Query-based Multidocument Summarization for Information Retrieval

Toshihiko Sakurai Akira Utsumi
The University of Electro-Communications
1-5-1 Chofugaoka, Chofushi, Tokyo 182-8585, Japan
toshi@utm.se.uec.ac.jp, utsumi@se.uec.ac.jp

Abstract

This paper presents a genre-independent method of generating a single summary from a set of the retrieved documents for information retrieval. The proposed method generates the core part of the summary from the most relevant document to a query, and then the additional part of the summary, which elaborates on the topics, from the other documents. In order to evaluate the validity of the proposed method, we participated in TSC (Text Summarization Challenge) in the 4th NTCIR Workshop. The performance was not satisfactory for the specific task, but we believe that our method would be useful for a set of documents including various genres, such as one retrieved by Web search engines.

Keywords: Multidocument Summarization, Information Retrieval, Query

1 Introduction

Finding relevant information in a set of retrieved documents is often difficult. For example, the existing Web search engines produce only a ranked list of links to Web pages (documents) with short descriptions consisting of some sentences including a query given by a user. The user has to browse these documents by viewing each at a time to get relevant information. Browsing is often a laborious task because information that the user searches for is scattered over the list of retrieved documents. Therefore, to show a single summary of the retrieved documents is highly helpful in finding information the user wants to get.

The purpose of this study is to automatically generate a single summary which includes both the information that a user searches for and the information characteristic of each of the retrieved documents. In order to realize the purpose and to eliminate the redundancy in a multidocument summary, we propose a multidocument summarization method which consists of two steps. At the first step, the core part of the summary, which identifies the content satisfying user's information need, is generated from the most relevant document to the queries. At the next step, the additional part of the summary, which elaborates on the core topics and reflects characteristic information included in the retrieved documents, is generated from the rest of the documents. These two summaries are merged into a single summary, which we consider as the informative summary based on the query.

2 Multidocument Summarization

2.1 Types of summary

There are two types of summaries according to the purpose: indicative summary and informative summary (Okumura and Nanba 1999). An indicative summary points to information of the document which helps the user to decide whether the document should be read or not. On the other hand, an informative summary provides all the relevant information to represent (and often replace) the original document. Since the purpose of this study is to provide a user with the topics of query, our method aims at generating an informative summaries.

2.2 Major approaches

There are the following three major problems introduced by having to handle multiple input documents for multidocument summarization (Radev, Hovy and Mckeown 2002).

1. recognizing and coping with redundancy
2. identifying important differences among documents
3. ensuring summary coherence

Since it is difficult to solve these problems comprehensively, we focus on the first two problems (removing redundancy and identifying differences). To identify redundancy in summary, various similarity measures are used. A common approach is to measure similarity between all pairs of sentences and then use clustering to identify themes of common information (Marcu and Gerber 2001).

3 Query-based Multidocument Summarization

This chapter presents a genre-independent method of generating a multidocument summary for information retrieval. In information retrieval, a user wants the topics about a query. IR systems provide a list of retrieved documents that contain query. The user must select documents that contain relevant information. However, the task of selecting relevant documents is very difficult because the retrieved documents are similar with respect to query. The idea underlying the proposed method is that to offer multidocument summarization to the user can reduce the burden of that task.

3.1 Outline

Basically, our method extracts important segments (sentences) from each retrieved document. In order to reduce redundancy, the extraction process is divided into two steps, as shown in Figure1. The algorithm of the method is as follows:

Step 1 Generating a core part of the summary ($Summary_Q$) from the most relevant document (D_1).

Step 2 Generating an additional parts of the summary ($Summary_C(D_i)$) which complement an information relevant to the query from the other documents (D_2 through D_n) in the order of document rank.

Step 1 extracts sentences ($Summary_Q$) which provide information about the query from the most relevant document. This idea is based on the assumption that such document includes the topics relevant to the query most abundantly. Step 2 extract sentences ($Summary_C(D_i)$) which complement information about query from the other retrieved documents by considering to reduce redundancy.

