Text Summarization Challenge

Text summarization evaluation at NTCIR Workshop2

Takahiro Fukusima
Otemon Gakuin University
fukusima@res.otemon.ac.jp

Manabu Okumura
Tokyo Institute of Technology
oku@pi.titech.ac.jp

Abstract

We describe the outline of Text Summarization Challenge (TSC hereafter), a text summarization evaluation conducted as one of the tasks at the NTCIR Workshop2. First, we introduce TSC explaining its background and purpose. Then we describe briefly types of summarization and summarization evaluation methods in general. Next, we focus on TSC, including the participants, the three tasks in TSC, data used, evaluation methods for each task, brief report on the results as well as the features of the Challenge. The future directions for TSC are mentioned at the end as conclusion.

Keywords: text summarization, summarization evaluation

1 Introduction

As research on automatic text summarization is becoming a hot topic in NLP, we also see the needs to discuss and clarify the issues on how to evaluate text summarization systems. SUMMAC in May 1998 as a part of TIPSTER (Phase III) project ([1], [2]) and on-going TIDES program shows the need and importance of the evaluation for text summarization.

In Japan, there has been a lot of research on automatic text summarization. However, since the evaluation of such systems was done individually with their own evaluation measures at universities and industrial research organizations, and there have been few discussions about evaluation measures, which makes it difficult to compare text summarization systems. Moreover, we did not have enough language resources such as human-prepared summaries.

Thus, we have chosen automatic text summarization as an NTCIR-2 task in order for the researchers in the field to collect and share text data for summarization, and to make clear the issues of evaluation measures for summarization of Japanese texts. Before we started TSC formally, we had chosen 33 organizing committee members and hold two preliminary meetings in 1999. We also initiated a mailing list specifically for discussing issues related to TSC. At the same time, we have set up a website for TSC written in Japanese and English at http://galaga.jaist.ac.jp:8000/tsc/. What we discussed at the two meetings and in the mailing list has been archived and can be seen in the website.

2 Types of summarization

We would like to introduce briefly types of text summarization, and summarization evaluation methods in this and the following sections.

Text summarization is a task of producing a shorter text from the source, while keeping the information content in the source, and summaries are the results of such task.

There are several ways to classify summaries. The following threes factors are considered to be important for text summarization research ([3]).

Input factors: text length, genre, single vs. multiple documents

Purpose factors: who is the user, the purpose of summarization

Output factors: running text or headed text etc.

Summaries can be classified with respect to the number of the source text (single text vs. multiple texts summarization), and as to whether they are tailored to particular users. If they are targeted for specific users, they may be called "user-focused", and if they are intended to users in general, such summaries are "generic".

In terms of summarization purposes, summaries can be "indicative" or "informative". Users can make use of indicative summaries before referring to the source, e.g. to judge relevance of the source text. On the other hand, users may use summaries in place of the source text (informative summaries).

Summaries can also be classified into extracts and abstracts in terms of how they are composed. Conventional text summarization systems produce summaries by using sentences or paragraphs as basic unit, giving them degree of importance, sorting them based on the importance, and gathering the important sentences. In short, summaries are a set of important sentences extracted from the source text and called "extracts".

In contrast, summaries may contain newly produced text, and they are called "abstracts" ([4]).

3 Evaluation methods

It is said that evaluation methods for text summarization can be largely divided into two categories: intrinsic and extrinsic.

The quality of summaries is judged directly with some norms, typically "ideal" summaries produced by hand, or important sentences selected by hand (intrinsic evaluation) ([5], [6]).

The quality of summaries can also be judged by measuring how it influences the achievement of some other task (extrinsic evaluation). Such tasks can be question-answering, comprehension task, as well as relevance judgement of a document to a topic ([2]).

In TSC, we have conducted evaluations for text summarization of varied lengths, varied genres, and single-document using both intrinsic and extrinsic methods.

4 Participants

We had 9 participating systems for Task A-1 and A-2, and 7 systems for Task B at the Dryrun. We have 10 participating systems for Task A-1, 9 systems for Task A-2, and 9 systems for Task B at the Formal run. As group, we had 9 participating groups that are all Japanese, from universities, governmental research institute or companies in Japan. Table 1 shows the breakdown of the groups. (Please note that one group consists of a company and a university.)

University	3
Governmental research institute	1
Company	5

Table 1 Breakdown of Participants

5 Three Tasks in TSC and its Schedule

TSC has three tasks. First two tasks are for intrinsic evaluation (called Task A) and they are summarization tasks where participants are given texts to be summarized automatically and summarization rates (or lengths of summary). They should submit the results of their summarization system according to the summarization rates.

