Corpus and Evaluation Measures for Multiple Document Summarization with Multiple Sources

Tsutomu HIRAO
NTT Communication Science Laboratories
hirao@cslab.kecl.ntt.co.jp

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

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

Chikashi NOBATA
Communication Research Laboratories
nova@crl.go.jp

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

Abstract

In this paper, we introduce a large-scale test collection for multiple document summarization, the Text Summarization Challenge 3 (TSC3) corpus. We detail the corpus construction and evaluation measures. The significant feature of the corpus is that it annotates not only the important sentences in a document set, but also those among them that have the same content. Moreover, we define new evaluation metrics taking redundancy into account and discuss the effectiveness of redundancy minimization.

1 Introduction

It has been said that we have too much information on our hands, forcing us to read through a great number of documents and extract relevant information from them. With a view to coping with this situation, research on 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 (Mani et al., 2002) 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) (Okumura et al., 2003). Multiple document summarization is now a central issue for text summarization research.

In this paper, we detail the corpus construction and evaluation measures used at the Text Summarization Challenge 3 (TSC3 hereafter), where multiple document summarization is the main issue. We also report the results of a preliminary experiment on simple multiple document summarization systems.

2 TSC3 Corpus

2.1 Guidelines for Corpus Construction

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

[1] http://duc.nist.gov

[2] http://www.lr.pi.titech.ac.jp/tsc

[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.

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 (Radev et al., 2003).

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.

In TSC3, since we have key data (a set of correct important sentences) for steps 1 and 2, we conducted automatic evaluation using a scoring program. We adopted an *intrinsic* evaluation by human judges for step 3, which is currently under evaluation. We provide details of the extracts prepared for steps 1 and 2 and their evaluation measures in the following sections. We do not report the overall evaluation results for TSC3.

2.2 Data Preparation for Sentence Extraction

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 (Fukusima and Okumura, 2001; Zechner, 1996; Paice, 1990).

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

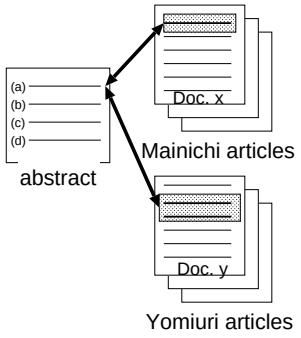

Figure 1: An example of an abstract and its sources.

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(Kupiec et al., 1995; Teufel and Moens, 1997; Marcu, 1999; Jing and McKeown, 1999).

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 (Figure 1).

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 the system outputs the sentences in the original documents that are annotated as corresponding to the same sentence in the abstract, it

[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	$\{s_1\}$␣$\{s_{10},s_{11}\}$
2	$\{s_3,s_5,s_6\}$
3	$\{s_{20}, s_{21},s_{23}\}$␣$\{s_1,s_{30},s_{60}\}$

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 Evaluation Metrics

We use both *intrinsic* and *extrinsic* evaluation. The *intrinsic* metrics are "Precision", "Coverage" and "Weighted Coverage." The *extrinsic* metric is "Pseudo Question-Answering."

3.1 *Intrinsic* Metrics

3.1.1 Number of Sentences System Should Extract

Precision and Recall are generally used as evaluation matrices for sentence extraction, and we used the PR Breaking Point (Precision = Recall) for the evaluation of extracts in TSC1 (Fukusima and Okumura, 2001). This means that we evaluate systems when the number of sentences in the correct extract is given. Moreover, in TSC3 we assume that the number of sentences to be extracted is known and we evaluate the system output that has the same number of sentences.

However, it is not as easy to decide the number of sentences to be extracted in TSC3 as in TSC1. We assume that there are correspondences between sentences in original documents and their abstract as in Table 1. An ASCII space, "␣", is the delimiter for the sets of corresponding sentences in the table. As shown in the table, we often see several sets of sentences that correspond to a sentence in the abstract in multiple document summarization.

An 'extract' here is a set of sentences needed to produce the abstract. For instance, we can obtain 'extracts' such as "s_1,s_3,s_5,s_6,s_{30},s_{60}", and "s_{10},s_{11},s_3,s_5,s_6,s_{20},s_{21},s_{23}" from Table 1 [6]. Often there are several 'extracts' and we must determine which of these is the best. In such cases, we define the 'correct extract' as the set with the least number of sentences needed to produce the abstract because it is desirable to convey the maximum amount of information with the least number of sentences.

