Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine

Tatsunori MORI and Masanori NOZAWA and Yoshiaki ASADA
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{mori,nozawa,asada}@forest.eis.ynu.ac.jp

Abstract

Recent years, the answer-focused summarization is paid attention to as a technology complementary to information retrieval and question answering. In order to realize multi-document summarization focused by multiple questions, we propose a method to calculate sentence importance using scores produced by a Question-Answering engine in response to multiple questions. We also describe an integration of it into a generic multi-document summarization system. The evaluation results show that the proposed method has better performance than not only several baselines but also other participants' systems in the evaluation workshop NTCIR4 TSC3 Formal Run, although we have to take notice of the fact that some of the other systems do not use the information of questions.

1 Introduction

Recently, a large amount of documents is available in many ways. It, however, is not easy to find needed information efficiently among them. Although we come to be able to obtain a set of relevant documents or answers using technologies of information retrieval (IR) and question answering(QA), it is still necessary to examine the original documents.

One of the complementary technologies is the automatic multi-document summarization for documents retrieved by IR systems. Recent years, especially *Answer-Focused Summarization* is paid attention to(Hirao et al., 2001; Wu et al., 2002). This is based on an empirical viewpoint that information need of a user can be described as a set of questions. One of the four tasks in DUC 2003 organized by NIST was to produce summaries of multiple documents in response to a single question(Over and Yen, 2003). Since, in multi-document summarization, some moderate amount of text is necessary so as for users to be able to understand the content, the amount of documents which users should finally read would not be small when a separate summary is made, one by one, for each matter which users want to know.

Based on the discussion described above, in order to realize multi-document summarization focused by multiple questions, we propose a method to calculate sentence importance using scores produced by a QA engine in response to multiple questions. We also describe an integration of it into a generic multi-document summarization system.

2 Overview of the Proposed Method

In this paper, we assumed that a set of documents to be summarized is given as a result of IR, and a set of questions, which corresponds to information need by a user, is also provided. As for questions, it is a more natural situation where a user gives questions one by one in the interaction with a system, and according to the input the system gradually outputs a part of summary including an answer by taking account of relations to previously displayed parts of summary. In this paper, however, we suppose that questions are given all at once, as a primary approximation. Under the situation, 1) extraction of important part according to information need, 2) reduction of redundancy in documents, and 3) detection of difference among documents, are necessary for multi-document summarization. For those necessary functions, we adopt the following techniques.

(a) Calculation of sentence importance based on scores of a QA Engine, which corresponds to 1). (See Section 3.)

(b) Calculation of sentence importance based on Information Gain Ratio (IGR) with respect to probabilistic distribution of words, which corresponds to 1) and 3). (See Section 4.)

(c) Control of redundancy in summary text based on Maximal Marginal Relevance (MMR), which corresponds to 2) and 3). (See Section 6.)

We also introduce (d) smoothing of sentence importance by a Hanning window function(See Section 5). Figure 1 illustrates the overview of the multi-document summarizer for Japanese documents based on the proposed method. The input to the system consists of a set of document IDs, a set of questions, which represents user's information need, and the summary length. The output of the system is an extract, namely a series of sentences extracted from documents. For example, in the case of the topic 0500 (Articles about clone sheep "Dolly") in the