There are two types of summarizing tasks in Task A.

Task A-1 is to extract important sentences. Summarization rate is given as a ratio between the number of chosen sentences and the total number of the sentences in the article. The rates are 10%, 30%, and 50%.

Task A-2 is to produce summaries in plain text to be compared with human-prepared summaries. Summarization rate is a rate between the number of characters in the summary and the total number of characters in the original article. The rates are about 20% and 40%.

The third and final task is for extrinsic evaluation, called here Task B where summaries are evaluated by conducting IR task. Given queries and retrieved documents based on the queries, participants submit summaries. The length of the summaries is not limited, however, the summaries should be in plain text. They should make one summary for each document (not a summary from multiple documents). The retrieved documents may include irrelevant documents to the queries.

The schedule of evaluations at TSC is as follows: the Dryrun was conducted in September 2000 and the Formal run was in December 2000. The final evaluation results have been reported to the participants by the end of December 2000.

6 Data Used for TSC

We use newspaper articles from the Mainichi newspaper database of 1994, 1995, 1998. As key data (human prepared summaries), we prepare the following types of summaries.

Extract-type summaries:

We asked captioners who are well experienced in summarization to select important sentences from each article. The summarization rates are 10%, 30%, and 50%. These summaries are used for Task A-1.

Abstract-type summaries:

We ask the captioners to summarize the original articles in two ways. The first is to choose important parts of the sentences recognized important in extract-type summaries (abstract-type type1). The second is to summarize the original articles "freely" without worrying about sentence boundaries, trying to obtain the main ideas of the articles (abstract-type type2). Both types of abstract-type summaries are used for Task A-2.

In terms of genre, we used editorials and articles on social issues for the formal run evaluation. The lengths are grouped into two kinds for the editorials: about 1200 and 2400 characters, for social issue articles, the lengths are three kinds: about 600, 900, and 1200 or more characters.

7 Evaluation Methods for each task

We use summaries prepared by human as key data for evaluation. We would like to describe evaluation methods for each task below.

7.1 Evaluation methods for Task A

For Task A-1, we use recall, precision, and F measures where:

Recall = the number of correct sentences marked by the system / the total number of correct sentences marked by human

Precision = the number of correct sentences marked by the system / the total number of sentences marked by the system

F-measures = 2 x Recall x Precision / (Recall + Precision)

After calculating these scores for each article, we compute the average of them and make it its final score. We also evaluate the results of two baseline systems. One is based on lead method (lead), and the other based on term frequency (tf).

For Task A-2, the evaluation is not as simple as A-1 above. The system results are compared with human-prepared summaries in two ways (A-2-1 and A-2-2) as explained below.

A-2-1 (content-based evaluation)

We try to find out how close the two summaries are by examining the content words ([7]). Morphological analysis is done to the system results and human summaries, and only content words *(keitaiso*) are selected. Then, the distance between the word-frequency vector of human summary and system result is computed.

We use both abstract-type summaries (abstract-type type 1 and typ e 2) as key.

A-2-2 (subjective evaluation)

We ask human judges, who are experienced in producing summaries, to evaluate and rank the system summaries in terms of two points of views.

- How much the system summary covers the important content of the original article.
- How readable the system summary is.

The judges are given 4 types of summaries to be evaluated and ranked in 1 to 4 scale (1 is the best, 2 for the second, 3 for the third best, and 4 for the worst). The first two types are human-produced abstract-type type1 and type2 summaries. The third is system result, and the fourth is summaries produced by lead-based method.

7.2 Evaluation method for Task B

For Task B, the evaluation is an extrinsic one using information retrieval task. Human subjects are given queries and summaries of the retrieved documents. They read the summaries and judge how relevant the documents are. The evaluation method is basically the same as SUMMAC ([2], [8]). The measures for evaluation are recall, precision and F-measures as well as the time to indicate how long it takes to carry out the task.

Recall = the number of documents judged relevant correctly by human subjects / the total number of relevant documents

Precision = the number of documents judged relevant correctly by human subjects / the total number of documents judged relevant by subjects

F-measures = 2 x Recall x Precision / (Recall + Precision)

8 Results

Here are the results for Task A-1 for selecting important sentences, and summarization rates are 10%, 30%, and 50%. Since recall and precision scores turned out to be the same for the all systems, F-measures scores are used in Figure 1.