Finding the minimum set of sentences to produce the abstract amounts to solving the constraint satisfaction problem. In the example in Table 1, we obtain the following constraints from each sentence in the abstract:

- $C_1 = s_1 \vee (s_{10} \wedge s_{11})$,
- $C_2 = s_3 \wedge s_5 \wedge s_6$,
- $C_3 = (s_{20} \wedge s_{21} \wedge s_{23}) \vee (s_1 \wedge s_{30} \wedge s_{60})$

With these conditions, we now find the minimum set that makes all the conjunctions true. We need to find the minimum set that makes $C_1 \wedge C_2 \wedge C_3 = true$. In this case, the minimum cover is $\{s_1, s_3, s_5, s_6, s_{30}, s_{60}\}$, and so the system should extract six sentences.

In TSC3, we computed the number of sentences that the system should extract and then evaluated the system outputs, which must have the same number of sentences, with the following precision and coverage.

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

[6] In fact, it is possible to produce the abstract with other sentence combinations.

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).

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

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

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

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:

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

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.

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

3.1.4 Weighted Coverage

Now we define 'Weighted Coverage' since each sentence in TSC3 is ranked A, B or C, where "A" is the best. This is similar to "Relative Utility" (Radev et al., 2003). We only use three ranks in order to limit the ranking cost. The definition is obtained by modifying equation (6).

$$\text{W.C.} = \frac{\sum_{i=1}^{n} w(r(i)) \times e(i)}{w(A)n(A) + w(B)n(B) + w(C)n(C)}, \quad (7)$$

where $r(i)$ denotes the ranking of the i-th sentence of the abstract and $w(r(i))$ is its weight. $n(rank)$ is the number of sentences whose ranking is $rank$ in the abstract. Suppose the first sentence is ranked A, the second B, and the third C in Table 1, and their weights are given as $w(A) = 1$,$w(B) = 0.5$ and $w(C) = 0.3$ [7].

As before, if the system extracts "s_{10},s_{11},s_5,s_{17},s_{60},s_{61}", then the Weighted Coverage is computed as follows:

$$\text{W.C.} = \frac{1 \times 1 + 0.5 \times 0.33 + 0.3 \times 0.33}{1 \times 1 + 0.5 \times 1 + 0.3 \times 1} = 0.689. \quad (8)$$

3.2 *Extrinsic* Metrics

3.2.1 Pseudo Question-Answering

Sometimes question-answering (QA) by human subjects is used for evaluation (Morris et al., 1992; Hirao et al., 2001). 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 (Mani et al., 2002).

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_{ed} 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]:

$$\text{S}_{\text{ed}} = \frac{\text{length of the sentence} - \text{edit distance}}{\text{length of the answer strings}}. \quad (9)$$

The score for a summary is the maximum value of the scores for sentences in the summary. The

[7] $w(r(i))$ may be computed differently. It is 1/rank (one divided by rank) here.

Table 2: Description of TSC3 Corpus.

# of doc. sets	30
# of articles (The Mainichi)	175
# of articles (The Yomiuri)	177
Total	352
# of Sentences	3587

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.

4 Preliminary Experiment

In order to examine whether our corpus is suitable for summarization evaluation, our evaluation measures significant information and redundancies in the system summaries.

Below we provide the details of the corpus, evaluation results and effectiveness of the minimization of redundant sentences.

4.1 Description of Corpus

According to the guidelines described in section two, we constructed extracts and abstracts of thirty sets of documents drawn from the Mainichi and Yomiuri newspapers published between 1998 to 1999, each of which was related to a certain topic. First, we prepared abstracts (their sizes were 5% and 10% of the total number of the characters in the document set), then produced extracts using the abstracts. Table 2 shows the statistics.

One document set consists of about 10 articles on average, and the almost same number of articles were taken from the Mainichi newspaper and the Yomiuri newspaper. Most of the topics are classified into a single-event according to McKeown (2001).

The following list contains all the topics.