3.2 Generating a core part of the summary

In this section, we explain an algorithm of Step 1, generation of a core part of the summary relevant to the query. We adopt the method that extracts important passages. We use a single-document summarization method we have proposed (Sakurai and Utsumi 2004).

The algorithm of Step 1 is as follows:

Step1-1 Calculate a weight $W_D(t)$ of the term t (we use only nouns[1]) in the most relevant document D_1, using the following measures.

- Term frequency ($TF_D(t)$)
- Similarity between query and term ($SQ_D(t)$)

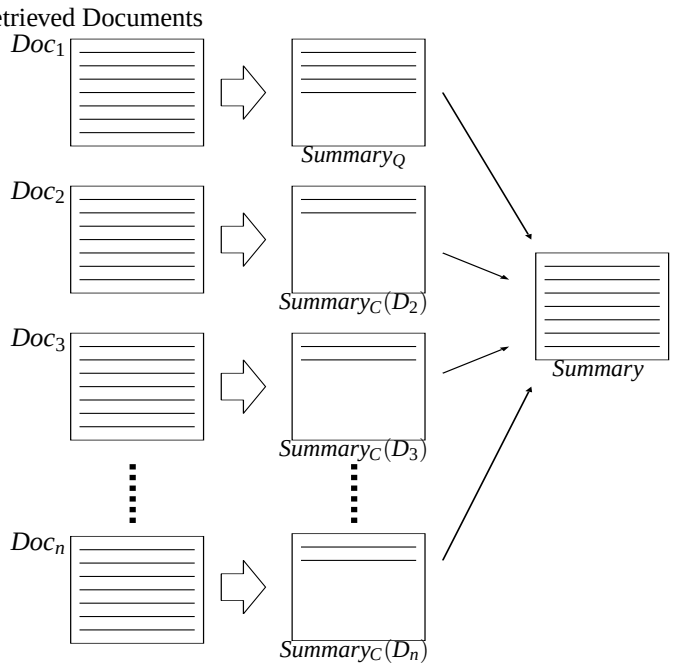

Figure 1. Query-based multidocument summarization method

- Cooccurrence level between query and term ($CO_D(t)$)

.

Step1-2 Calculate the degree of importance $I_D(S)$ of each sentence S using the weight of term $W_D(t)$.

Step1-3 Extract important sentences until the length of these sentences reaches 40% of the specified length of summary L, and output the extracted sentences as a core part of the summary ($Summary_Q$).

3.2.1 Calculation of a weight of term

Term frequency

Term frequency $TF_D(t)$ is the number of times the term t occurs in the document D. It provides one measure of how well that term describes the document contents.

Similarity between query and term

The degree of similarity $SQ_D(t)$ between the term t and the query q is calculated on the basis of the distance in the Japanese thesaurus "Bunrui-Goi-Hyo" provided by The National Institute for Japanese Language. When that term is grouped into the same category with the query in the thesaurus, its weight of the

[1] We use a Japanese morphological analyzer "ChaSen" to identify nouns (without ' 非自立',' 特殊',' 副詞可能', 助動詞語幹).

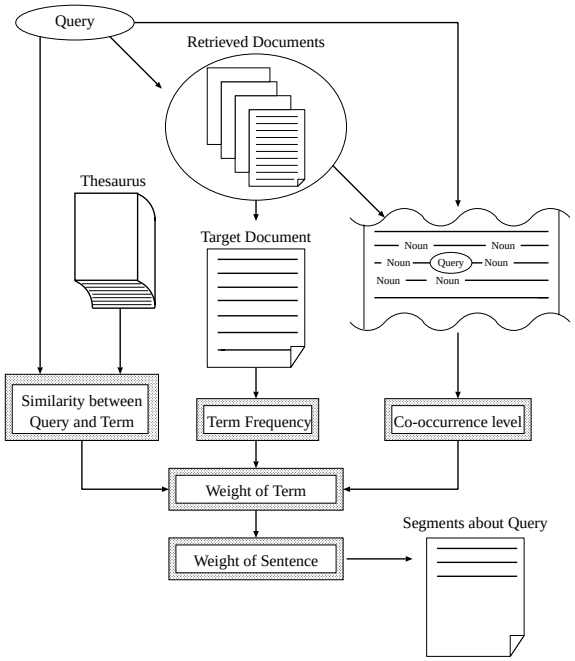

Figure 2. Extraction of a core summary about query

term by similarity $SQ_D(t)$ is calculated according to Table1. When the user gives multiple queries, $SQ_D(t)$ is calculated as the sum of similarity value to each query.