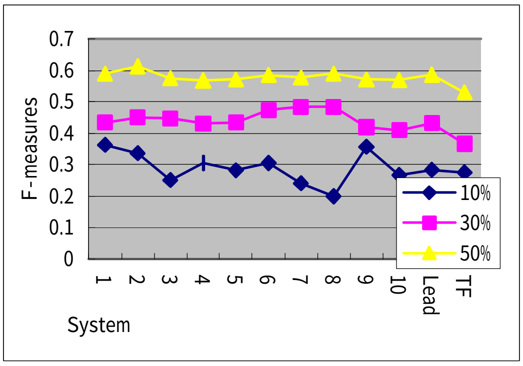

Figure 1 Evaluation Results for Task A-1 (F-measures for 10, 30, 50%)

The next is the results of Task A-2. Next, here are results for Task A-2 with content-based evaluation. First, we show the results for important-part summaries (abstract type 1), and for free summaries (abstract type 2)

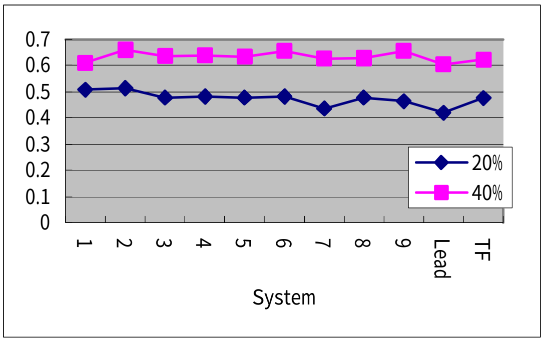

Figure 2 Evaluation Results for Task A-2-1 (abstract-type 1)

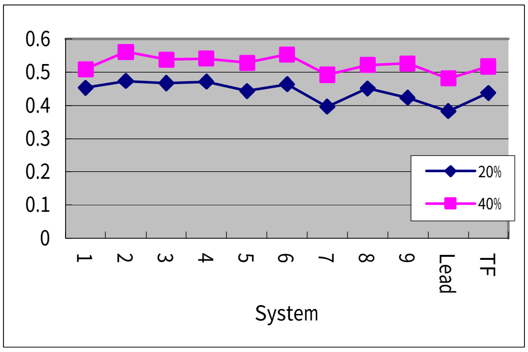

Figure 3 Evaluation Result for Task A-2-1 (abstract-type 2)

For Task A-2 with subjective evaluation, we had the following results. In Figure4, digits stand for summarization rates (20%, 40%), and R for readability, and C for content information.

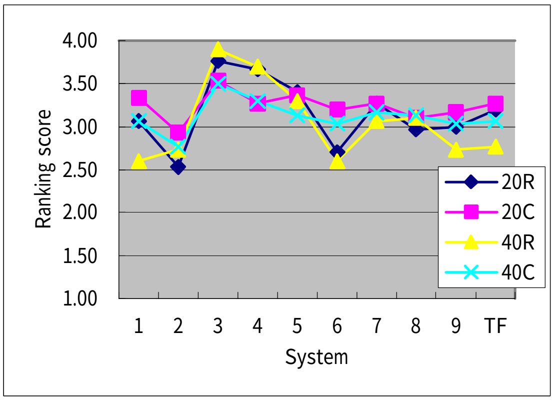

Figure 4 Evaluation Results for Task A-2-2

For Task B, we have the following results. First, the result for level-A and B documents

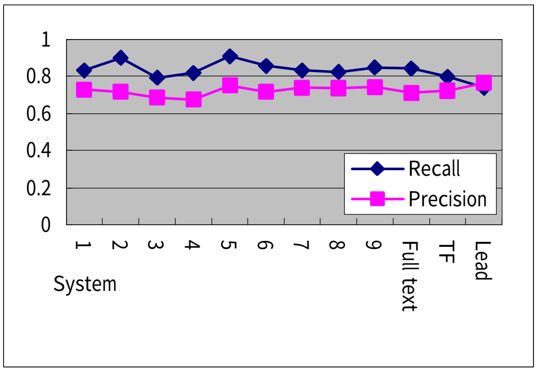

Figure 5 Task B results (level A documents)

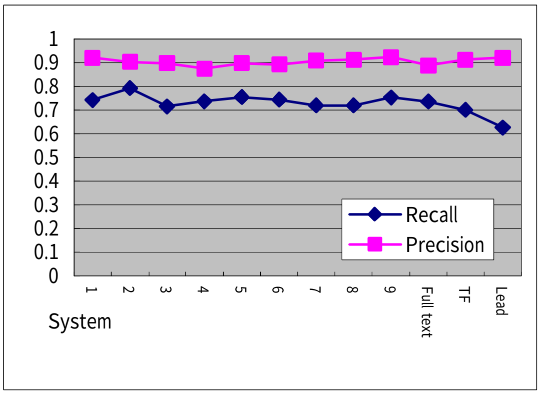

Figure 6 Task B results (level B documents)

And here is the result of time to finish the task. The scores are average time for conducting the IR task for 50 texts (one query) in minutes.