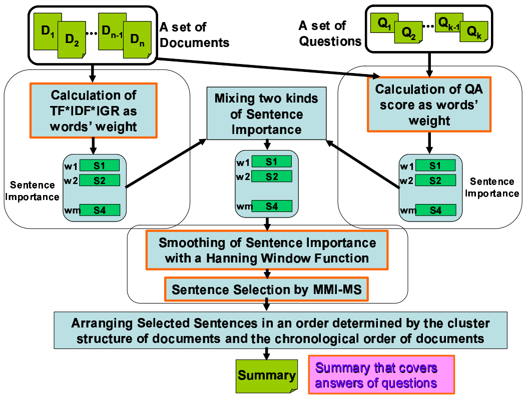


Figure 1: Multi-Document Summarization using a QA Engine

test collection of NTCIR4 TSC3[1], the input includes the nine document IDs in Figure 2, the ten questions in Figure 3, and the number '491' for the summary length in character. When the input is given to the system, it firstly computes the importance score for each sentence by integrating two kinds of importance: (a) sentence importance based on scores of a QA Engine, and (b) sentence importance based on IGR with respect to probabilistic distribution of words. Secondly, the sentence importance is smoothed by a Hanning window function to maintain cohesion in an output summary. Using MMR, thirdly, it re-orders sentences so as to reduce the redundancy in summary text while taking account of the sentence importance. Fourthly, it selects the top-n sentences. Finally it orders the selected sentences according to the cluster structure of original documents and the chronological order of documents, and outputs the ordered sentences as a summary. Figure 4 shows an example output summary, in which each phrase in italic face corresponds to one of the answers to the questions in Figure 3.

3 Calculation of Sentence Importance using a QA engine

In order to deal with given questions, we adopt a Japanese QA engine proposed by Mori et al.(Mori et al., 2003). It does not require any preprocessing on documents, and users can therefore use various search engines by making simple wrappers. Since it performs computationally expensive processing including Japanese morphological analysis[2], parsing, extraction of Named Entities, and so on, after a question is submitted, they introduce a search mechanism for finding answers controlled by the A* algorithm and propose a method to approximately estimate words' score with some less expensive processing. The mechanism reduces the computational cost on irrelevant parts of documents and real-time processing is achieved.

[1]See Section 8.

[2]The morphological analysis for Japanese sentences consists of the word segmentation and the POS tagging.

The QA engine gives each word in documents a score which represents the appropriateness of the word in terms of the answer to a question. Under the assumption that a word under consideration is the answer, namely, the word is tied up to the interrogative in the question sentence, the score is calculated as the degree of matching between the rest part of the question and the rest part of the sentence in which the word appears. The measure of matching they proposed is the linear combination of the following measures: the number of shared character bigrams, the number of shared words, the degree of case matching, the degree of matching between dependency structures, and the degree of matching between the type of NE in sentences and the type of the question. We propose the utilization of the QA score as the weight of word and the calculation of sentence importance from those scores. We expect that the proposed method will achieve better performance in summary generation than other existing methods, which usually use the frequency of words in questions, the question types, and/or the frequency of NEs in documents.

In this paper, we assume that a set of questions are supplied to the system. Consequently, not one QA score but a tuple of QA scores is assigned to each word. Each of QA scores in the tuple is linked up with one of questions. The range of QA score depends on the complexity of question sentence and the type of question. Thus it is originally meaningless to compare scores for different questions. We however want to assign a single unified value of importance to each word with respect to a set of questions. We therefore normalize the range of QA scores to make them comparable. In this paper, we adopt the T-score in Formula (1) as a normalization method to examine the deviation from the average of score of words because, in answering a question, it is important not to obtain absolute values of words' score, but to find the relative order among them:

$$T(x, D) = \frac{0.1 * (x - average(D))}{standard\_deviation(D)} + 0.5 \qquad (1)$$

where x is a score value to be normalized, D is a set of score values for a question, which includes x. Let $score^n(w, q)$ be the normalized score of the word w with respect to the question q, then the importance value $Imp^n_{QA}(S_i)$ of the sentence S_i is calculated as follows:

$$Imp^n_{QA}(S_i) = \max_{w \in W_{S_i}, q \in Q} score^n(w, q) \qquad (2)$$

where Q is the set of given questions and W_{S_i} is the set of words appearing in the sentence S_i. Since the sentence importance is depends on whether the sentence includes at least one of answers, we adopt the maximum function to integrate normalized scores.

