Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

Text Summarization Challenge 3 –Text Summarization Evaluation at NTCIR Workshop 4–

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

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

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

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

Abstract

In this paper, we describe the overview of Text Summarization Challenge 3 (TSC3 hereafter), sequel test summarization evaluation conducted as one of the tasks at the NTCIR Workshop 4.
We have defined two kinds of task in TSC3; extracting important sentences and abstracting from multiple documents. We have prepared 30 document sets which concerned with certain topic and conducted Formal run evaluation with nine participants.
keywords *Automatic Summarization, Summarization Evaluation, extrinsic evaluation, intrinsic evaluation*

1 Introduction

Automatic text summarization has attracted a lot of attention recently and there have been many studies in this field. There is a particular need to establish methods for the automatic summarization of multiple documents rather than single documents.

There have been several evaluation workshops on text summarization. In 1998, TIPSTER SUMMAC [4] took place and the Document Understanding Conference (DUC)[1] has been held annually since 2001. DUC has included multiple document summarization among its tasks since the first conference. The Text Summarization Challenge (TSC)[2] has been held once in one and a half years as part of the NTCIR (NII-NACSIS Test Collection for IR Systems) project since 2001. Multiple document summarization was included for the first time as one of the tasks at TSC2 (in 2002) [7]. Multiple document summarization is now a central issue for text summarization research.

[1] http://duc.nist.gov
[2] http://www.lr.pi.titech.ac.jp/tsc

In TSC3, we have defined two kinds of task; extracting important sentences and abstracting from multiple documents. We have prepared 30 document sets which concerned with certain topic and conducted Formal run evaluation with nine participants.

2 Task Definition

We briefly explain the task definition of TSC3. The tasks are similar to the TSC2 Task B. The first task is extraction, *i.e.*, the system extracts important sentences from given document sets. The second is abstraction, *i.e.*, the system generate summaries whose number of characters was less than a fixed number.

We give the participants followings:

- document sets (30 sets),
- titles of document sets,
- length of extract and abstract (2 kinds).

As the target documents, we use both Mainichi and Yomiuri newspapers published between 1998 and 1999.

We believe that multiple document summarization system need following:

important sentence extraction,

redundant sentence reduction,

rewrite the result of extraction to reduce the size of the summary to the specified number of characters or less.

The extraction task evaluate sentence extraction and reduction and the abstraction task evaluate techniques to reduce the summary size.

3 Evaluation Methods

We employ both *intrinsic* and *extrinsic* evaluation. For extraction, we define "Precision" and "Coverage." For abstraction, we use subjectivity evaluation and pseudo question-answering.

3.1 *Intrinsic* Metrics for Extraction

Multiple document summarization from multiple sources, *i.e.*, several newspapers concerned with the same topic but with different publishers, is more difficult than single document summarization since it must deal with more text (in terms of numbers of characters and sentences). Moreover, it is peculiar to multiple document summarization that the summarization system must decide how much redundant information should be deleted[3] .

In a single document, there will be few sentences with the same content. In contrast, in multiple documents with multiple sources, there will be many sentences that convey the same content with different words and phrases, or even identical sentences. Thus, a text summarization system needs to recognize such redundant sentences and reduce the redundancy in the output summary.

However, we have no way of measuring the effectiveness of such redundancy in the corpora for DUC and TSC2. Key data in TSC2 was given as abstracts (free summaries) whose number of characters was less than a fixed number and, thus, it is difficult to use for repeated or automatic evaluation, and for the extraction of important sentences. Moreover, in DUC, where most of the key data were abstracts whose number of words was less than a fixed number, the situation was the same as TSC2. At DUC 2002, extracts (important sentences) were used, and this allowed us to evaluate sentence extraction. However, it is not possible to measure the effectiveness of redundant sentences reduction since the corpus was not annotated to show sentence with same content. In addition, this is the same even if we use the SummBank corpus [9].