System	minutes
1	9:41
2	12:48
3	6:25
4	6:44
5	8:33
6	9:01
7	10:16
8	9:16
9	9:31
Full text	13:46
TF	8:44
Lead	7:32

Table 2 Time needed for Task B

Though the results are computed as scores, we do not consider they are definite. Described below as one of the features of TSC, we would like to focus on round-table evaluation at the workshop, which we plan to include in the final paper of the workshop.

9 Participant system technologies

The following table shows brief descriptions of the participating systems. It is based on the eight papers submitted to the workshop by each participating group. Please note that system number in the next table does not correspond to those in the other tables and figures in the paper or in the evaluation result section.

System	Reported methods and features
I	The system uses a scoring function integrating sentence location, sentence length, TF-IDF, similarity to the title.
II	The system uses word frequency, location, and content of the sentence (opinion or not) for extract type summaries. The system for abstract type summaries deletes multiple modifiers and illustration with paraphrasing technique.
III	Key concept is phrase-represented summarization. It selects core relations among words, and constructs phrases as summary
IV	Uses similarity information among original documents by hierarchical clustering and information gain ratio for Task B.
V	The system uses a hybrid method of term-frequency based method and lead sentence extraction.
VI	The approach is based on passage importance with Hanning window (density of words) for Task A-1 and Task B. For Task A-2, the system makes use of dependency structure analysis, TF-IDF and similarity to lead sentences.
VII	The system is a combination of the original system and lead-based method. The original system computes scores for ranking sentences and paragraphs based on relevance degrees between two sentences or paragraphs.
VIII	The approach utilizes keyword-based sentence extraction algorithm and thematic hierarchy based sentence extraction algorithm.

Table 3 System Technologies

10 Features of TSC evaluation

For Task A, we use several summarization rates ([9]), and prepared various lengths and genres of the texts and used them for evaluations. The lengths vary from 600, 900, 1200 and 2400 characters, and the genres include business news, social issues as well as editorials.

As for Task A-2, since it is difficult to have intrinsic evaluation for informative summaries, we will hold “round-table evaluation” where we just present the results of content-based and subjective evaluations, and offer all the participants the opportunity to discuss issues, including how to evaluate summaries, at NTCIR workshop 2.

11 Conclusions

We have described the outline of the Text Summarization Challenge. We are planning to continue our efforts for text summarization evaluation as TSC2. Though the full details are not yet decided, we would like to continue to produce human summaries of the same kind with varied lengths and genres of newspaper articles, and we like to conduct multiple-document evaluation in the second Challenge.

Acknowledgements

We would like to thank all the members of the organizing committee, and to give special thanks to all the people who helped us with the evaluations, and Dr. Kiyoaki Shirai at Tokyo Institute of Technology who is not an organizing member, however, helped us much.

References

[1] Proceedings of The Tipster Text Program Phase III, Morgan Kaufmann, 1999.
[2] Mani, I., et al. The TIPSTER SUMMAC Text Summarization Evaluation, Technical Report, MTR 98W0000138 The MITRE Corp. 1998.
[3] Sparck-Jones, K. “Automatic Summarizing: Factors and Directions” in Mani, I. And Maybury, M., editors, Advances in Automatic Text Summarization, pp.1-12. MIT Press, 1999.

[4] Mani, I. And Maybury, M., editors, Advances in Automatic Text Summarization, MIT Press, 1999.
[5] Paice, C. Constructing literature abstracts by computer: Techniques and prospects. Information Processing Management, 26(1) 1990.
[6] Edmundson, H.P. New methods in automatic abstracting. The Association for Computational Machinery, 16(2) 1969.
[7] Donaway, R.L., Drummey, K.W. and Mather, L.A, "A Comparison of Rankings Produced by Summarization Evaluation Measures", Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization. Pp. 69-78.
[8] http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/index.html
[9] Jing, H. and Barzilay, R. and McKeown, K. and Elhadad, M, "Summarization Evaluation Methods: Experiments and Analysis", pp. 51-59 in Intelligent Text Summarization Technical Report SS-98-06, AAAI Press 1998.