JY-19990402J1TYEUG0400060, JY-19990527J1TYMAJ1400040, JY-19980424J1TYMAK1400070, JY-19980723J1TYMAJ1400050, JY-19980301J1TYMAP1400050, 980110135, 980723029, 980424152, 980215018

Figure 2: Example of input (1) : a set of document IDs (NTCIR4 TSC3 Topic 0500)

What is “Dolly”? / Where was Dolly born? / What kind of clone is Dolly? / What is the origin of Dolly? / What is identical between Dolly’s cell and the cell of the ewe that is the origin of Dolly?/ Who is the director of the Rosslyn Institute, UK?/ What kind of criticism has been presented about the fact that Dolly’s origin is the mammary-gland cell gathered from a sheep under pregnancy?/ How long is Dolly’s longevity in comparison with ordinary sheep?/ When did Dolly give birth?/ What is confirmed by Dolly’s childbirth?

Figure 3: Example of input (2) : a set of questions (English translation of NTCIR4 TSC3 Topic 0500)

According to the Evening Standard on the eighth, *the Rosslyn Institute*, U.K. announced that they made the world’s first *cloned sheep* “Dolly”, which is female and was born by putting a soma into an ovum, copulate with a male sheep. The Rosslyn Institute in Edinburgh, U.K., which succeeded in making the world’s first cloned sheep Dolly in the year before last, announced on *the 23rd* that Dolly gave birth recently and it was consequently confirmed that she had *a normal fertility*. Details will be published in British science magazine “Nature” of the issue on the 23rd. Dolly was made from *a mammary gland cell picked from a pregnant sheep*. Although this ewe is already dead, a part of the tissue is cryopreserved in U.K. The research group of Dr. *Ian Wilmut* at the Rosslyn Institute, U.K. and the research group at University of Leicester, U.K. separately confirmed by *the DNA test* that Dolly was a clone by the soma nucleus transplant of *an adult ewe*, and published it in British science magazine“Nature” of the issue on the 23rd. Two male and one female were bone, and all of them were spry. The team of the Rosslyn Institute, U.K., which made Dolly born, will publish it in British science magazine “Nature” of the issue on the 27th. She has already have got pregnant two times, and has given birth to spry children.

Figure 4: Example of output : a summary (extract) of documents (English translation)

4 Sentence Importance based on Information Gain Ratio

Mori et al.(Mori, 2002) proposed a method to calculate sentence importance using a word weighting technique based on IGR. It extracts the similarity information among given documents by a hierarchical clustering and assigns a importance value to each word according to whether the probabilistic distribution of the word is consistent with the cluster structure of documents. We adapt the method in order to obtain the sentence importance with respect to given documents. Let C_i be the i-th sub-cluster of cluster C, then the IGR of the probabilistic distribution of the word w in the cluster C is defined as follows.

$$
\begin{aligned}
IGR(w,C) &= \frac{info(w,C) - info_{div}(w,C)}{split\_info(C)} \\
info(w,C) &= -p(w|C)\log_2 p(w|C) \\
&\quad - (1-p(w|C))\log_2(1-p(w|C)) \\
info_{div}(w,C) &= \sum_i \frac{|C_i|}{|C|} info(w,C_i) \\
split\_info(C) &= -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|}
\end{aligned}
$$

Here, note the following points.

1. When the set of documents to be summarized is a result of information retrieval, the contrast between it and the set of unretrieved documents conveys important information in terms of word weighting. As shown at the top of Figure 5, we introduce another cluster above the cluster of retrieved documents. In the cluster, words that are relevant not to unretrieved documents but to retrieved documents have higher weights.
2. We obtain one weight value for each word in each cluster. On the other hand, one document may belong to multiple clusters in hierarchical clustering. We therefore have to integrate a set of weights for a word into a single value. We adopt the average of IGR values appearing on the path from the root cluster to a document and denote it as $IGR\_ave(w,D)$ for the word w in the document D.

