w,C_i)
\end{aligned}
$$

$$
split\_info(C) = -\sum_i \frac{|C_i|}{|C|}\log\frac{|C_i|}{|C|}
$$

where $freq(w, C)$, C_i and $|C|$ are the frequency of the word w in C, the i-th sub-cluster of C, and the number of words in $|C|$, respectively. The value (1) represents the degree of consistency between the probabilistic distribution of a words and the partition of sub-clusters.

Here, we have to take account of the following points:

1. If the target of summarization is an IR result, it is important to compare the target documents with the set of documents which are *not* retrieved but exist in the document database, in order to obtain the information what words really have contribution in the retrieval. Therefore, as shown in Figure 2 we introduce another layer of cluster, which corresponds to the whole document database. The cluster consists of two sub-clusters. One sub-cluster is the cluster of retrieved documents, which is the target of further clustering, and the other one corresponds to the rest of database. Because of the contrast in the introduced cluster, higher weights are given to words specific to the retrieved documents. It may include the effect of query bias.

2. As shown in Figure 2, we obtain one IGR value for each word in each cluster. In order for every partition of cluster to be reflected in term weight, we have to integrate all of IGR values. In this paper, we adopt the integration IGR_sum given by the summation shown in (2):

$$
IGR\_sum(w,D) = \sum_{C\in Cset(D)} gain\_r(w,C), \qquad (2)
$$

where $Cset(D)$ is the set of all clusters to which the document D belongs.

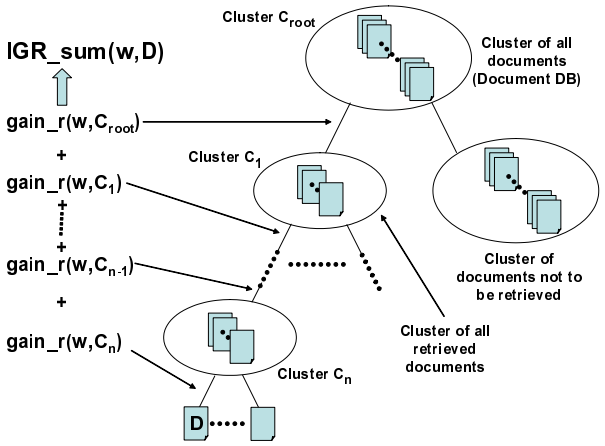

Figure 2. Term weight IGR_sum based on Information Gain Ratio

We define the final weight $weight(w, D)$ of the word w in the document D as the combination of the three types of fundamental weights, TF, IDF and IGR_sum. Since the weights TF, IDF and IGR_sum reflect the independent characteristics, each document, the document database, and the cluster structure respectively, we adopt the following integration:

$$weight(w, D) = IGR\_sum(w, D) \cdot TF(w, D) \cdot IDF(w) \quad (3)$$

We adopt the *maximum distance algorithm* as the clustering method for deriving IGR values. The algorithm has one parameter, α, which has the range [0.5,1] and controls the number of sub-clusters in a cluster implicitly. The smaller the value of the parameter is, the more clusters each cluster is divided into. In the experiment described below, we set the parameter to 0.8.

3.2 MMR

As described in Section 1, Goldstein et al. proposed a multiple document summarization method for IR results, called MMR-MD (Maximal Marginal Relevance – Multi-Document)[4]. To treat redundancy among retrieved passages, passages are re-ranked according to not only the relevance of passages to a query but also the similarity among passages. MMR-MD, therefore, can do both detection of shared parts of documents (reduction of redundancy) and extraction of different parts (improvement of comprehensibility), simultaneously. The redundancy control of MMR-MD is based on the notion of MMR(Maximal Marginal Relevance) proposed by Carbonell et al.[2]. As defined in (4), MMR selects the next passage in the set R of passages relevant to the query Q.

$$MMR(R, A) \stackrel{\text{def}}{=} Arg \max_{D_i \in R \setminus A} [\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in A} Sim_2(D_i, D_j)] \quad (4)$$

where A, Sim_1 and Sim_2 are the set of passages already selected, the similarity between a query and a passage, the similarity between two passages. After initializing S to the empty set and assigning an appropriate value to the parameter λ, we may have a ranking of passages under the control of redundancy by calculating (4) repeatedly.

In (4), the first term and the second term of the right hand side may be considered as the term for extracting the parts relevant to the query and the term for eliminating redundancy, respectively. The value of parameter λ coordinate those two effects.

