A Multiple-Document Summarization System with User Interaction

Hiroyuki SAKAI
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku,
Toyohashi 441-8580,
Japan,
sakai@smlab.tutkie.tut.ac.jp

Shigeru MASUYAMA
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku,
Toyohashi 441-8580,
Japan,
masuyama@tutkie.tut.ac.jp

Abstract

We propose a multiple-document summarization system with user interaction. Our system extracts keywords from sets of documents to be summarized and shows the keywords to a user on the screen. Among them, the user selects some keywords reflecting his/her needs. Our system controls the produced summary by using these selected keywords. For evaluation of our method, we participated in TSC3 of NTCIR4 workshop by letting our system select 12 best keywords regarding scoring by the system. Our participated system attained the best performance in content evaluation among systems not using sets of questions. Moreover, we evaluated effectiveness of user interaction in our system. With user interaction, our system attained both higher coverage and precision than that without user interaction.

1 Introduction

Recent rapid progress of computer and communication technologies enabled us to access enormous amount of machine-readable information easily. However, this has caused the information overload problem. In order to solve this problem, automatic summarization methods have been studied (Mani and T.Maybury, 1999). In particular, the necessity for a multiple-document summarization has been increasing and the multiple-document summarization technology has been intensively studied recently (Mani, 2001).

In this paper, we define multiple-document summarization as a process for producing a summary from a relevant document set. Such a document set may be very large and may contain a number of topics. It is preferable that a summary produced by a multiple-document summarization system from the document set covers all topics contained in the document set. However, it is difficult to produce a summary that covers all the topics in the document set with a small number of characters. For example, a document set relevant to "releasing AIBO" contains some topics, e.g., what is AIBO?, how to sell AIBO?, etc. Moreover, sentences recognized as important sentences considerably differ person to person (Nomoto and Matsumoto, 2001). This is because "summarization need", i.e., topics a different person wants to read, may differ. Hence, we propose a multiple-document summarization system with user interaction for coping appropriately with user's summarization need. Our system extracts keywords from a document set to be summarized and shows the keywords to a user. Among them, the user selects keywords reflecting user's summarization need. Our system controls a produced summary by using the keywords selected by the user. For realizing our purpose, we have devised a scoring method for keywords extraction specialized to our purpose. We would like to emphasize here the fact that scoring of words for extracting keywords shown to a user is crucial for the system performance as well as different from those used in usual automatic indexing.

We participated in TSC3 (Text Summarization Challenge - 3) of NTCIR4 workshop [1] and attained the best performance in content evaluation among systems not using sets of questions. Note that our system participated in TSC3 is an automatic summarization system without user interaction by letting our system with user interaction select 12 best keywords regarding scoring by the system. Moreover, we evaluated effectiveness of user interaction and that with user interaction attained both higher coverage and precision than that without user interaction.

2 Feature of our multiple-document summarization system

Our multiple-document summarization system proposed in this paper is different from previ-

[1] http://www.lr.pi.titech.ac.jp/tsc/index-en.html

ously proposed multiple-document summarization methods (see, e.g.,(Barzilay et al., 1999), (Mani and Bloedorn, 1999), (Goldstein et al., 2000), (Ando et al., 2000), (Lin and Hovy, 2002), (Nobata and Sekine, 2002), (Hirao et al., 2003)) in that: (1) Our system can produce a summary coping appropriately with each user's summarization need by letting a user select keywords reflecting user's summarization need. (2) The keywords are extracted automatically from a document set to be summarized by calculating a score to each noun contained in the document set. The formula to calculate scores consists of not only frequency of nouns and document frequency used in $tf \cdot idf$ but also distribution of nouns in the document set and location of nouns in documents or the document set. The reason why such factors are used will be explained in the next section. (3) Our system deletes redundant adnominal verb phrases in sentences to reduce the number of characters in a sentence. The deletable adnominal verb phrases are decided statistically by using entropy based on a probability that verbs modify noun, etc. Our previous method (Sakai and Masuyama, 2002) adjusted to multiple-document summarization so that more deletable adnominal verb phrases are recognized, is used in this system.

The interactive summarization system has been introduced for the first time by (Saggion and Lapalme, 2002). The system proposed in (Saggion and Lapalme, 2002) is based on shallow syntactic and semantic analysis, conceptual identification and text re-generation, while, our system is based on a statistical method.

