Text Summarization Challenge 2
Text summarization evaluation at NTCIR Workshop 3

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

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

Hidetsugu Nanba
Hiroshima City University
nanba@its.hiroshima-cu.ac.jp

Tsutomu Hirao
NTT Communication Science Laboratories, NTT Corp.
hirao@cslab.kecl.ntt.co.jp

Abstract

We report the outline of Text Summarization Challenge 2 (TSC2 hereafter), a sequel text summarization evaluation conducted as one of the tasks at the NTCIR Workshop 3. First, we describe briefly the previous evaluation, Text Summarization Challenge (TSC1) as introduction to TSC2. Then we explain TSC2 including the participants, the two tasks in TSC2, data used, evaluation methods for each task, and brief report on the results. Lastly we describe plans for the next evaluation, TSC3.

1 Introduction

As research on automatic text summarization is being 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 Document Understanding Conference (DUC) ([3]) in the United States show the need and importance of the evaluation for text summarization.

In Japan, Text Summarization Challenge (TSC1), a text summarization evaluation, the first of its kind, was conducted in the years of 1999 to 2000 as a part of the NTCIR Workshop 2. It was realized in order for the researchers in the field to collect and share text data for summarization, and to make clearer the issues of evaluation measures for summarization of Japanese texts ([4],[5],[6]). TSC1 used newspaper articles and had two tasks for a set of single articles with intrinsic and extrinsic evaluations. The first task (task A) was to produce summaries (extracts and free summaries) for intrinsic evaluations. We used recall, precision and F-measure for the evaluation of the extracts, and content-based as well as subjective methods for the evaluation of the free summaries. The summarization rates for task A were as follows: 10, 30, 50% for extracts and 20, 40% for free summaries.

The second task (task B) was to produce summaries for information retrieval (relevance judgment) task. The measures for evaluation were recall, precision and F-measure to indicate the accuracy of the task, as well as the time to indicate how long it takes to carry out the task.

We also prepared human-produced summaries including key data for the evaluation. In terms of genre, we used editorials and business news articles at TSC1's dryrun, and editorials and articles on social issues at the formal run evaluation.

In comparison, TSC2 uses newspaper articles and has two tasks (single- and multi-document summarization) for two types of intrinsic evaluations. In the following sections, we describe TSC2 in detail.

2 Participants

We had 4 participating systems for Task A, and 5 systems for Task B at dryrun. We have 8 participating systems for Task A and 9 systems for Task B at formal run. As group, we had 8 participating groups, which are all Japanese, of universities, governmental research institute or companies in Japan. Table 1 shows the breakdown of the groups.

3 Two Tasks in TSC2 and its Schedule

TSC2 has two tasks. They are single document summarization (task A) and multi-document summarization (task B).

Task A: We ask the participants to produce summaries in plain text to be compared with human-prepared summaries from single documents. 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%. This task is the same as task A-2 in TSC1.

Task B: In this task, more than two (multiple) documents are summarized for the task. Given a set of documents, which has been gathered for a pre-defined topic, the participants produce summaries of the set in plain text format. The information that was used to produce the document set, such as queries, as well as summarization lengths are given to the participants. Two summarization lengths are specified, short and long summaries for one set of documents.

The schedule of evaluations at TSC2 was as follows: dryrun was conducted in December 2001 and formal run was in May 2002. The final evaluation results were reported to the participants by early July 2002.

4 Data Used for TSC2

We use newspaper articles from the Mainichi newspaper database of 1998, 1999. 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%.

Abstract-type summaries:
We asked 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 idea of the articles (abstract-type type2). Both types of abstract-type summaries are used for task A. The summarization rates are 20% and 40%.

Both extract-type and abstract-type summaries are summaries from single articles.

Summaries from more than two articles:
Given a set of newspaper articles that has been selected based on a certain topic, the captioners produced free summaries (short and long summaries) for the set. Topics are various, from kidnapping case to Y2K problem.

5 Evaluation Methods for each task

We use summaries prepared by human as key data for evaluation. The same two intrinsic evaluation methods are used for both tasks. They are evaluation by ranking summaries and by measuring the degree of revisions. Here are the details of the two methods. We use 30 articles for task A and 30 sets of documents (30 topics) for task B at formal run evaluation.

5.1. Evaluation by ranking

This is basically the same as the evaluation method used for TSC1 task A-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.