0310 Two-and-half-million-year old new hominid species found in Ethiopia.
0320 Acquisition of IDC by NTT (and C&W).
0340 Remarketing of game software judged legal by Tokyo District Court.
0350 Night landing practice of carrier-based aircrafts of the Independence.
0360 Simultaneous bombing of the US Embassies in Tanzania and Kenya.
0370 Resignation of President Suharto.
0380 Nomination of Mr. Putin as Russian prime minister.
0400 Osama bin Laden provided shelter by Taliban regime in Afghanistan.
0410 Transfer of Nakata to A.C. Perugia.
0420 Release of Dreamcast.
0440 Existence of Japanese otter confirmed.
0450 Kyocera Corporation makes Mita Co. Ltd. its subsidiary.
0460 Five-story pagoda at Muroji Temple damaged by typhoon.
0470 Retirement of aircraft YS-11.
0480 Test observation of astronomical telescope 'Subaru' started.
0500 Dolly the cloned sheep.
0510 Mass of neutrinos.
0520 Human Genome Project finishes decoding of the 22nd chromosome.
0530 Peace talks in Northern Ireland at the end of 1999.
0540 Debut of new model of bullet train (700 family).
0550 Mr. Yukio Aoshima decides not to run for gubernatorial election.
0560 Mistakes in entrance examination of Kansai University.
0570 Space shuttle Endeavour, from its launch to return.
0580 40 million-year-old fossil of new monkey species found by research group at Kyoto University.
0590 Dead body of George Mallory found on Mt. Everest.
0600 Release of SONY's AIBO.
0610 e-one, look-alike of iMac.
0630 Research on Kitora tomb resumes.
0640 Tidal wave damage generated by earthquake in Papua New Guinea.
0650 Mistaken bombing of the Chinese embassy by NATO.

4.2 Compared Extraction Methods

We used the lead-based method, the TF·IDF-based method (Zechner, 1996) and the sequential pattern-based method (Hirao et al., 2003), and compared performance of these summarization methods on the TSC3 corpus.

Lead-based Method

The documents in a test set were sorted in chronological and ascending order. Then, we extracted a sentence at a time from the beginning of each document and collected them to form a summary.

TF·IDF-based Method

The score of a sentence is the sum of the significant scores of each content word in the sentence. We therefore extracted sentences in descending order of importance score. The sentence score $\mathrm{S}_{\mathrm{tfidf}}(s_i)$ is defined by the following.

$$\mathrm{S}_{\mathrm{tfidf}}(s_i) = \sum_{t \in s_i} w(t, DS), \tag{10}$$

where $w(t, DS)$ is defined as follows:

$$w(t, DS) = tf(t, DS) \cdot \log \frac{|DB|}{df(t)}. \tag{11}$$

$tf(t, DS)$ is the frequency of word t in the document set, $df(t)$ is the document frequency of t, and $|DB|$ is the total number of documents in the set. In fact, we computed these using all the articles published in the Mainichi and Yomiuri newspapers for the years 1998 and 1999.

Sequential Pattern-based Method

The score of a sentence is the sum of the significant scores of each sequential pattern in the sentence. The patterns used for scoring were decided

Table 3: Evaluation results for "Precision", "Coverage" and "Weighted Coverage."

Method	Length	Prec.	Cov.	W.C.
Lead	Short	.426	.212	.326
	Long	.539	.259	.369
TF·IDF	Short	.497	.292	.397
	Long	.604	.325	.434
Pattern	Short	.613	.305	.403
	Long	.665	.298	.418

Table 4: Evaluation results for "Pseudo Question-Answering."

Method	Length	Exact	Edit
Lead	Short	.300	.589
	Long	.275	.602
TF·IDF	Short	.375	.643
	Long	.393	.659
Pattern	Short	.390	.644
	Long	.370	.640

by using a statistical significance test such as the χ^2 metric test and using 1,000 patterns. This is an extension of Lin's method (Lin and Hovy, 2000). The sentence score $\mathrm{S_{pat}}(s_i)$ is defined by the following.

$$\mathrm{S_{pat}}(s_i) = \sum_{p \in s_i} w(p), \tag{12}$$

where $w(p)$ is defined as follows:

$$w(p) = \frac{\log(f(p, DS) + 1) \cdot \log(\frac{|AS|}{f(p, AS)})}{len(p)}. \tag{13}$$

$f(p, DS)$ is the sentence frequency of pattern p in the document set and $f(p, AS)$ is the sentence frequency of pattern p in all topics. $|AS|$ is the number of sentences in all topics and $len(p)$ is the pattern length.

4.3 Evaluation Result

Table 3 shows the *intrinsic* evaluation result. All methods have lower Coverage and Weighted Coverage scores than Precision scores. This means that the extracted sentences include redundant ones. In particular, the difference between "Precision" and "Coverage" is large in "Pattern."

Although both "Pattern" and "TF·IDF" outperform "Lead," the difference between them is small. In addition, we know that "Lead" is a good extraction method for newspaper articles; however, this is not true for the TSC3 corpus.