3.3 MMI-MS

In this section, we will discuss about introduction of redundancy control mechanism of MMR into the important sentence extraction based on IGR. The unit of MMR is originally passage or document, and it adopted passages' relevance to a query as initial ranking Sim_1. Therefore, MMR may be incorporated into the scheme of important sentence extraction, if we change the unit from passage to sentence and use the importance of sentence as an initial ranking method. In this paper, we call the extended scheme scheme of important sentence extraction *MMI-MS* (Maximal Marginal Importance – Multi-Sentence) and define it as follows:

$$\textit{MMI-MS} \stackrel{\text{def}}{=} Arg \max_{S_i \in SS \setminus A} [\lambda Imp(S_i) - (1 - \lambda) \max_{S_j \in A} Sim_s(S_i, S_j)] \quad (5)$$

where SS, $Imp(S_i)$ and Sim_s are the set of all sentences in target documents, the importance of sentence S_i and a certain measure of similarity between sentences, respectively. As $Imp(S_i)$ and Sim_s, we adopt the sentence importance based on IGR and the cosine correlation between sentence vectors.

3.4 Normalization of sentence importance and similarity between sentences

Th ranges of $Imp(S_i)$ and Sim_s should be comparable in order for MMI-MS to work appropriately in elimination of redundancy. The range of Sim_s is $[0, 1]$. On the other hand, $Imp(S_i)$ does not have such a fixed range in our importance calculation. Therefore, we perform the following normalization of $Imp(S_i)$ and Sim_s.

1. Normalization of sentence importance in a document

   In document D, importance of sentence is converted to *deviation score* whose average is 0.5 as follows:

   $$Imp_s^d(S_i) = 0.5 + \frac{Imp_s(S_i) - Imp_s^{ave}(D)}{\sigma(D)} \quad (6)$$

   where $Imp_s^{ave}(D)$ and $\sigma(D)$ are the average and standard deviation of sentence importance in document D.

   This normalization equalize the averages of sentence importance in documents under the hypothesis that all documents is equally important.

2. Normalization of sentence importance and sentence similarity in all of documents

   This normalization is to make sure that MMI-MS takes effect. Firstly, $Imp_s(S_i)$ and $Sim_s(S_i, S_j)$ are divided by their maximum values, respectively. Then, they are converted to deviation scores whose average is 0.5.

3.5 Choice of the value of parameter λ in MMR-MS

The parameter λ has the range $[0, 1]$. The closer to 0 the parameter is, the more effective the elimination of redundancy is. Since the adequate value of λ may depend on the target set of documents, the value should be selected adequately.

For example, if the document set includes many different sub-topics and scarcely has overlaps, λ should be set to the value close to 1 to weaken the redundancy elimination. On the contrary, if the document set has many overlaps like articles about one specific topic, the value should be smaller value to promote the elimination of redundancy.

In this paper, we pay attention to the average of cosine correlation as the measure for redundancy among sentences. For example, the situation that the average has higher value shows the fact that there are many similar sentences. In such a situation , the parameter λ is set to a smaller value in order to promote elimination of redundancy by MMI-MS.

We adopt the equation (7) for λ, which has the value range $[0.5, 1]$, in order for the sentence importance to be more dominant than elimination of redundancy.

$$\lambda \quad = \quad 0.5 + 0.5(1 - Sim_s^{ave}) \qquad (7)$$

where Sim_s^{ave} is the average of $Sim_s(S_i, S_j)$.

4 Generation of Text of Summary

So far, we describe the method to extract important sentences from multiple documents. In this section, from the viewpoint of readability we will discuss how the extracted sentences should be arranged to be one result summary.

In generating summary, we have to pay attention not only naturalness as a text but also coherence among extracted sentences. Each sentence in a summary should have an adequate context in order for readers not to be misled. In the scheme of sentence extraction for summary generation, we need the following steps to make a good cohesive summary.

1. Extract important sentences with taking the coherence into account. For example, extraction unit may be a series of sentence connected to each other with reference or lexical cohesion. We simply call the unit of extraction *passage* hereafter. Paraphrasing reference expressions into their referent would also effective to make sentences self-contained.

2. Place extracted passages in order according to relations among passages like time order, topical coherence, and so forth.

3. Generate expressions to clarify relations between passages and insert them to the summary, if needed.

4. Finally, adjust the draft of summary by paraphrases or simplification to increase uniformity of expressions.

Mani et al.[5] proposed a method to improve informativeness and readability of original summary, which is represented as a list of syntactic trees with coreference information. Nanba et al.[8] discussed a method to generate summary easy to read by rewriting extracted important sentences. Otsuka et al.[9] improve the coherence among extracted sentences by replacing reference expressions with their antecedents.