In any case, because many of the current summarization systems for multiple documents are based on sentence extraction, we believe these corpora to be unsuitable as sets of documents for evaluation.

On this basis, in TSC3, we assumed that the process of multiple document summarization consists of the following three steps, and we produce a corpus for the evaluation of the system at each of the three steps[4] .

Step 1 Extract important sentences from a given set of documents

Step 2 Minimize redundant sentences from the result of Step 1

Step 3 Rewrite the result of Step 2 to reduce the size of the summary to the specified number of characters or less.

We have annotated not only the important sentences in the document set, but also those among them that have the same content. These are the corpora for steps 1 and 2. We have prepared human-produced free summaries (abstracts) for step 3.

3.1.1 Number of Sentences System Should Extract

We begin with guidelines for annotating important sentences (extracts). We think that there are two kinds of extract.

1. A set of sentences that human annotators judge as being important in a document set [1, 11, 8].

2. A set of sentences that are suitable as a source for producing an abstract, *i.e.*, a set of sentences in the original documents that correspond to the sentences in the abstracts[**?**, 10, 5, 3].

When we consider how summaries are produced, it seems more natural to identify important segments in the document set and then produce summaries by combining and rephrasing such information than to select important sentences and revise them as summaries. Therefore, we believe that second type of extract is superior and thus we prepared the extracts in that way.

However, as stated in the previous section, with multiple document summarization, there may be more than one sentence with the same content, and thus we may have more than one set of sentences in the original document that corresponds to a given sentence in the abstract; that is to say, there may be more than one key datum for a given sentence in the abstract[5] .

we have two sets of sentences that correspond to sentence a in the abstract.

(1) s_1 of document x, or
(2) a combination of s_2 and s_3 of document y

This means that s_1 alone is able to produce a, and a can also be produced by combining s_2 and s_3.

We marked all the sentences in the original documents that were suitable sources for producing the sentences of the abstract, and this made it possible for us to determine whether or not a summarization system deleted redundant sentences correctly at Step 2. If

[3] It is true that we need other important techniques such as those for maintaining the consistency of words and phrases that refer to the same object, and for making the results more readable; however, they are not included here.

[4] This is based on general ideas of a summarization system and is not intended to impose any conditions on a summarization system.

[5] We use 'set of sentences' since we often find that more than one sentence corresponds to a sentence in the abstract.

Table 1. Important Sentence Data.

Sentence ID of Abstract	Set of Corresponding Sentences
1	$\{s_1\}_\sqcup\{s_{10},s_{11}\}$
2	$\{s_3,s_5,s_6\}$
3	$\{s_{20},s_{21},s_{23}\}_\sqcup\{s_1,s_{30},s_{60}\}$

the system outputs the sentences in the original documents that are annotated as corresponding to the same sentence in the abstract, it has redundancy. If not, it has no redundancy. Returning to the above example, if the system outputs s_1,s_2,and s_3, they all correspond to sentence a in the abstract, and thus it is redundant.

3.1.2 Precision

Precision is the ratio of how many sentences in the system output are included in the set of the corresponding sentences. It is defined by the following equation.

$$\text{Precision} = \frac{m}{h}, \tag{1}$$

where h is the least number of sentences needed to produce the abstract by solving the constraint satisfaction problem and m is the number of 'correct' sentences in the system output, *i.e.*, the sentences that are included in the set of corresponding sentences. For example, the sentences listed in Table 1 are 'correct.' If the system output is "$s_{10}, s_{11}, s_5, s_{17}, s_{60}, s_{61}$", then the Precision is as follows:

$$\text{Precision} = \frac{4}{6} = 0.667. \tag{2}$$

for "$s_1, s_{10}, s_{11}, s_3, s_5, s_{60}$", the Precision is as follows:

$$\text{Precision} = \frac{6}{6} = 1. \tag{3}$$

3.1.3 Coverage

Coverage is an evaluation metric for measuring how close the system output is to the abstract taking into account the redundancy found in the set of sentences in the output.