Table 4 shows the *extrinsic* evaluation results. Again, both "Pattern" and "TF·IDF" outperform "Lead", but the difference between them is small. We found a correlation between the *intrinsic* and *extrinsic* measures.

Table 5: Effects of clustering ("Precision", "Coverage", "Weighted Coverage").

Method	Length	Prec.	Cov.	W.C.
TF·IDF	Short	.430	.297	.377
	Long	.533	.345	.455
Pattern	Short	.531	.289	.390
	Long	.620	.338	.456

Table 6: Effects of clustering (Pseudo Question-Answering).

Method	Length	Exact	Edit
TF·IDF	Short	.401	.650
	Long	.377	.648
Pattern	Short	.392	.650
	Long	.380	.655

4.4 Effect of Redundant Sentence Minimization

The experiment described in the previous section shows that a group of sentences extracted in a simple way includes many redundant sentences. To examine the effectiveness of minimizing redundant sentences, we compare the Maximal Marginal Relevance (MMR) based approach (Carbonell and Goldstein, 1998) with the clustering approach (Nomoto and Matsumoto, 2001). We use 'cosine similarity' with a bag-of-words representation for the similarity measure between sentences.

Clustering-based Approach

After computing importance scores using equations (10) and (12), we conducted hierarchical clustering using Ward's method until we reached h (see Section 3.1.1) clusters for the first $3h$ sentences. Then, we extracted the sentence with the highest score from each cluster.

Table 5 shows the results of the *intrinsic* evaluation and Table 6 shows the results of the *extrinsic* evaluation. By comparison with Table 3, the clustering-based approach resulted in TF·IDF and Pattern scoring low in Precision, but high in Coverage. When comparing Table 4 with Table 6, the score is improved in most cases. These results imply that redundancy minimization is effective for improving the quality of summaries.

MMR-based Approach

After computing importance scores using equations (10) and (12), we re-ranked the first $3h$ sentences by MMR and extracted the first h sentences.

Table 7 and 8 show the *intrinsic* and *extrinsic* evaluation results, respectively. We can see the effectiveness of redundancy minimization by MMR. Notably, in most cases, there is a large improvement in both the *intrinsic* and *extrinsic* evaluation results as compared with clustering.

Table 7: Effects of MMR ("Precision", "Coverage", "Weighted Coverage").

Method	Length	Prec.	Cov.	W.C.
TF·IDF	Short	.469	.306	.403
	Long	.565	.376	.475
Pattern	Short	.469	.332	.429
	Long	.577	.377	.500

Table 8: Effects of MMR (Pseudo Question-Answering).

Method	Length	Exact	Edit
TF·IDF	Short	.386	.647
	Long	.405	.667
Pattern	Short	.417	.663
	Long	.390	.656

These results show that redundancy minimization has a significant effect on multiple document summarization.

5 Conclusion

We described the details of a corpus constructed for TSC3 and measures for its evaluation, focusing on sentence extraction. We think that a corpus in which important sentences and those with the same content are annotated for multiple documents is a new and significant feature for summarization corpora.

It is planned to make the TSC3 corpus available (even if the recipient is not a TSC3 participant) by exchanging memoranda with the National Institute of Informatics in Japan. We sincerely hope that this corpus will be useful to researchers who are interested in text summarization and serve to facilitate further progress in this field.

References

J. Carbonell and J. Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21th ACM-SIGIR*, pages 335–336.

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

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

T. Hirao, J. Suzuki, H. Isozaki, and E. Maeda. 2003. Multiple Document Summarization using Sequential Pattern Mining (in Japanese). In *The Special Interest Group Notes of IPSJ (NL-158-6)*, pages 31–38.

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

J. Kupiec, J Petersen, and F. Chen. 1995. A Trainable Document Summarizer. In *Proc. of the 18th SIGIR*, pages 68–73.

C-Y. Lin and E. H. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proc. of the 16th COLING*, pages 495–501.

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

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

K. McKeown, R. Barzilay, D. Evans, V. Hatzivassilogou, M. Y. Kan, B. Schiffman, and S. Teufel. 2001. Columbia Multi-Document Summarization: Approach and Evaluation. In *Proc. of the Document Understanding Conference 2001*.

A. H. Morris, G. M. Kasper, and D.A. Adams. 1992. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension. *Information System Research*, 3(1):17–35.

T. Nomoto and M. Matsumoto. 2001. A New Approach to Unsupervised Text Summarization. In *Proc. of the 24th ACM-SIGIR*, pages 26–34.

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

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

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

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

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