In single document summarization, we may achieve the minimum requirement of cohesion in a summary by making an effort to maintain the cohesion which a target document originally has.

On the other hand, in multi-document summarization, we have to find cohesion across target documents, which do not appear explicitly in the documents. Moreover, the style of document would differ from each other. However, many of existing researches do not tackle these problems directly, but merely place sentences or passages in some order by using information in original documents, like date information, position in document, and so forth.

For example, methods proposed by Goldstein et al. [4], Stein et al. [13] and Radev et al. [12, 11] arrange the extracted units in time order of original documents.

Stein's method, moreover, reorder the units (each summary of single document) so as to increase similarity between adjacent units. These can be regarded as attempts to make cohesion among units in easy ways.

Note that the methods described above aim to summarize documents into one *continuous* reading. This type of summary would be suitable for documents of one specific topic like articles about one crime. In summarization of IR result, however, we usually obtain documents in several different points of view, namely sub-topics, even if the query is about one topic. We think that, in such a case, it would be preferable to show the outline of sub-topics explicitly and present sub-summaries for sub-topics along with the outline.

Barzilay et al.[1] reported the following findings by the experience in which subjects arrange the same set of extracted important sentences so as to maximize readability of text.

- Firstly, the total order of sentences depends on subjects, and there are several possibilities of order of sentences.

- The order, however, is not totally free, but several blocks of sentences can be found.

- Sentences belonging to same block are related to each other in terms of a sub-topic.

While the order of blocks and the order of sentences in each block may depend on each subject, there is general agreement among subjects that each block consists of sentences related to one sub-topic.

Based on above discussion, we propose a method of summary generation with following three steps according to characteristics of document set:

1. Make sub-summaries by classifying sentences according to sub-topics,
2. Generate a *sub-title*, or a list of keywords, for each sub-summary,
3. And arrange sub-summaries in order of date along with sub-titles.

4.1 Clustering important sentences according to sub-topics

As described in Viewpoint 3, there are, at least, two types of structure of sub-topics in a document set:

1. Set of documents that describes one specific topic in time series like articles about one crime.
2. Set of documents that describes one broad topic from several different views. Therefore, the set may consists of several sub-topics.

For a document set like 1, the strategy to arrange sentences in time series would be suitable. On the other hand, for a document set like 2, we need to adopt some strategy to classify sentences into sub-topics, and arrange them in time series, if necessary. We may deal with both of these cases by classifying the original documents into some clusters according to relevance to sub-topics.

Now, what kind of clustering algorithm is suitable for our purpose? The focus of topic may gradually shift according to the progress of incident, even if the target documents are in the case of 1, where they have to be classified into one cluster. In such a case, it is important for clustering algorithm to minimize not the size of each cluster but the distance between each document and its nearest document.

Therefore, we adopt the *single link method*, in which each document is classified into the cluster that has the closest document. As for the similarity (or distance) between documents, we use the similarity of original documents. Similarity of sub-summaries of documents is not suitable for the clustering, because in this stage our method has already eliminated the redundancy in summaries.

Based on the above discussion, we will make clusters of extracted sentences and line them up by the following process:

1. Obtain the summary of each document by gathering extracted important sentences document by document. The summary of each document can be regarded as *sub-summary* and is the unit of the following processing.
2. Make clusters of sub-summaries according to the similarity among original documents ($cosine$ values of document vectors), where the clustering algorithm is the single link method, and the threshold similarity for merging is 0.5.
3. Let the *cluster date* be the date of the oldest document. Line up clusters in time order of cluster date.
4. In each cluster, line up sub-summaries in time order according to document date.

4.2 Keyword generation for brief summaries of clusters

The very brief summary for each cluster of sub-summary may useful for readers to understand the outline of the whole summary. These short summaries work as separator of sub-topics, and may contribute for improving the readability.

As brief summary of each cluster, in this paper, we use keywords extracted from each cluster by the following process.

1. Select the most important word in each document. The most important word is defined as the word of highest weight in the most important sentence. If the most important word is the part of longer compound word, the longest compound word is selected as keyword of document. Otherwise, the most important word is the keyword of document.
2. The brief summary of cluster is made as a set of keywords of documents that belongs to the cluster.

The reason to select compound words in keyword selection is that such compound word may carry more specific information. In the summarization, we display the keywords at the beginning of sub-summary of each cluster.