Cooccurrence of query and term

A term which frequently cooccurs with the query in the retrieved documents can be seen as implicitly related to the query. We assume that cooccurrence frequency of the term t provides a measure of how well that term describes the query-related topics. The weight $CO_D(t)$ of the term t by cooccurrence frequency is defined by the equation (1).

$$CO_D(t) = log_2\left(\frac{COF(t)}{TF_D(t)} + 1\right) \qquad (1)$$

In this equation, $COF(t)$ denotes the number of times the term t cooccurs with the queries within a 10-word window in all the retrieved documents. The reason for normalizing $COF(t)$ by the term frequency $TF_D(t)$ is that $COF(t)$ is likely to increase in proportion to the frequency of that term in the document.

Using these three values mentioned above, the weight $W_D(t)$ of the term t is calculated by the following equation.

$$W_D(t) = TF_D(t) + SQ_D(t) + CO_D(t) \qquad (2)$$

3.2.2 Calculation of importance of sentence

The degree of importance $I_D(S)$ of the sentence S in the document D is calculated by the following equation.

$$I_D(S) = \sum_{t \in S} \frac{W_D(t)}{n} \qquad (3)$$

In order to avoid a unreasonable effect of the sentence length, the degree of importance is normalized by the number of words n contained in the sentence S.

3.3 Generating an additional part of the summary

After a core part of the summary is generated from the most relevant document D_1 at Step 1 as described in Section3.2, an additional part of the summary is generated from the other documents. We assume that these documents contain a variety of information which elaborates on the topics related to the query. In this section, we explain an algorithm of Step 2 for extracting segments that contain detailed information on the query-related topics.

The algorithm of Step 2 is shown as follows:

Step2-1 Calculate a weight $W_{D_i}(t)$ of the term t in the ith relevant document D_i using the following measures.

- Difference weight of discriminative term ($PTF_{D_i}(t)$)

Table 1. Similarity between the term and the query using a thesaurus

Matching digit number	Example	$SQN_D(t)$
Topside four figures of grouping number	貿易 (commerce)1.4760,2,1,3 売買 (trade)1.4761,1,1,2	2
Grouping number	サッカー (soccer)1.3374,8,2,1 テニス (tennis)1.3374,7,7,1	4
Grouping number Paragraph number	野球 (baseball)1.3374.9,1,2 奪三振 (strikeout)1.3374,9,14,1	7
Grouping number Paragraph number Intra-paragraph number	議会 (legislature)1.2730,1,1,2 国会 (parliament)1.2730,1,1,3	10

- Inverse document frequency ($IDF_{D_i}(t)$)

Step2-2 Calculate the degree of importance $I_{D_i}(S)$ of each sentence S using the weight of term $W_{D_i}(t)$ by the equation (3) shown in Section 3.2.2.

Step2-3 Extract important sentences ($Summary_C(D_i)$) until the length of these sentences reaches the calculated length L_{D_i}, increase i by 1, and return to Step 2-1 until all documents are processed.

Step2-4 Output the extracted sentences ($Summary_C(D_2)$ through $Summary_C(D_n)$) as an additional part of the summary ($Summary_C$).

In the rest of this section, we describe the detail of Step 2-1 and a method of calculating the length of segment to extract from D_i in Step 2-3.

3.3.1 Calculation of weight of term

Step 2-1 extracts important sentences from the document D_i using the frequency of the terms included in the extracted sentences from the documents D_1 through D_{i-1}. In this section, we denote by SQ' the extracted sentences from D_1 through D_{i-1}, i.e., $Summary_Q$ and $\Sigma^{i-1}Summary_C(D_n)$.