The set of sentences in the original documents that corresponds correctly to the i-th sentence of the human-produced abstract is denoted here as $A_{i,1}, A_{i,2}, \cdots, A_{i,j}, \cdots, A_{i,\ell}$. In this case, we have ℓ sets of corresponding sentences. Here, $A_{i,j}$ indicates a set of elements each of which corresponds to the sentence number in the original documents, denoted as $A_{i,j} = \{\theta_{i,j,1}, \theta_{i,j,2}, \cdots, \theta_{i,j,k}, \cdots\}$. For instance, from Table 1, $A_{1,2} = \theta_{1,2,1}, \theta_{1,2,2}$ and $\theta_{1,2,1} = s_{10}, \theta_{1,2,2} = s_{11}$.

Then, we define the evaluation score $e(i)$ for the i-th sentence in the abstract as equation (1).

$$e(i) = \max_{1 \le j \le \ell} \left(\frac{\sum_{k=1}^{|A_{i,j}|} v(\theta_{i,j,k})}{|A_{i,j}|} \right), \tag{4}$$

where $v(\alpha)$ is defined by the following equation.

$$v(\alpha) = \begin{cases} 1 & \text{if the system outputs } \alpha \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Function e returns 1 (one) when any $A_{i,j}$ is outputed completely. Otherwise it returns a partial score according to the number of sentences $|A_{ij}|$.

Given function e and the number of sentences in the abstract n, Coverage is defined as follows:

$$\text{Coverage} = \frac{\sum_{i=1}^{n} e(i)}{n}. \tag{6}$$

If the system extracts "s_{10},s_{11},s_5,s_{17},s_{60},s_{61}", $e(i)$ is computed as follows:

$$\begin{aligned} e(1) &= \max(0,1) = 1 \\ e(2) &= \max(0.33) = 0.33 \\ e(3) &= \max(0,0.33) = 0.33 \end{aligned}$$

and its Coverage is 0.553. If the system extracts "s_1,s_{10},s_{11},s_3,s_5,s_{60}", then the Coverage is 0.780.

$$\begin{aligned} e(1) &= \max(1,1) = 1 \\ e(2) &= \max(0.67) = 0.67 \\ e(3) &= \max(0,0.67) = 0.67 \end{aligned}$$

3.2 *Intrinsic* Metrics for Abstraction

We employ subjectivity evaluation for both content information and readability of summaries.

3.2.1 Content Evaluation

Human judges match summaries they produced with system results at sentence level, and evaluate the results based on the degree of the matching (how well they match). The sentences in the human-produced summaries have values that show the degree of importance, and these values are taken into account in coming up with final evaluation.

3.2.2 Readability Evaluation

We use "Quality Questions (QQ)" for readability evaluation. We modified DUC's QQ for Japanese text. There are sixteen questions as following:

q00 How many redundant or unnecessary sentences are there?

q01 How many places are there where (zero) pronouns or referring expressions to be used?
q02 How many pronouns are there whose antecedents are missing?
q03 How many proper nouns which appeared in the unsuitable position are there?
q04 How many expressions which have same meanings but different term are there?
q05 How many of the sentences are missing important constituents?
q06 How many places are there where conjunctions should be supplied or conjunctions should be deleted?
q07 How many unnecessary words (adverbs, adjectives, etc.) are there?
q08 Does the summary has wrong chronological ordering?
q09 How many sentences which should unify writing style (polite style or ordinaly style) are there?
q10 How many redundant verbs are there?
q11 How many sentences which has wrong concord expression are there?
q12 How many sentences have incorrect word order?
q13 How many incorrect inflection words are there?
q14 How many complex sentences are there that had better be divided?
q15 How many sentences are there that had better be unified?

