Sentence Extraction by tf/idf and Position Weighting from Newspaper Articles

Yohei SEKI
Dept. of Informatics, The Graduate University for Advanced Studies (Sokendai)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Dept. of Integrated Information Technology, Aoyama Gakuin University
6-16-1 Chitosedai Setagaya-ku, Tokyo 157-8572, Japan
seki@grad.nii.ac.jp/seki@it.aoyama.ac.jp

Abstract

Recently, lots of researchers are focusing their interests on the development of summarization systems from large volume sources combined with knowledge acquisition techniques such as information extraction, text mining or information retrieval. Some of these techniques are implemented according to the specific knowledge in the domain or the genre from the source document. In this paper, we will discuss Japanese Newspaper Domain Knowledge in order to make a summary. My system is implemented with the sentence extraction approach and weighting strategy to mine from a number of documents.
Keywords: NTCIR, multi-document summarization, newspaper article structure, and TSC (Text Summarization Challenge).

1 Introduction

There is a year long workshop being held by the National Institute of Informatics in Japan called NTCIR-3. We participated in the 'Text Summarization Challenge' (TSC) formalrun in the spring of 2002: Japanese Summarization tasks. We created an experimental system for the Japanese Summarization to compute the importance value for each sentence based on Japanese newspaper terms. Our input data was collected from Mainichi Newspaper articles from 1998 and 1999. This included about 230,000 articles. In this paper, we discuss mainly multi document summarization[6, 7, 9, 10] related problematic issues from several Japanese newspaper articles.

This paper consists of five sections. We explain the tasks of TSC in Section 2, and discuss details of our system strategy for sentence selection in Section 3. Section 4 contains a brief evaluation of our system with TSC problems. In Section 5, another strategy for better results is detailed. Finally, we present our conclusions in Section 6.

2 Summarization Tasks in TSC

The Text Summarization Challenge (TSC) consists of two tasks. Task A consists of single document summarization problems for two patterns of sizes from 200 to 500 character numbers. Two pattens of sizes corresponds to 20% and 40% condensed rates. Task A has 30 problem source articles. Task B also consists of multi document summarization problems for two patterns of sizes from 125 to 1000 character numbers. It addresses 30 topics as well, which each contain 3 to 17 source articles. Source article IDs are previously specified. In addition, 1 to 4 keywords are also specified. The formalrun with these tasks was held on four consecutive days in May, 2002.

3 Sentence Extraction Strategy

3.1 Single Document Summarization

My summarization approach is based on a sentence weighting ordering approach. Sentence importance weights are computed in the following.

1. The tf/idf values of nouns in the sentence are biased.
   First, the tf/idf values of all nouns in the document except some stop words are computed. The tf/idf value equation is as follows: $TermFrequency \times log(AllDocumentNumbers \div DocumentFrequency)$. According to each document, the sum of all the tf/idf values of nouns in the document is computed. The importance value of a sentence is computed by the sum of tf/idf values of sentences containg nouns divided by the sum of all tf/idf values in the document.
2. The weights of phrases included in the Heading are biased.
   If a sentence contains phrases in the heading, the number of phrases is divided by the total phrases in the heading. That value is then multiplied by the constant 0.1, and adds to the sentence weights.
3. Position Weights are biased.
   The line number of the sentence in the document divided by the number of all lines in the document corresponds to the position value in the document from 0 to 1. We surveyed 10% summary data from the Japanese NTCIR2-SUMM corpus in the Mainichi Newspaper from July to November in 1998. The correct summary position weights are shown in Figure 1. These values are multiplied by the importance value in Step 2. These are the resultant importance values of each sentence.
4. Summarization.
   Finally, important sentences whose sum of characters exceeds the restricted character amount, are eliminated. The remaining sentences are then sorted as they appeared in the original document.

3.2 Multiple Document Summarization

In order to generate one summarization document from multiple articles, one approach is to compute each sentence's importance weight within each document. The simplest strategy is to extract important sentences equally from every related document according to the rates of summarization and arrange them chronologically. By weighing sentence importance with tf/idf value of a contained lexical set or words in the heading, we can extract sentences specific to each document.

Another method is weighing each sentence across the document set. In order to implement this strategy, the importance value of each sentence is adjusted from 0 to 1 by dividing the sum of tf/idf values contained in each sentence and comparing sentences' importance values across all documents.

I take the first approach. First, the character numbers required for a summary are divided by the sum of character numbers from all the source articles. This value corresponds almost equally to compress rates. According to these rates, the line number needed for extraction from each document is computed. These sentences are extracted from every source article by following the single document summarization strategy. This process proceeds in the document publishing time stamp order, and if the sum of characters contained in the summary is over the restricted character number, the summary construction process stops. If sentences are extracted from whole source articles and the sum of the character numbers is still under the restricted character amount, the subsequent important sentences are extracted one by one from each article chronologically.