1. Content: How much the system summary covers the important content of the original article.

2. Readability: How readable the system summary is.

The judges are given 4 types of summaries to be evaluated and rank them in 1 to 4 scale (1 is the best, 2 for the second, 3 for the third best, and 4 for the worst).

For task A, the first two types are human-produced abstract-type type1 and type2 summaries. The third is system results, and the fourth is summaries produced by lead method.

For task B, the first is human-produced free summaries of the given set of documents, and the second is system results. The third is the results of the first baseline system based on lead method where the first sentence of each document is used. The fourth is the results of the second baseline system using Stein method ([7]).

5.2. Evaluation by revision

It is a newly introduced evaluation method in TSC2 to evaluate the summaries by measuring the degree of revision to system results. The judges read the original documents and revise the system summaries in terms of the content and readability. The revisions are made by one of three

editing operations (insertion, deletion, replacement). The degree of the revision is computed based on the number of the operations and the number of revised characters.

As baseline for task A, human produced summaries (abstract type1 and abstract type 2) as well as lead-method results are used. And as baseline for task B, human produced summaries that are different from the key data, lead-method results, and the results based on the Stein method are used.

When more than half of the document needs to be revised, the judges can 'give up' revising the document.

6 Results

6.1. Results of Evaluation by ranking

Table 1 shows the result of evaluation by ranking for task A and Table 3 shows the result of evaluation by ranking for task B. Each score is the average of the scores for 30 articles for task A, and 30 topics for task B at formal run.

System No	Content 20%	Readability 20%	Content 40%	Readability 40%
F0101	2.53	2.87	2.60	2.77
F0102	2.67	2.97	2.50	2.77
F0103	2.80	2.93	2.90	2.90
F0104	2.77	2.73	2.80	2.90
F0105	2.70	2.73	2.60	2.77
F0106	2.73	2.57	2.63	2.67
F0107	2.70	2.60	2.50	2.53
F0108	2.40	2.83	2.60	2.77
TF	3.30	3.30	3.20	3.10
Human	2.33	2.20	2.10	2.03

Table 1 Ranking evaluation (task A)

In Table 1, 'TF' indicates a baseline system based on term-frequency method, and 'Human' indicates human-produced summaries that are different from the key data used in ranking judgement.

In Table 2 'Human' indicates human-produced summaries that are different from the key data used in ranking judgement.

System No	Content Short	Readability Short	Content Long	Readability Long
F0201	2.70	3.17	2.50	3.23
F0202	2.73	2.70	2.77	2.93
F0203	2.60	2.33	2.97	3.03
F0204	2.63	2.90	2.80	3.03
F0205	2.53	3.10	2.73	3.30
F0206	3.20	3.00	3.47	3.30
F0207	2.40	2.87	2.63	3.27
F0208	2.93	2.70	2.53	2.80
F0209	2.83	2.73	2.53	2.87
Human	2.00	2.17	1.83	2.33

Table 2 Ranking evaluation (task B)

6.2. Results of Evaluation by revision

Table 3 shows the result of evaluation by revision for task A at rate 40%, and Table 4 shows the result of evaluation by revision for task A at rate 20%. Table 5 shows the result of evaluation by revision for task B long, and Table 6 shows the result of evaluation by revision for task B short. All the tables show the evaluation results in terms of average number of revisions (editing operations) per document.

Please note that UIM stands for unimportant, RD for readability, IM for important, C for content in Tables 3 to 6. They mean the reason for the operations, e.g. 'unimportant' is for deletion operation due to the part judged to be unimportant.

System	Deletion		Insertion		Replacement	
	UIM	RD	IM	RD	C	RD
F0101	2.0	0.1	1.5	0.4	0.5	0.7
F0102	1.6	0.4	1.5	0.4	0.4	0.8
F0103	2.3	0.2	2.4	0.2	0.4	0.5
F0104	2.4	0.4	2.7	0.5	0.4	0.5
F0105	2.0	0.3	1.7	0.1	0.7	0.7
F0106	2.8	0.2	2.3	0.4	0.3	0.6
F0107	2.5	0.6	1.8	0.2	0.1	0.5
F0108	2.0	0.4	2.4	0.1	0.4	0.6
ld	2.9	0.1	0.7	0.1	0.4	0.1
free	0.4	0.4	1.2	0.4	0.1	0.3
part	0.7	0.6	0.9	0.3	0.1	0.4
edit	0.3	0.1	0.4	0.3	0.1	0.2
ALL	1.9	0.3	1.8	0.3	0.3	0.5