3.3 *extrinsic* Metrics for Abstraction

Sometimes question-answering (QA) by human subjects is used for evaluation [6, 2]. That is, human subjects judge whether predefined questions can be answered by reading only a machine generated summary. However, the cost of this evaluation is huge. Therefore, we employ a pseudo question-answering evaluation, *i.e.*, whether a summary has an 'answer' to the question or not. The background to this evaluation is inspired by TIPSTER SUMMAC's QA track [4].

3.3.1 Pseudo Question-Answering

Sometimes question-answering (QA) by human subjects is used for evaluation [6, 2]. That is, human subjects judge whether predefined questions can be answered by reading only a machine generated summary. However, the cost of this evaluation is huge. Therefore, we employ a pseudo question-answering evaluation, *i.e.*, whether a summary has an 'answer' to the question or not. The background to this evaluation is inspired by TIPSTER SUMMAC's QA track [4].

For each document set, there are about five questions for a short summary and about ten questions for long summary. Note that the questions for the short summary are included in the questions for the long summary. Examples of questions for the topic "Release of SONY's AIBO" are as follows: "How much is AIBO?", "When was AIBO sold?", and "How many AIBO are sold?".

Now, we evaluate the summary from the 'exact match' and 'edit distance' for each question. 'Exact match' is a scoring function that returns one when the summary includes the answer to the question. 'Edit distance' measures whether the system's summary has strings that are similar to the answer strings. The score $S_{\mathrm{e}d}$ based on the edit distance is normalized with the length of the sentence and the answer string so that the range of the score is [0,1]:

$$\mathrm{S_{ed}} = \frac{\text{length of the sentence} - \text{edit distance}}{\text{length of the answer strings}}. \quad (7)$$

The score for a summary is the maximum value of the scores for sentences in the summary. The score is 1 if the summary has a sentence that includes the whole answer string.

It should be noted that the presence of answer strings in the summary does not mean that a human subject can necessarily answer the question.

Table 2. System-ID

SYS-ID	organization name
SOUKEN	The Graduate University for Advanced Studies
CRLNYU	Communications Research Laboratory / New York University
smlab	Toyohashi University of Technology
MOGS	The University of Tokyo
forest	Yokohama National University
KLEIR	Pohang University of Science & Technology
DBLAB	Hokkaido University
UEC	The University of Electro-Communications
UYDI	Ritsumeikan University

Table 3. Task Participation.

ID	abstraction	extraction
SOUKEN	$*$	$**$
CRLNYU	$*$	$**$
smlab	$*$	$*$
MOGS	$*$	$*$
forest	$*$	$*$
KLEIR	$*$	–
DBLAB	$*$	$*$
UEC	$*$	$*$
UYDI	$*$	$*$

4 Task Participants

In TSC3, there were nine participants. Table 2 show the system-id and the organization name. Table 3 show the task participation. The symbol "$*$" indicates that the participant submitted one result to the task. The two symbols denotes two kinds of submitted results.

Table 4. Evaluation Results (Extraction)

ID	Short		Long	
	Cov.	Prec.	Cov.	Prec.
SOUKEN(a)	0.315	0.494	0.355	0.554
SOUKEN(b)	0.372	0.591	0.363	0.587
CRLNYU(a)	0.222	0.314	0.313	0.432
CRLNYU(b)	0.293	0.378	0.295	0.416
smlab	0.328	0.496	0.327	0.535
MOGS	0.283	0.406	0.341	0.528
forest	0.329	0.567	0.391	0.680
DBLAB	0.308	0.505	0.339	0.585
UEC	0.181	0.275	0.218	0.421
UYDI	0.251	0.476	0.247	0.547
LEAD	0.212	0.426	0.259	0.539

Table 5. Results on Content Evaluation.