4 Evaluation

TSC results were evaluated with score ranking with content and readability for 20% and 40% condensed rates. The results of our present system are shown as follows.

line position/all lines	$0 < x \leq 0.1$	$0.1 < x \leq 0.2$	$0.2 < x \leq 0.3$	$0.3 < x \leq 0.4$	$0.4 < x \leq 0.5$
Distributed Probability	0.17	0.23	0.14	0.08	0.05
line position/all lines	$0.5 < x \leq 0.6$	$0.6 < x \leq 0.7$	$0.7 < x \leq 0.8$	$0.8 < x \leq 0.9$	$0.9 < x \leq 1$
Distributed Probability	0.04	0.06	0.04	0.04	0.15

Figure 1. Distributed Probability of Important Sentences

4.1 Task A (Single Document Summarization)

Task A had 30 problems. This task is evaluated with human ranking from the best (1 point) to the worst (4 point). My system's evaluation did not vary whether 20% or 40% summary. The difference between Content evaluation and Readability evaluation also did not change significantly. The total score ranges are shown in Table 1.

4.1.1 Good Results for Single Document Summary

In the summaries with the article ID 990424046, 990624052, and 990629039, the long summary got the first score. In contrast to human-written summaries, my system's summary include necessary elements and news background, reason, origin and explanation. Short summary, however, did not contain necessary elements in the topic sentence.

In the summaries with the article ID 990402040 and 990403032, the short summary also got the first score. In contrast to human-written summaries which contain many clauses that lacked detailed information, my system had an effective combination of facts and opinion.

4.1.2 Problems for Single Document Summary

Some documents include a multi-topic sentence like "We have three perspectives" in the article ID 990624050. For this type documents, short summaries must contain compact explanations for each perspective and reduce some background or explanation elements according to desired summary size. In addition, autobiographic type articles like the article ID 990109032 include some

ArticleID	C 20%	R 20%	C 40%	R 40%
990109032	3	4	3	4
990117039	3	4	3	4
990120043	3	3	3	4
990129047	2	2	3	3
990130032	3	3	3	3
990201036	3	3	3	3
990202041	3	3	3	3
990105044	3	3	3	3
990211049	3	3	3	3
990305053	3	3	3	3
990311036	3	3	3	3
990313042	3	3	3	3
990313046	3	1	1	1
990402040	1	1	1	1
990403032	1	1	1	1
990410033	3	3	3	3
990428029	3	3	3	3
990430039	3	3	3	3
990501040	2	2	1	2
990502043	3	3	3	3
990531030	3	4	3	4
990603040	3	3	3	3
990604040	3	3	3	3
990605036	3	3	3	3
990616038	3	3	3	3
990618040	3	3	3	3
990624050	3	3	4	4
990624052	2	2	1	1
990629039	2	2	1	1
990630039	2	2	2	2
avg	2.70	2.73	2.60	2.77

Table 1. Scoring in Task A

important events and their reference expressions, but this type importance cannnot be computed from term-based weighting. These two types had better be categorized or discriminated for different summarization strategies, such as multi document summarization for DUC (Document Understanding Conference)[1].

4.1.3 For Better Single Document Summarization

One of the problematic issues is to cut the abstract important sentence. In Japanese language, this sentence corresponds to the conclusion in the tail part of the article. Conclusion part in Japanese newspaper articles is sometimes not equal to introduction, and this fact must be considered in case of to make a summary. My system's positional weighting approach take it into a consideration, but a different approach may be effective.

One of the good scoring case to make a short summary as in [2, 3, 8] is to take more facts than facts in the sample summaries. In case of a short summary, strategy to select many short sentences may be more effective. In contrast, in case of a long summary, the resultant summary consisted of many short sentences would be hard to read.

In addition, if detailed short modified terms like "The business in 'Oogata village, Akita prefecture'" or "The judgement 'last November'" had better not be eliminated. From the viewpoint of the construction, many supporting stories of an experience and concise conclusion would be better than a loose style discussion. These results mean sharp and smart construction with real experiences and numeric data would be easy to read for a summary.

4.2 Task B (Multiple Document Summarization)

Task B had 30 topics to make a summary with a few keywords from multiple article sources. The result is evaluated as in Task A. My system scores are shown in Table 2.

TopicID	C Short	R Short	C Long	R Long
0010	1	1	4	4
0020	2	2	2	2
0030	4	3	3	2
0040	2	4	4	4
0050	2	2	2	2
0060	2	4	3	4
0070	2	1	4	3
0080	2	3	3	3
0090	2	4	2	4
0100	2	2	3	3
0110	4	4	3	3
0120	2	4	2	2
0130	4	4	4	4
0140	2	2	4	3
0150	3	3	2	2
0160	4	4	3	4
0170	3	2	2	4
0180	3	4	4	4
0190	3	4	3	3
0200	4	4	1	2
0210	4	4	4	4
0220	4	4	3	4
0230	3	3	3	3
0240	1	1	1	3
0250	3	2	2	2
0260	2	2	2	2
0270	2	2	1	2
0280	2	1	3	2
0290	2	3	3	3
0300	3	4	4	4
avg	2.63	2.90	2.80	3.03

Table 2. Scoring in Task B

4.2.1 Good Reults and Problems for Multi Document Summary

The short summaries in the topic ID 0010, 0070, 0240, and 0280 got the first score for readability. These summaries consisted of a few good sentences and consistency is good. The readability for long summaries in these articles, however, did

[1] http://www-nlpir.nist.gov/projects/duc/index.html

not score well because a topic sentence appeared in the middle of the discourse. The long summaries in the topic ID 0200 and 0270 scored well. In these two summaries, several topics from several documents were well detailed and the topic's description consistency were good. Another problem is sentence ordering from several documents[1]. Some time event expressions in several documents jumped their chronological ordering.