Table 3 Evaluation by revision (task A 40%)

System	Deletion		Insertion		Replacement	
	UIM	RD	IM	RD	C	RD
F0101	1.4	0.4	1.3	0.2	0.5	0.3
F0102	1.2	0.4	1.0	0.0	0.4	0.5
F0103	0.8	0.1	1.2	0.0	0.2	0.1
F0104	0.8	0.1	1.2	0.1	0.1	0.2
F0105	1.2	0.1	0.7	0.0	0.4	0.2
F0106	2.1	0.2	1.7	0.1	0.1	0.2
F0107	0.8	0.6	0.9	0.1	0.2	0.1
F0108	1.4	0.1	1.1	0.1	0.2	0.6
ld	1.9	0.1	1.3	0.0	0.0	0.0
free	0.6	0.4	1.1	0.1	0.2	0.1
part	0.7	0.3	1.1	0.1	0.1	0.2
edit	0.2	0.1	0.5	0.1	0.2	0.2
ALL	1.1	0.3	1.1	0.1	0.2	0.3

Table 4 Evaluation by revision (task A 20%)

In Table 3 and Table 4, ‘ld’ means a baseline system using lead method, ‘free’ is free summaries produced by human (abstract type 2), and ‘part’ is human-produced (abstract type1) summaries, and these three are baseline scores for task A.

System	Deletion		Insertion		Replacement	
	UIM	RD	IM	RD	C	RD
F0201	3.8	0.7	7.2	1.4	1.1	0.9
F0202	5.2	0.6	3.5	0.4	0.7	0.5
F0203	5.1	0.6	3.8	0.5	0.9	0.6
F0204	4.2	0.6	3.4	0.7	1.4	0.7
F0205	8.1	0.6	5.4	1.7	3.0	1.3
F0206	3.2	0.2	4.7	0.7	0.8	0.6
F0207	7.0	1.1	4.1	1.1	1.1	1.1
F0208	4.8	0.7	4.0	0.4	0.8	0.9
F0209	4.6	0.5	3.9	0.5	0.5	0.5
human	3.0	0.9	3.4	7.8	1.0	1.2
ld	5.7	0.9	2.9	0.4	0.7	0.5
stein	4.0	0.5	2.2	0.3	0.8	0.5
edit	3.0	1.2	2.9	0.7	0.7	1.1
ALL	4.9	0.7	4.0	1.3	1.1	0.8

Table 5 Evaluation by revision (task B long)

System	Deletion		Insertion		Replacement	
	UIM	RD	IM	RD	C	RD
F0201	3.5	0.5	4.3	0.8	1.1	0.7
F0202	3.5	0.4	2.4	0.2	0.7	0.2
F0203	3.6	0.3	2.8	0.2	0.5	0.4
F0204	2.7	0.5	2.3	0.2	1.2	0.7
F0205	5.5	0.4	2.5	0.8	2.0	0.7
F0206	2.0	0.4	3.4	0.6	0.4	0.4
F0207	3.5	0.4	2.7	0.3	0.6	0.6
F0208	2.4	0.5	2.3	0.4	0.2	0.3
F0209	2.5	0.5	2.2	0.2	0.3	0.4
human	1.9	0.8	2.4	2.0	0.9	0.7
ld	2.8	0.7	2.4	0.2	0.5	0.4
stein	3.0	0.3	1.8	0.2	0.4	0.3
edit	2.2	0.8	2.5	0.6	1.0	1.2
ALL	3.1	0.5	2.6	0.5	0.7	0.5

Table 6 Evaluation by revision (task B short)

In Table 5 and Table 6, 'human' means human-produced summaries which are different from the key data, and 'ld' means a baseline system using lead method, 'stein' means another baseline system using Stein method, and these three are baseline scores for task B.

To determine the plausibility of the judges' revision, the revised summaries were again evaluated with the evaluation methods in section 5. In Tables 3 to 6, 'edit' means the evaluation results for the revised summaries.

We also measure as degree of revision the number of revised characters for the three editing operations, and the number of documents that are given up revising by the judges.

7 Discussion

7.1. Discussion for Evaluation by ranking

We here further look into how the participating systems perform by analysing the ranking results in terms of differences in scores for content and those for readability.