ID	Short	Long
SOUKEN	0.228	0.214
CRLNYU	0.188	0.240
smlab	0.247	0.258
MOGS	0.230	0.248
forest	0.291	0.323
KLEIR	0.222	0.210
DBLAB	0.207	0.247
UEC	0.131	0.233
UYDI	0.197	0.221
LEAD	0.160	0.159
HUMAN	0.385	0.402

5 Evaluation Results

5.1 Results on extraction

Table 4 show the results of extraction. The score is the average scores for 30 document sets. "LEAD" denotes the baseline system based on lead-based method. All methods have lower Coverage scores than Precision scores. This means that the extracted sentences include redundant ones. In addition, we know that "Lead" is a good extraction method for newspaper articles; however, this is not true for the TSC3 corpus.

5.2 Results on abstraction

5.2.1 Content Metrics

Table 5 show the results on content evaluation by human subject. "HUMAN" indicates human-produced summaries that are different from model summaries.

Table 7. Results for Pseudo Question-Answering.

ID	Short		Long	
	exact	edit	exact	edit
SOUKEN	0.394	0.677	0.399	0.706
CRLNYU	0.257	0.556	0.266	0.602
smlab	0.367	0.653	0.356	0.677
MOGS	0.342	0.614	0.327	0.630
forest	0.439	0.710	0.442	0.751
KLEIR	0.321	0.601	0.313	0.611
DBLAB	0.390	0.684	0.356	0.633
UEC	0.133	0.427	0.201	0.549
UYDI	0.304	0.579	0.308	0.628
LEAD	0.300	0.589	0.275	0.602
HUMAN	0.461	0.716	0.426	0.721

5.2.2 Readability Metrics

Table 6 show the results on readability evaluation using "Quality Questions."

5.2.3 Pseudo Question-Answering Metrics

Table 7 show the results on pseudo question-answering.

6 Conclusion

We described the outline of TSC3. We defined two kinds of task using Mainichi and Yomiuri newspapers published between 1998 and 1999. We reported the results of two tasks.

References

[1] T. Fukusima and M. Okumura. Text Summarization Challenge: Text Summarization Evaluation in Japan. In *Proc. of the NAACL 2001 Workshop on Automatic summarization*, pages 51–59, 2001.

[2] T. Hirao, Y. Sasaki, and H. Isozaki. An Extrinsic Evaluation for Question-Biased Text Summarization on QA tasks. In *Proc. of the NAACL 2001 Workshop on Automatic Summarization*, pages 61–68, 2001.

[3] H. Jing and K. McKeown. The Decomposition of Human-Written Summary Sentences. *Proc. of the 22nd ACM-SIGIR*, pages 129–136, 1999.

[4] I. Mani, G. Klein, D. House, L. Hirschman, T. Firman, and B. Sundheim. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68, 2002.

[5] D. Marcu. The Automatic Construction of Large-scale Corpora for Summarization Research. *Proc. of the 22nd ACM-SIGIR*, pages 137–144, 1999.

[6] A. H. Morris, G. M. Kasper, and D. Adams. The Effects and Limitations of Automatic Text Condensing

Table 6. Results on Readability Evaluation.