Difference weight of discriminative term

We define a discriminative term in the document D_i as the term which does not occur or hardly occurs in the segment SQ', but occurs frequently in D_i. The discriminative term can be seen as an indicator of the characteristic contents of the document D_i which are not described in the summary SQ'. In order to judge whether a term t is a discriminative one, Step 2-1 calculates the difference of the normalized frequencies of that term between the document D_i and the segment SQ, i.e., $TF_{D_i}(t) - TF_{SQ'}(t)$. If this value is positive, that term is judged to be discriminative.

According to the above idea, the difference weight $PTF_{D_i}(t)$ of the term t in the document D_i is calculated by the following equation.

$$PTF_{D_i}(t) = \begin{cases} TF_{D_i}(t) - TF_{SQ'}(t) & TF_{D_i}(t) > TF_{SQ'} \\ 0 & TF_{D_i}(t) \leqq TF_{SQ'} \end{cases} \quad (4)$$

Figure3 shows an example of $PTF_{D_i}(t)$ calculation.

Inverse document frequency

Inverse document frequency $IDF_{D_i}(t)$ of the term t is calculated by following equation.

$$IDF_{D_i}(t) = \log \frac{N}{df(t)} + 1 \quad (5)$$

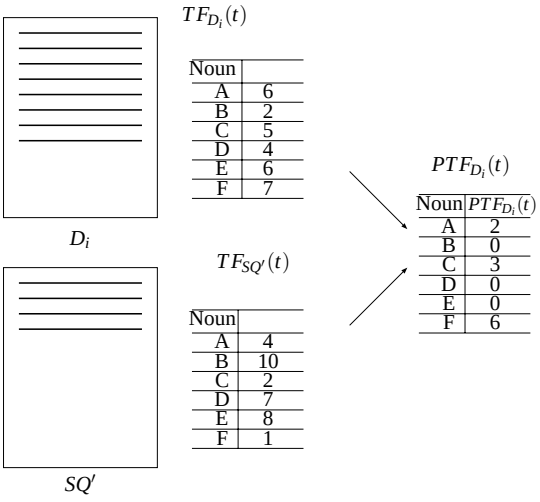

Figure 3. Calculation of discriminative term

In equation (5), N is the number of the retrieved documents and $df(t)$ is the number of documents in which the term appears.

Using these two values mentioned above, the weight $W_{D_i}(t)$ of the term t is calculated by following equation.

$$W_{D_i}(t) = PTF_{D_i}(t) \times IDF_{D_i}(t) \quad (6)$$

3.3.2 Length of summary

In order for the generated summary to be within the given summary length L_T, the length L_{D_i} of $Summary_C(D_i)$ is calculated by following equation.

$$L_{D_i} = (L_T - L_{SQ'}) \times (0.3 + 0.7\frac{i-2}{N-1}) \quad (7)$$

In this equation, N is the number of the retrieved documents.

3.4 Example of summary

Figure 4 shows an example of the generated summary, given the keywords “天体, 望遠鏡, すばる, 試験, 観測, 開始” as query. The description in parentheses at the end of each paragraph denotes the original document from which that segment was extracted. As we can see from Figure 4, the core part of the summary $Summary_Q$ contains the topics about query such as “すばる” and “観測”. On the other hand, the additional part of the summary $Summary_C$ hardly contains the query-related topics, but it includes the information that is not described in $Summary_Q$. This exmaple of summary is generated from seven documents out of eight ones in a document set, but it may hardly contain redundant sentences.

４００億円をかけて建設されたすばるは口径８・２メートルの望遠鏡で，１枚の鏡としては世界最大．観測装置は全部で７台あるが，そのうちの２台を使って試験観測を実施している．すばるが建つマウナケア山頂は海抜４２００メートル．このため，山頂には各国の望遠鏡が並んでおり，すばるを含めて１３基を数える．大型望遠鏡は複数の鏡を組み合わせた米国のケック（口径１０メートル）と，近くファーストライトを迎える米英加チリなどの共同観測施設ジェミニ（同８・１メートル）だ．($Summary_Q$)