First, task A. When we comapre evaluation results for content and readability of the same percentage, the readability scores tend to be higher than those for content, and it is especially clearer for 40% summarization. On the other hand, when the evaluation results of content 20% and 40%, and readability 20% and 40% are compared, the ranking scores for 20% summarization tend to be higher than those for 40%, and this is true with the baseline system and human summaries as well.

Next, task B. When the evaluation results of the same length are compared, he scores for readability tend to be higher, thence, the differences are in minus values, than those for content for both short and long summaries. In addition, the differences are larger than the differences we saw for task A. Again when the evaluation results of content short and long, readability short and long are checked, unlike task A, the scores for short summaries tend to be lower than those for long summaries. This tendency is very clear for the readability ranking scores.

Intuitively longer summaries can have better readability since they have more words to deal with. However, it is not the case with task B ranking results. Longer summaries had worse scores, especially in readability evaluation.

7.2. Discussion for Evaluation by revision

To determine the plausibility of the judges' revision, the revised summaries were again evaluated with the evaluation methods in section 5. As Tables 3 to 6 show, the degree of the revisions for the revised summaries is rather smaller than that for the original ones and is almost same as that for human summaries.

Tables 7 and 8 show the results of evaluation by ranking for the revised summaries at task A and B respectively. They show that the scores for the revised summaries are rather smaller than those for the original ones and are almost same as those for human summaries. From these results, the quality of the revised summaries is considered as same as that of human summaries.

System No	Content 20%	Readability 20%	Content 40%	Readability 40%
edit	2.37	2.43	2.33	2.33

Table 7 Ranking evaluation (task A)

Sytem No	Content Short	Readability Short	Content Long	Readability Long
edit	1.93	2.23	2.13	2.50

Table 8 Ranking evaluation (task B)

8. Plans for TSC3

We would like to describe breifly the plans for the third evaluation (TSC3). The outline of the task will be as follows (see our web page [4] for more details):

(1) Task: Multiple Document Summarization

The participants produce two kinds of summaries from sets of documents which are considered to be relevant to queries. This task is basically the continuation of TSC2 task B.

What is given to the participants:

- document sets
- titles of the document sets
- sets of questions about important information of the document sets
- summary lengths (2 kinds)

What the participants submit:

- two kinds of summaries

Subtask (optional task): Sentence extraction from the document sets
The participants select relevant sentences and delete redundant sentences in them. (cf. TREC Novelty Track)

What is given to the participants:

- in addition to the above, the number of sentences to be extracted(2 kinds)

What the participants submit:

- result of evaluation which is described below in evaluation 1.

We think multiple-document summarization system needs at least the following:
1. important sentence extraction technique
2. technique to measure the degree of closeness (or redundancy) of the extracted sentences
3. technique of shortening the sentences after deleting the redundant information.
(The subtask aims at evaluating 1. and 2.)

(2) Evaluation

1. Coverage, Precision for systems which produce extracts in making summaries.
We will provide the scoring tool and the human-produced extracts. This is the evaluation for the subtask. The participants should submit the evaluation results by the specified date.

2. Intrinsic evaluation
a. Content evaluation
b. Readability evaluation
Subjective evaluation based on quality questions. (cf. DUC 2002)

3. Extrinsic evaluation
The participants measure how much the system summaries include the passages which are answers to the given sets of questions. We will make available the human-selected passages which are answers to the sets of questions and provide the scorer. The participants should submit the evaluation results by the specified date.

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] http://www-nlpir.nist.gov/projects/duc/.

[4] http://oku-gw.pi.titech.ac.jp/tsc/index-en.html.

[5] Takahiro Fukusima and Manabu Okumura, "Text Summarization Challenge –Text Summarization Evaluation at NTCIR Workshop 2", In Proceedings of NTCIR Workshop 2, pp.45-50, 2001.

[6] Takahiro Fukusima and Manabu Okumura, "Text Summarization Challenge – Text Summarization Evaluation in Japan", North American Association for Computational Linguistics (NAACL2001), Workshop on Automatic Summarization, pp.51-59, 2001.

[7] Gees C. Stein, Tomek Strazalkowski and G. Bowden Wise, "Summarizing Multiple Documents using Text Extraction and Interactive Clustering", Pacific Association for Computational Linguistics, pp.200-208, 1999.