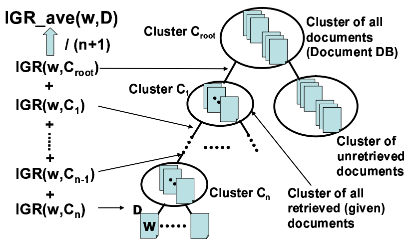


Figure 5: Word Weight $IGR\_ave(w,D)$ based on Information Gain Ratio

The weight of word described above is combined with the other existing word weights, i.e. TF and IDF. We define the importance $Imp_{IGR}(S_i)$ of the sentence S_i as the average weight of nominals in the sentence as Formula (3). The importance value is normalized across a document by T-score and is denoted as $Imp^n_{IGR}(S_i)$. Finally the integrated importance $Imp^n(S_i)$ of the sentence S_i is defined as the linear combination of $Imp^n_{IGR}(S_i)$ and $Imp^n_{QA}(S_i)$ in Formula (4), where α is the mixing factor.

$$Imp_{IGR}(S_i) = \frac{\sum_{w \in Noun(S_i)} \mathrm{TF}(w, D) \cdot \mathrm{IDF}(w) \cdot \mathrm{IGR}\_ave(w, D)}{|Noun(S_i)|} \quad (3)$$

$$Imp^n(S_i) = \alpha \cdot Imp^n_{QA}(S_i) + (1 - \alpha) \cdot Imp^n_{IGR}(S_i) \quad (4)$$

5 Smoothing of sentence importance by a Hanning window function

In the method described so far, each sentence is processed separately. When, however, a relatively large number of documents are given, there seems to be less cohesion between sentences in the generated summary, because the method tends to equally choose a small number of important sentences from all of documents. If we generate a long summary, it is necessary to improve the cohesion between sentences while taking account of importance of sentences. We therefore introduce a method to smooth importance using a Hanning window function. The sentence importance smoothed with the function of the window size W is defined as follows.

$$Imp^n_c(S_i) = \sum_{j=i-\frac{W}{2}}^{i+\frac{W}{2}} \frac{1 + cos 2\pi \frac{j-i}{W}}{2} \cdot Imp^n(S_j) \quad (5)$$

In typical case, when a sentence with moderate importance is placed between two very important sentences, the weight of the sentence is increased by the window function. As the result, the series of three sentences is likely to be selected as a part of summary. In this case, the adoption of the second sentence may improve the cohesion between those two important sentences.

6 Control of Redundancy in Summary based on MMR

We introduce a redundancy control mechanism based on Maximal Marginal Relevance (MMR) proposed by Carbonell et. al. (Carbonell and Goldstein, 1998). MMR is originally a method to reorder documents or passages by taking account of both redundancy and relevance to a query. As shown in Formula (6), it is applicable to choosing important sentences with considering redundancy, when the unit of selection is changed to a sentence and the initial order of sentences is given by sentence importance:

$$\mathrm{MMI\text{-}MS}(SS, A) \stackrel{\mathrm{def}}{=} Arg \max_{S_i \in SS \setminus A} [\lambda Imp^n_c(S_i) - (1 - \lambda) \max_{S_j \in A} Sim_s(S_i, S_j)] \quad (6)$$

where SS is the set of sentences to be summarized, A is the set of sentences already selected in the summary, Sim_s is the similarity between two sentences, and λ is the parameter to control the degree of redundancy. In this paper, we call it MMI-MS (Maximal Marginal Importance – Multi-Sentence). When we iteratively apply Formula (6) to SS after assigning an empty set to A, we can obtain an ordered list of sentences by taking account of both the importance and the redundancy. As for Sim_s, we adopt a simple cosine similarity between sentence vectors. Each element of a sentence vector represents an weight of a noun in a document.

7 Generation of Summary

Since we, in this paper, focused on the extraction of important sentences, the phase of summary generation is extraction-based and quite simple. It selects the top most important sentences until the total length/number of selected sentences reaches a given summary length.

The selected sentences are arranged in an order determined by the cluster structure and the chronological order of input documents as follows. Firstly, the input documents are non-hierarchically clustered using the single link clustering method. Then, the obtained clusters and the documents in each cluster are arranged in the chronological order. Here, the date of cluster is defined as the date of the oldest document in it.