5 Experiment

We participated in NTCIR3 TSC2 and evaluated our proposal with the Task B[14]. Each participant generated summaries according to topic information given by the task organizers, and submitted a set of summaries to the task organizers. Each topic corresponds to one IR result, which consists of the following information:

- Topic ID
- List of keywords for query in IR.
- Brief description of the information needs.
- Set of document IDs, which are target documents of summarization. The number of documents varies from 3 to 17 according to topic.
- Length of summary to be generated. There are two lengths of summary, 'Long' and 'Short'. The length is given in character and the character 'Newline' is not included in the length. While the length may varies according to the topic, the length of 'Long' is twice of 'Short'.

To each participant, the task organizers supplied the information about system evaluation. In the rest of this section, we will describe the evaluative information.

5.1 Evaluation by Ranking

The task organizer prepared the following four summaries for each topic and each length:

1. Summary made by a human ('*Manual*' hereafter).

2. Summary generated by a system to be evaluated.
3. Summary generated by Lead method (Baseline No.1. '*Lead*' hereafter).
4. Summary generated by Stein's method (Baseline No.2. '*Stein*' hereafter).

Next, one of twenty human assessors reads the original set of documents and four summaries for each topic and each summary length, then ranks the four summaries in the following evaluation points of view:

Comprehensibility The summary has all of important contents, or not. ('*C*' hereafter)

Readability The summary is easy to read, or not ('*R*' hereafter)

Therefore, every system will have four types of ranking, namely, *C Short*, *R Short*, *C Long*, and *R Long*, for each topic. Note that smaller value means the better system in this evaluation.

5.2 Evaluation by Human's Correction

In this evaluation, one of human assessors corrects summaries of system outputs in terms of the evaluation viewpoints described above. The each step of correction should be one of three operations, namely, insertion, deletion, substitution of strings. If more than 50% of summary should be correct, assessors may give it up.

6 Experimental Results

6.1 Result of Evaluation by Ranking

Table 1 and 2 show the evaluation by average rank. Note that, strictly speaking, the *average* of rank is *not* a value suitable for comparison of systems. The ranking is not an interval scale but an ordinal scale. In other words, the number of position and the difference of position in ranking do not have special meaning, while the order of items is meaningful. Thus, we should consider the average of rank as not a strict measure but a lax measure. Therefore, we compared our system with each of baselines topic by topic. The result is shown in Table 3.

We also evaluate only the topics that have 7 documents or less, because by examination of each topic we found out that our method works well for smaller document sets. The result is shown in Table 4 and 5.

6.2 Result of Evaluation by Human's Correction

Table 6 shows the average amount of human's correction.

7 Discussion

7.1 Comprehensibility

According to experimental results, proposed method is superior to the baselines, namely the lead method and Stein's method, in comprehensibility, especially under the following situations:

Table 1. Evaluation by Ranking for Proposed method (All of 30 topics)

	C Short	R Short	C Long	R Long
Average Ranking(AR)	2.53	3.10	2.73	3.30
Ranking of AR	2	8	5	8

Average Rank: Average ranking (AR) of our system compared with three baselines.
Ranking of AR: Ranking of AR in all nine participating systems

Table 2. Average Ranking compared with Baselines (All of 30 topics)

	C Short	R Short	C Long	R Long
Proposed	2.53	3.10	2.73	3.30
Manual	1.70	2.2	1.77	2.13
Lead	2.97	1.97	3.00	2.63
Stein	2.43	1.97	2.37	1.67

Table 3. Number of Win and Lose compared with Baselines (All of 30 topics)

	C Short			R Short		
	W	L	T	W	L	T
v.s. Manual	7	22	1	8	20	2
v.s. Lead	**15**	**12**	**3**	6	21	3
v.s. Stein	**16**	**12**	**2**	7	22	1
	C Long			R Long		
	W	L	T	W	L	T
v.s. Manual	9	21	0	7	21	2
v.s. Lead	**17**	**13**	**0**	10	20	0
v.s. Stein	11	18	1	3	27	0

W: win, L: lose, T: tie

Table 4. Average Ranking compared with Baselines (15 Topics with 7 documents or less only)

	C Short	R Short	C Long	R Long
Proposed	1.93	2.60	2.13	3.00
Manual	1.87	2.33	2.00	2.06
Lead	3.13	2.20	2.80	2.87
Stein	2.87	2.33	2.93	1.87

Table 5. Number of Win and Lose compared with Baselines (15 Topics with 7 documents or less only)

	C Short			R Short		
	W	L	T	W	L	T
v.s. Manual	6	9	0	6	9	0
v.s. Lead	**11**	**4**	**0**	6	7	0
v.s. Stein	**13**	**1**	**1**	7	8	0
	C Long			R Long		
	W	L	T	W	L	T
v.s. Manual	7	8	0	5	10	0
v.s. Lead	**10**	**5**	**0**	7	8	0
v.s. Stein	**10**	**4**	**1**	3	12	0