最近では，ハッブル宇宙望遠鏡が，光を弱める大気の妨げがない地球の外側で，目覚ましい観測成果を上げている．光を集める能力はハッブル宇宙望遠鏡の十倍以上．($Summary_C(D_2)$)

紀宮さまは「遠く星々をはるかに超えて銀河のかなたを旅する冒険を連想させます」と応じ，すばるの可能性に期待する気持ちを述べられた．九年に及ぶ工事では地元作業員四人が事故で亡くなり，式典で紀宮さまも冥福（めいふく）を祈られた．ドームの基礎には四人の名を刻んだプレートも埋め込まれている．($Summary_C(D_3)$)

中心にある４個の明るい星の周囲に，生まれてまもない小質量の星がたくさん見える．右上の赤い星雲は中心部にある太陽の３０倍の重さの原始星がガスを激しく噴き出している現場で，可視光では見えない．($Summary_C(D_4)$)

新たな宇宙の姿をとらえた映像や，思いがけない発見を続々と届けてくれそうだ．($Summary_C(D_5)$)

２８日夜（日本時間２９日昼）には，初めての観測の成果が，日本と現地で同時発表される．($Summary_C(D_7)$)

望遠鏡は鏡の口径が大きいほど集光力が大きく，より暗い天体を観測できる．($Summary_C(D_8)$)

Figure 4. Example of summary generated by the proposed method

4 Application of our method to TSC3

In order to evaluate the validity of the proposed method, we participated in TSC3 (Text Summarization Challenge 3) at the 4th NTCIR Workshop.

The method proposed in Section 3 is originally developed to summarize multiple documents retrieved by information retrieval systems based on keyword-based queries such as Web search engines. However, for TSC3 tasks we were provided with document sets, titles of document sets (sentence queries), and question sets. Therefore, we must extract keyword-based queries from the given information and rank the documents in order to perform the TSC3 tasks by the proposed method.

4.1 Keyword-based query extraction

We selected the following noun words as keyword-based queries.

A: Nouns which occur in the title of the document set.

B: Nouns in the question set whose term frequency TF is more than 7 % of the sum of all TFs in that question set.

4.2 Document ranking

We rank the provided documents by the following algorithm.

Step1 Selecting as the most relevant document D_1 the document in which the queries occur the most frequently in the three longest documents of the document set.

Step2 Ranking the other documents according to the frequency of low-frequent terms, which occur less than three times in the question set.

5 Evaluation

In this chapter, we report the evaluation result of our method obtained by performing two tasks of TSC3: important sentence extraction and multidocument abstraction. Summaries (extracts and abstracts) were generated from the document sets provided by TSC. In this chapter, we show the result of the proposed method (our participant ID is F0310), together with those of the baseline system by the lead-based method (LEAD) and human-generated summaries (HUMAN). We also show the rank of our method among the nine participants.

5.1 Intrinsic metrics for extraction

Table 2 shows the results of extraction. The numbers in parentheses of Table 2 denote the rank of our method.

Table 2. Evaluation results (Extraction)

	Short		Long	
Method	Cov.	Prec.	Cov.	Prec.
Ours	0.181	0.275	0.218	0.421
(Rank)	(10)	(10)	(10)	(9)
LEAD	0.212	0.426	0.259	0.539

Table 3. Results on content evaluation

Method	Short	Long
Ours	0.131	0.233
(Rank)	(9)	(6)
LEAD	0.160	0.159
HUMAN	0.358	0.402

5.1.1 Precision

The overall result of our method is poor, but one interesting result deserves special mention. Our method achieves a better result in the long extract than in the short extracts, and the difference of precision between them is the largest in the participants. This result suggests that our two-step summarization method is suitable for long summarization.

5.1.2 Coverage

The score of coverage is low in both the short extract and the long extract. Moreover, our generated summaries contain sentences which is not semantically related to the query (the title of documents). The reason for the poor result may be that our method is developed for information retrieval based on keyword-based queries, and highly depends on the characteristic information of information retrieval. There is a room for improvement about the our method, especially a method of document ranking and determination of a condensation rate.