8 Experimental Evaluation

We evaluate the proposed method with NTCIR4 TSC3 formal run. NTCIR TSC is a series of evaluation workshops of text summarization organized by National Institute of Informatics, Japan(Fukusima and Okumura, 2001). The latest workshop, or NTCIR4 TSC3, was just held on June, 2004(Hirao et al., 2004). In this paper, we evaluate the proposed method from the following points of view: 1) the sentence precision and coverage of the system's extracts with respect to the model extracts, and 2) the answer coverage of the system's summaries with respect to the model abstracts. The model extracts and the model abstracts were prepared by the task organizers.

The evaluation set of the formal run has 30 topics. Each topic information consists of a list of document IDs of Japanese newspaper articles to be summarized, a title of topic, two types ('Short' and 'Long') of length of summary, and two sets of questions (for Short and for Long) that represent the items to be included in the summary. The compression ratio for 'Short' and 'Long' is about 5% and 10%, respectively.

The parameters of the system are manually tuned to fit the set of five example topics which was supplied by the task organizers before the formal run. While we do not apply the smoothing by a Hanning window function to the generation of 'Short' summaries, the width of the

Hanning window is four for 'Long'. The mixing factor α is set to 0.8 (for 'Sort') or 0.7 (for 'Long'). The parameter λ of MMI-MS is set to $0.4 + 0.5 \cdot (1 - Sim_{ave})$, where Sim_{ave} is the average similarity among sentences in a topic.

8.1 Performance of Important Sentence Extraction

In this experiment, we use the model extracts that are prepared by the task organizers according to the corresponding model abstracts. Each model abstract is created by one of five specialists, who are ex-newspaper journalists.

Figure 6 (a) and (b) show the average sentence coverage and the average sentence precision of the extracts generated by the proposed system. In these figures, the label 'IGR+MMR+QA' represents the proposed method. The labels 'IGR+MMR', 'IGR+MMR+QB' and 'IGR+MMR+QB+NE' correspond to three baseline methods we prepared. IGR+MMR is the same method as IGR+MMR+QA except for suppressing the function of the QA engine. IGR+MMR+QB is a query-biased method in which the sentence importance $Imp^n_{QB}(S_i)$ is used instead of Formula (2). $Imp^n_{QB}(S_i)$ is calculated by normalizing the importance in Formula (7) with T-score. IGR+MMR+QB+NE is the same method as IGR+MMR+QB except for adding extra weight to sentences which include named entities (ENAMEX, TIMEX and NUMEX), as shown in Formula (8). The major difference between the proposed system and those baselines is whether to use the QA engine, namely information about answers, or not.

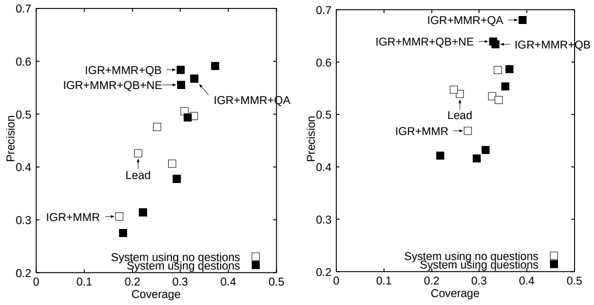
$$Imp_{QB}(S_i) = \sum_{w \in noun(S_i) \cap noun(Q)} \text{TF}(w, S_i) * (1.0 + log_2 \text{TF}(w, Q)) \quad (7)$$

$$Imp_{QB+NE}(S_i) = Imp_{QB}(S_i) + 2.0 * |NE(S_i)| \quad (8)$$

On the other hand, the label 'Lead' represents the lead method, which selects the beginning of each document. It is a baseline method provided by the task organizers. Other plots correspond to other participants' systems. Note that it is left to participants to decide whether to use the questions in topic information for generating summaries. Thus, there may exist systems which do not use the questions. It also has to be noted that there are no manually-created extracts supplied by the task organizers other than the model extracts, unlike the evaluation in the following section.