Short

ID	q00	q01	q02	q03	q04	q05	q06	q07	q08	q09	q10	q11	q12	q13	q14	q15
SOUKEN	0.333	1.033	0.600	0.333	2.333	0.900	0.633	0.933	−0.567	0.500	1.767	0.100	0.000	0.000	0.100	0.133
CRLNYU	0.033	0.567	0.700	0.667	1.567	1.400	0.500	0.267	−0.500	0.100	0.367	0.033	0.000	0.033	0.000	0.100
smlab	0.200	1.333	0.533	0.333	3.067	0.467	0.733	1.067	0.000	0.033	2.467	0.000	0.000	0.033	0.067	0.233
MOGS	0.067	0.700	0.433	0.300	2.433	0.933	0.933	0.500	−0.133	0.100	1.267	0.000	0.000	0.033	0.067	0.100
forest	0.700	0.633	1.200	0.600	2.367	1.267	0.767	0.567	−0.300	0.200	0.967	0.067	0.000	0.000	0.067	0.100
KLEIR	0.100	1.067	0.433	0.400	2.433	0.500	0.567	0.867	0.200	0.267	1.633	0.100	0.000	0.000	0.067	0.100
DBLAB	0.167	1.100	0.133	0.300	1.433	0.667	0.667	0.867	0.133	0.033	1.867	0.000	0.000	0.033	0.033	0.133
UEC	1.967	0.200	1.767	0.400	0.633	3.800	1.333	0.167	−0.600	0.233	0.000	0.200	0.033	0.000	0.000	0.133
UYDI	0.167	1.233	0.767	0.267	2.567	2.800	0.600	0.667	−0.600	0.567	1.833	0.000	0.000	0.067	0.033	0.233
LEAD	1.500	1.267	0.267	0.267	1.667	0.067	0.767	1.533	0.267	0.067	1.667	0.000	0.000	0.033	0.033	0.200
HUMAN	0.033	0.267	0.000	0.000	0.433	0.400	0.400	0.000	0.933	0.500	0.033	0.000	0.000	0.033	0.033	0.033

Long

ID	q00	q01	q02	q03	q04	q05	q06	q07	q08	q09	q10	q11	q12	q13	q14	q15
SOUKEN	0.200	1.600	1.133	0.433	5.533	2.100	1.000	1.900	−0.833	0.733	2.967	0.500	0.000	0.000	0.067	0.300
CRLNYU	0.300	0.500	2.100	0.333	2.667	3.600	1.500	0.467	−0.900	0.133	0.233	0.000	0.067	0.000	0.033	0.233
smlab	0.433	1.700	0.933	0.100	5.133	1.300	1.100	1.800	−0.500	0.167	3.900	0.033	0.033	0.033	0.033	0.333
MOGS	0.167	1.367	0.833	0.400	4.200	1.100	1.100	1.300	−0.533	0.133	3.000	0.133	0.000	0.033	0.100	0.200
forest	1.100	1.200	1.367	0.633	4.667	1.233	1.133	1.233	−0.633	0.167	3.067	0.033	0.000	0.033	0.033	0.300
KLEIR	0.333	1.567	1.067	0.533	4.567	1.000	0.933	1.967	−0.133	0.200	3.300	0.000	0.000	0.000	0.033	0.100
DBLAB	0.367	1.500	0.600	0.467	2.833	2.033	0.967	1.567	−0.267	0.067	3.533	0.100	0.000	0.067	0.033	0.233
UEC	2.367	0.133	2.500	0.500	2.567	4.667	1.833	0.333	−0.733	0.333	0.300	0.233	0.067	0.067	0.000	0.233
UYDI	0.700	1.433	1.400	0.400	5.133	4.500	0.767	1.967	−0.933	0.933	4.367	0.067	0.033	0.033	0.033	0.400
LEAD	2.833	2.100	0.633	0.367	4.300	0.300	1.033	4.333	−0.333	0.067	5.133	0.000	0.000	0.000	0.067	0.433
HUMAN	0.033	0.167	0.100	0.000	1.133	0.467	0.433	0.067	0.800	0.567	0.000	0.033	0.000	0.000	0.033	0.100

on Reading Comprehension. *Information System Research*, 3(1):17–35, 1992.

[7] M. Okumura, T. Fukusima, and H. Nanba. Text Summarization Challenge 2, Text Summarization Evaluation at NTCIR Workshop 3. In *Proc. of the HLT/NAACL 2003 Text Summarization Workshop*, pages 49–56, 2003.

[8] C. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[9] D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu, and E. Drabek. Evaluation challenges in large-scale document summarization. In *Proc. of the 41st ACL*, pages 375–382, 2003.

[10] S. Teufel and M. Moens. Sentence Extraction as a Classification Task. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 58–65, 1997.

[11] K. Zechner. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proc. of the 16th COLING*, pages 986–989, 1996.