5.2 Intrinsic metrics for abstraction

5.2.1 Content evaluation

Table 3 shows the results on content evaluation by human subjects. Our result of short abstract is worse than those of the lead method and other participants' methods. On the other hand, the result of long abstract is not bad; it is higher than the lead method and almost the same as the other methods. It again suggests that our method is suitable for long summarization.

5.2.2 Readability evaluation

Table 4 shows the results on readability evaluation by "Quality Questions". For redundancy of the generated summaries measured by "q00" score, our method does not yield a satisfactory performance. It suggests that our method on the basis of discriminative terms does not work well. On the other hand, our method achieves good scores in some questions ("q01", "q04", "q07", "q10", "q14"). Especially, our method achieves the best score of "q04" in all the methods. It means that our two-step method can reasonably reduce semantically similar expressions.

5.3 Extrinsic metrics for abstraction

Table 5 shows the results on pseudo question-answering. The result of our method is worse than any other participants in both the short and the long abstract. This result is not surprising because we did not use any method specific to question answering; we only used information of nouns extracted from question sentences.

Table 4. Results on readability evaluation

	Short				Long			
	Our Method	(Rank)	LEAD	HUMAN	Our Method	(Rank)	LEAD	HUMAN
q00	1.967	(9)	1.500	0.033	2.367	(9)	2.833	0.033
q01	0.200	(1)	1.267	0.267	0.133	(1)	2.100	0.167
q02	1.767	(9)	0.267	0.000	2.500	(9)	0.633	0.100
q03	0.400	(6)	0.267	0.000	0.500	(7)	0.367	0.000
q04	0.633	(1)	1.667	0.433	2.567	(1)	4.300	1.133
q05	3.800	(9)	0.067	0.400	4.667	(9)	0.300	0.467
q06	1.333	(9)	0.767	0.400	1.833	(9)	1.033	0.433
q07	0.167	(1)	1.533	0.000	0.333	(1)	4.333	0.067
q08	-0.600	(8)	0.267	0.933	-0.733	(6)	-0.333	0.800
q09	0.233	(6)	0.067	0.500	0.333	(7)	0.067	0.567
q10	0.000	(1)	1.667	0.033	0.300	(2)	5.133	0.000
q11	0.200	(9)	0.000	0.000	0.233	(8)	0.000	0.033
q12	0.033	(9)	0.000	0.000	0.067	(8)	0.000	0.000
q13	0.000	(1)	0.033	0.033	0.067	(8)	0.000	0.000
q14	0.000	(1)	0.033	0.033	0.000	(1)	0.067	0.033
q15	0.133	(5)	0.200	0.033	0.233	(3)	0.433	0.100

Table 5. Evaluation results (Extraction)

	Short		Long	
Method	exact	edit	exact	edit
Ours	0.133	0.427	0.201	0.549
(Rank)	(9)	(9)	(9)	(9)
LEAD	0.300	0.589	0.275	0.602
HUMAN	0.461	0.716	0.426	0.721

6 Conclusions

In this paper, we have proposed the method of generating a multidocument summary for information retrieval. We also evaluated the validity of the proposed method by participating in TSC3. The overall performance of TSC3 was not satisfactory, but the result of precision suggests that our method may have a beneficial effect on long summaries. Of course, we recognize that many improvements still remain to be done. Since our method has been developed for summarizing multiple documents retrieved by IR systems like Web search engines, it must deal with ill-formed documents such as Web pages.

References

[1] Marcu, D. and Gerber, L.: An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, 1–8 (2001).

[2] Radev, D., Hovy, E. and Mckeown, K.: Introcuction to the special issue on summarization. *Computational Linguistics*, Vol.28, No.4, 399–408 (2002).

[3] Sakurai, T. and Utsumi, A.: Query-based summarization for information retrieval. In *Proceedings of The Tenth Annual Meeting of The Association for Natural Language Processing*, 265–268 (2004), in Japanese.

[4] Tombros, A. and Sanderson, M.: 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 (SIGIR'98)*, 2–10. (1998).

[5] Okumura, M. and Nanba, H.: Automated text summarization: A survey. *Journal of Natural Language Processing*, Vol.6, No.6, 1–26. (1999), in Japanese.