8.2 Performance in terms of Coverage of Answers

Figure 7 (a) and (b) show the average answer coverage, which represents how many answers in model abstracts are included by summaries generated by a system. Task organizers provided two kinds of index of answer coverage; 1)



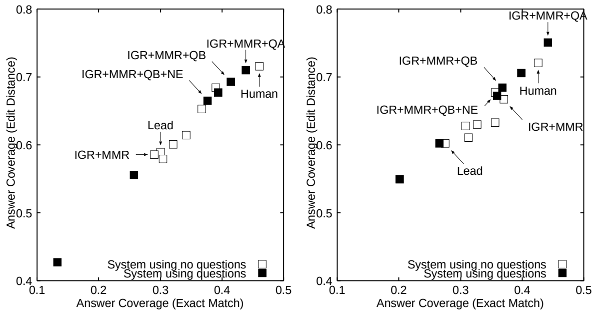
(a) Short (b) Long

Figure 6: Average Coverage and Average Precision of Sentence Extraction

Exact Match: the average ratio that the summaries include exact answer strings in model abstracts, 2) *Edit Distance*: the average score that is defined based on the edit distance $EditD()$ between an answer string Ans_i and a sentence string S as Formula (9):

$$Cov_{ED}(Ans_i) = \max_{S} \frac{Len(S) - EditD(S, Ans_i)}{Len(Ans_i)} \quad (9)$$

where the function $Len()$ returns the length of a string. The label 'Human' in the figures corresponds to a set of summaries, each of which is created by one of five specialists who does not create the corresponding model abstract.



(a) Short (b) Long

Figure 7: Average Answer Coverage for Questions

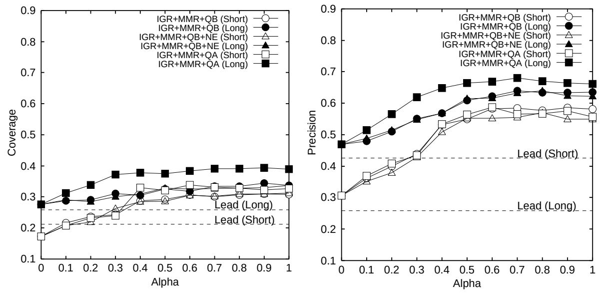
8.3 Mixing Factor of two kinds of Sentence Importance

We conducted the same experiments as the preceding sections except varying in the value of the parameter α from 0.0 to 1.0. Figure 8 and 9 show the performance of sentence extraction and the average answer coverage for questions, respectively.

9 Discussion

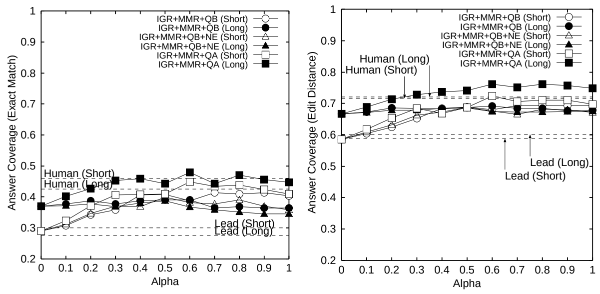
9.1 Performance of Important Sentence Extraction

As shown in Figure 6 (a), when the length of summary is 'Short', the performance of the proposed method (IGR+MMR+QA) is almost same as the baselines IGR+MMR+QB and



(a) Average Coverage (b) Average Precision

Figure 8: Performance of Sentence Extraction with the Mixing Factor α Varying from 0.0 to 1.0



(a) Exact Match (b) Edit Distance

Figure 9: Average Answer Coverage with the Mixing Factor α Varying from 0.0 to 1.0

IGR+MMR+QB+NE, while IGR+MMR+QA outperforms the lead method. It might be concluded that, for short summaries, words in questions conveys enough information for summarization. On the other hand, in the case of 'Long', the proposed method is predominance over all of baselines and other participating systems as shown in Figure 6 (b), although we have to take notice of the fact that some of the other systems do not use the information of questions. In comparison with IGR+MMR, the improvement of the proposed method is remarkable, and it shows that the sentence weight by the QA engine works effectively.