W: win, L: lose, T: tie

Table 6. Average amount of human's correction (in character per document)

	Short		Long	
	Proposed	Average	Proposed	Average
Deletion				
C(%)	24.4	18.6	21.5	16.9
R(%)	0.4	1.6	0.9	1.7
Insertion				
C(%)	16.2	25.9	18.1	19.5
R(%)	1.4	1.0	0.5	1.3
Substitution				
C(Del)(%)	4.0	2.0	2.9	1.5
C(Ins)(%)	9.1	2.9	4.6	2.0
R(Del)(%)	1.0	0.6	0.5	0.5
R(Ins)(%)	1.0	1.2	0.4	0.3

1. The output are short summaries.
2. The target is a small number of documents.

The result would shows that MMR is effectively integrated into the term-weighting method based on IGR.

On the other hand, our method is not rated highly in the other situations. One of reasons may be the incompleteness of automatic adjustment for parameter λ, which controls redundancy of summary. As shown in (7), λ has the bias 0.5, and we put more stress on important sentences selection. That is, MMI-MS tends to be suppressed. When the number of target documents becomes larger, the method tends to select important sentences from every document evenly if the effect of MMR is weak.

7.2 Readability

According to human's correction, the total amount of correction of our summaries is larger than outputs of other systems. One of main reasons is the fact that our system is based on important sentence extraction, and does not use any other methods such as a technique to increase cohesion among sentences, rewriting to shorten sentences.

We also have to note that the viewpoint of readability of TSC2 task organizers is different from ours. The task organizers evaluated readability of summaries *in terms of single document*. However, we aim at *readability in terms of information access*, and we do not stick to make a single document. Actually, almost all of subtitles in our summaries are deleted in human's correction.

8 Conclusion

In this paper, we proposed a method to make a summary from multiple documents with taking account of comprehensibility and readability. As for comprehensibility, we showed an integration of MMR into the term-weighting method based on IGR. As for readability, we propose a method to generate a summary based on clustering important sentences according to sub-topics and making a keyword list as a very brief summary for each cluster.

The evaluation in NTCIR3 TSC2 showed that the proposed method works well to generate comprehensive summaries when the length of summary is short and the target is a small (7 or less) number of documents.

In our future works, we will consider the improvement of summary generation including improvement of readability, treatment of cohesion, paraphrases.

References

[1] R. Barzilay, N. Elhadad, and K. R. MaKeown. Sentence ordering in multidocument summarization. In *Proceedings of the the first International Conference on Human Language Technology Research (HLT 2001)*, pages 149–155, 2001.

[2] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[3] D. R. Cutting, D. R. Karger, and J. O. P. J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR '92: 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[4] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-Document Summarization by Sentence Extraction. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, pages 40–48, 2000.

[5] I. Mani, B. Gates, and E. Bloedorn. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL 99)*, pages 558–565, 1999.

[6] T. Mori. Information gain ratio as term weight — the case of summarization of ir results —. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 02)*, pages 688–694, Aug. 2002.

[7] T. Mori. A term weighting method based on information gain ratio for summarizing documents retrieved by ir systems. *Journal of Natural Language Processing*, 9(4):3–32, 7 月 2002. (in Japanese).

[8] H. Nanba and M. Okumura. Producing more readable extracts by revising them. SIG Notes NL-133-8, Information Processing Society of Japan, 1999. (In Japanese).

[9] T. Otsuka, A. Utsumi, and K. Hirota. Youyaku-bunseisei-ni-okeru shouou-shori (anaphora resolution in summary generation). In *Proceedings of seventh annual meeting of the Association for Natural Language Processing (NLP 2001)*, pages 425–428, Mar. 2001. (In Japanese).

[10] J. R. Quinlan. *C4.5: progarms for machine learning*. Morgan Kaufmann Publishers, May 1993.

[11] D. R. Radev and W. Fan. Automatic summarization of search engine hit lists. In *Proceedings of ACL Workshop on Recent Advances in NLP and IR*, 2000.

[12] D. R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, 2000.

[13] G. C. Stein, T. Strazalkowski, and G. B. Wise. Summarizing Multiple Documents using Text Extraction and Interactive Clustering. In *Proceedings of the sixth Pacific Association for Computational Linguistics (PACLING 99)*, pages 200–208, 1999.

[14] TSC Committee. NTCIR 3 automatic text summarization task/TSC 2(text summarization challenge 2) web page. http://lr-www.pi.titech.ac.jp/tsc/tsc2-en.html, 2001.