4.2.2 For Better Multi Document Summarization

In case of multi document summarization, the different evaluation between a long summary and its short summary is more remarkable. Adding topic start sentences or repetitious referent elements and shifting focus according to the source document might cause the difficulties to read for a long summary. In order to avoid this problem, size change process according to the source document or to eleminate or sort repetitious referents across the documents might be required.

In contrast, in case of topicID 0200, a short summary has a bad evaluation while its long summary is not. This is because a proper topic start sentece is added into the first line and a proper supporting facts is inserted into other middle lines. This means a summary must be constructed from source documents according to each sentence position and its role. We would try a 'sentence-role based summarization' technique [8] for multiple document summarization from this findings.

5 Another Strategy For Better Results

We propose important sentences filtering techniques with sentence roles[8]. Some rhetorical sentence roles for newspaper articles are proposed as in [5]. We try a proposal in Kando[4] and tagged 5 specific sentence roles in newspaper articles as 'Main Description', 'Background', 'Elaboration of Main Description', 'Prospectives' and 'Opinion'. Sentence role tagging algorithm is in the following and this algorithm contains an overriding process for tagged roles.

STEP 1: 'Elaboration' roles are tagged according to whether a sentence contains heading noun phrases.

STEP 2: 'Background' roles are tagged according to whether a sentence contains numeric-data or time-related suffixes (Year, Day, or $).

STEP 3: 'Prospective' roles are tagged by some auxiliary elements (will, may, or seem) or noun elements (prospective, assumption or possibility).

STEP 4: 'Opinion' roles are tagged by some auxiliary elements (want, desire or think), quoting ending mark, some adjectives and juding auxiliary elements.

STEP 5: Of the remeining 'Elaboration' role sentences, the most important sentence is assigned to 'Main Description' role.

STEP 6: All sentence importance weights are biased by phrases in 'Main Description' sentence.

The combination of 'Main Desciption' and 'Background' or 'Prospective' are effective for evaluation with the simple weighting approach. My proposal is to select important sentences with each role according to summary size and construct summary from different role sentences with the same topic.

6 Conclusions and Future Direction

We tested my experimental summarization system mainly using pure weighting approach biased with positional weights and its application to the multi-document summarization based technique. The results showed that the concluding part could not be extracted from the source document only by biasing positional weights. In addition, its application to multi-document summarization has a defect for inserting sentences from several documents. This insertion process caused breaking a discourse construction. Pure weighting approach limitation is shown in this result.

In order to improve my results, some semantic information for summarization may be required to reduce the redundancy and to make a constructive summary. In addition, if the assigned threshold for summarization is changed according to the

sentences already containing in a summary, better results will follow. In order to determine precise extracted summary, we will try 'sentence-role' based summarization work.

Acknowledgements

We thank the National Institute of Informatics and organizing members held for NTCIR-3 TSC, and also thank Patricia Lipska for re-reading the English.

References

[1] R. Barzilay, N. Elhadad, and K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17:35–55, 8 2002.

[2] B. Boguraev, R. Bellamy, and C. Swart. Summarization miniaturisation: Delivery of news to hand-held. In the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001) Workshop on Automatic Summarization, pages 99–108, Pittsburgh, June 2001.

[3] S. Corston-Oliber. Text compaction for display on very small screens. In the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001) Workshop on Automatic Summarization, pages 89–98, Pittsburgh, June 2001.

[4] N. Kando. Text structure analysis based on human recognition: Cases of japanese newspaper articles and english newspaper articles (in japanese). National Center for Science Information Systems Reports, 8:107–126, 1996.

[5] E. D. Liddy. Development and implementation of a discourse model for newspaper texts. In the Dagstuhl on Summarizing Text for Intelligent Communication, Saarbruken, Germany, 1995.

[6] I. Mani. Automatic Summarization, volume 3 of Natural Language Processing. John Benjamins, Amsterdam, Philadelphia, first edition, 2001.

[7] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In the 18th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 74–82, Seattle, WA USA, July 1995.

[8] Y. Seki and N. Kando. Dynamic document generation based on tf/idf weighting. In Mobile Personal Information Retrieval Workshop at the 25th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 57–63, Tampere, Finland, August 2002.

[9] G. C. Stein, T. Strzalkowski, and G. B. Wise. Summarizing multiple documents using text extraction and interactive clustering. In Pacific Association for Computational Linguistics (PACLING-1999), 1999.

[10] G. C. Stein, T. Strzalkowski, G. B. Wise, and A. Bagga. Evaluating summaries for multiple documents in an interactive environment. In 2nd Int. Conf. on Language Resources & Evaluation (LREC2000), 2000.