It is also remarkable that the baseline IGR+MMR+QB shows comparatively better performance than other baselines and other participating systems. The reason may be that relatively many questions are available in the evaluation set. On the other hand, IGR+MMR+QB+NE is rather worse than IGR+MMR+QB. Since there are many questions in one topic, we do not select named entities according to question types, and utilize all of named entities for sentence weighting. It therefore may not be effective.

While, as for 'Long', the average precision of IGR+MMR+QA is quite high (0.680), the average coverage is relatively low (0.391). The reason is that the proposed method tends to extract sentences which have the same content as the sentences already selected. Figure 10 shows the average number of almost identical sentences in a summary. It is a part of subjective assessment for summaries. According to the

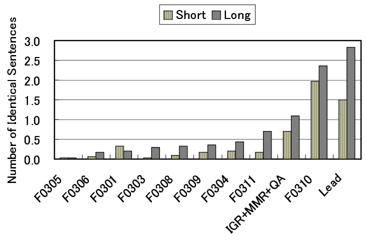


Figure 10: Average Number of Identical Sentences in a Summary

figure, the proposed method sometimes missed eliminating obviously redundant sentences. We think that one of reasons is the way to measure sentence similarity. In MMI-MS, we adopt the cosine measure between sentence vectors, each element of which is a weight of noun. Since a noun may have different weight values in different documents, the similarity between two sentences in different documents may be less than 1 even if they are identical. It is necessary to refine the calculation method of similarity.

9.2 Performance in terms of Coverage of Answers

According to Figure 7, the performance of the proposed method is better than that of baselines for both 'Short' and 'Long'. Note that the set of summaries labeled by 'HUMAN' is created by the five specialists without referring to questions in the topic information.

9.3 Effect of Mixing two kinds of Sentence Importance

Figure 8 and 9 show that the sentence importance derived from questions, like questions themselves and their answers, is predominant over the sentence importance based on IGR. We, however, can also find that there is a peak in each of curves at $\alpha = 0.6 \sim 0.8$. Thus we may conclude that mixing two kinds of importance takes effects to some degree. Especially, it is interesting that the peaks are not located at the point $\alpha = 1.0$ even though in the answer coverage curves. One of the reasons of it would be that the QA system we employ is insufficient in accuracy. The MRR[3] of the QA system is about 0.5 for the test sets of NTCIR QAC1 and QAC2[4]. The sentence importance based on the probabilistic distribution of words can, therefore, compensate for the inaccuracy.

10 Related Work

The proposed method is related with two kinds of summarization researches. One is the multi-document summarization, and the other is the question-focused summarization.

As for multi-document summarization, we are using several existing methods. We employ the word weighting method based on

[3] Mean Reciprocal Rank.

[4] QAC is the series of evaluation workshops concerning QA technologies in NTCIR.

IGR(Mori, 2002) and MMR for redundancy control(Carbonell and Goldstein, 1998). Hirao et al.(Hirao et al., 2001) inspired us to utilize the Hanning window function for text summarization.

In terms of the question-focused summarization, we introduce the following new methodologies: 1) generation of summary that can answer the multiple questions, and 2) integration of output of QA engine into conventional scheme of important sentence extraction by using QA score as word weight. Especially, as far as we know, there are only a few researches on text summarization using QA systems. The group of Columbia University employs a QA system in DUC 2003, they, however, did not describe the detail of that(Nenkova et al., 2003). Gaizauskas et al.(Gaizauskas, 2003) also have announced a project called 'Cubreporter', which aims to integrate the technologies including question answering and multi-document summarization. They, however, have not published the papers about that.

The basic methodology for the question-biased summarization is to give higher weight to words in a question(Tombros and Sanderson, 1998; Berger and Mittal, 2000; Nobata and Sekine, 2003). Okumura et al.(Okumura and Mochizuki, 2000) focused on the lexical chains with respect to words in a question sentence. Hirao et al.(Hirao et al., 2001) pay attention to not only words in a question but also named entities that correspond to the question type. Wu et al.(Wu et al., 2002) examined the relation between types of questions and the unit length of text fragment in summary generation. Our experiments described in Section 8, however, show that the proposed method, which uses the information of the answers from a QA system, outperforms the baselines that depend only on the information of questions.

11 Conclusion

In order to realize multi-document summarization focused by multiple questions, we introduced a calculation of sentence importance using a QA system. We also proposed an integration of it into a generic multi-document summarization system. The evaluation results showed that the proposed method has better performance than not only several baselines but also other participants' systems in NTCIR4 TSC3 Formal Run.

As stated earlier, it is observed that the proposed method sometimes missed eliminating obviously redundant sentences. Our future work, therefore, should include refining the calculation of sentence similarity. Another point at issue is the computational cost of the QA system. In the current implementation, the system calculates exact scores evenly for all of words in documents to be summarized. It takes about tens of seconds per one question on average PC hardware. Fortunately, the QA engine has the feature of controlled search in finding answers and can terminate the calculation after obtaining n-best answers. We plan to utilize the feature and the approximate scores to reduce the cost. The proposed method should be also improved so as to allow a more natural situation where a user gives questions one by one in the interaction with a system, and according to the input the system gradually outputs a part of summary including an answer by taking account of relations to previously displayed parts of summary.

References

Adam Berger and Vibhu O. Mittal. 2000. Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 294–301.

Jaime Carbonell and Jade Goldstein. 1998. 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.

Takahiro Fukusima and Manabu Okumura. 2001. Text Summarization Challenge Text summarization evaluation in Japan (TSC). In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, June.

Robert Gaizauskas. 2003. Cubreporter web page. http://nlp.shef.ac.uk/cubreporter/index.html.

Tsutomu Hirao, Yutaka Sasaki, and Hideki Isozaki. 2001. An extrinsic evaluation for question-biased text summarization on qa tasks. In *Proceedings of the NAACL 2001 workshop on Automatic Summarization*, pages 61–68.

Tsutomu Hirao, Manabu Okumura, Takahiro Fukushima, and Hidetsugu Nanba. 2004. Text Summarization Challenge 3 — Text summarization evaluation at NTCIR Workshop 4 —. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 407–411, June.

Tatsunori Mori, Tomoharu Ohta, Katsuyuki Fujihata, and Ryutaro Kumon. 2003. An A* search in sentential matching for question answering. *IEICE Transactions on Information and Systems*, E86-D(9):1658–1668, September. Special Issue on Text Processing for Information Access.

Tatsunori Mori. 2002. 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, August.

Ani Nenkova, Barry Schiffman, Andrew Schlaiker, Sasha Blair-Goldensohn, Regina Barzilay, Sergey Sigelman, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2003. Columbia at the Document Understanding Conference 2003. In *Proceedings of Document Understanding Conference 2003*.

Chikashi Nobata and Satoshi Sekine. 2003. Results of CRL/NYU system at DUC-2003 and an experiment on division of document sets. In *Proceedings of Document Understanding Conference 2003*.

Manabu Okumura and Hajime Mochizuki. 2000. Query-biased summarization based on lexical chaining. *Computational Intelligence*, 16(4):578–585.

Paul Over and James Yen. 2003. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of Document Understanding Conference 2003*.

A. Tombros and M. Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10.

Harris Wu, Dragomir R. Radev, and Weiguo Fan. 2002. Towards answer focused summarization. In *Proceedings of the 1st International Conference on Information Technology and Applications*.