3 The method to extract relevant keywords

A relevant document set S to be summarized may be regarded as a document set obtained by a hypothetical query from the entire document set Δ to be considered. In TSC3, the entire document set consists of newspaper articles, Mainichi newspaper and Yomiuri newspaper, Japanese daily newspapers, from January 1 to December 31, 1998, 1999. We explain a method to extract keywords relevant to such a hypothetical query from document set S. Here, we define such keywords as relevant keywords t_i, $i = 1, 2, \ldots, k$. We assign scores to nouns contained in document set S and nouns assigned a large score are extracted as relevant keywords. A large score is assigned if a noun fulfills the following four conditions.

1. The noun that appears frequently in the document set S to be summarized.
2. The noun that appears uniformly in each document $d \in S$.
3. The noun that appears in the beginning of a document (i.e., the 1st sentence) and in the beginning of the document set in chronological order (i.e., the 1st document).
4. The noun that does not appear frequently in entire document set Δ.

Our method for extracting relevant keywords consists of the following two steps.

Step 1: Calculate score $W(t_i, S)$ of noun t_i $(i = 1, \ldots, n)$ contained in document set S.

Step 2: Extract the nouns with k largest score $W(t_i, S)$ as relevant keywords. □

The score $W(t_i, S)$ is calculated by formula 1.

$$\begin{aligned} W(t_i, S) &= (0.5 + \frac{Tf(t_i, S)}{max_{i=1,\ldots,n} Tf(t_i, S)}) \\ &\times (0.5 + \frac{En(t_i, S)}{max_{i=1,\ldots,n} En(t_i, S)}) \\ &\times max_{d \in S} \frac{1 + nl(d) - nlf(t_i, d)}{nl(d)} \\ &\times max_{d \in S} \frac{1 + |S| - rt(t_i, d)}{|S|} \\ &\times idf(t_i, \Delta) \end{aligned} \quad (1)$$

where,

$Tf(t_i, S)$: frequency of noun t_i contained in document set S. This is calculated by formula 2.

$$Tf(t_i, S) = \sum_{d \in S} tf(t_i, d) \quad (2)$$

where, $tf(t_i, d)$ is a frequency of noun t_i in document d.

$En(t_i, S)$: entropy based on the probability that noun t_i appears in document $d \in S$. This is calculated by formula 3 to be introduced later.

$nl(d)$: the number of sentences in document $d \in S$.

$nlf(t_i, d)$: the line number of a sentence containing noun t_i for the first time in document $d \in S$.

$rt(t_i, d)$: the document number of document d containing noun t_i for the first time in document set S in chronological order.

$idf(t_i, \Delta)$**:** idf(Baeza-Yates and Ribeiro-Neto, 1999) value assigned to noun t_i in entire document set Δ.

$En(t_i, S)$ is an entropy based on a probability that noun t_i appears in document $d \in S$. For example, $En(t_i, S)$ assigned to noun t_i contained only in one document $d \in S$ is 0. Though such noun t_i may be an important noun for document d, it may be an irrelevant noun for document set S. Hence, noun t_i that is assigned small entropy value should not be extracted as a relevant keyword. However, a noun that appears uniformly in each document contained in document set S has a large entropy value. $En(t_i, S)$ is calculated by formula 3.

$$En(t_i, S) = -\sum_{d \in S} P(t_i, d) \log_2(P(t_i, d)) \quad (3)$$

$$where,\ P(t_i, d) = \frac{tf(t_i, d)}{Tf(t_i, S)} \quad (4)$$

The 3rd term in formula 1 is to assign a large value to a noun appearing in the beginning of a document. The 4th term in formula 1 is to assign a large value to a noun appearing in the beginning of a document set in chronological order. The reason why these members are included is that the 1st sentence in the 1st document frequently contains important information (see, e.g.,(Nobata and Sekine, 2002)).

4 The method to extract important sentences

The method to extract important sentences measures similarity between a sentence and the set of relevant keywords selected by a user, and extracts sentences assigned large similarity as important sentences. The similarity is calculated as cosine metric between a vector of a sentence and a vector of the set of relevant keywords. If the same noun as relevant keywords is contained frequently in a sentence, the cosine metric assigned to the sentence has a large value. The method to extract important sentences is summarized as follows: Here, we define relevant keywords shown to a user as keyword set K and define relevant keywords selected by a user as keyword set U.

Step 1: Re-calculate score of relevant keywords t_i's by the following formula. Here, we define the number of keywords shown to a user to be k.

$$W'(t_i, S) = \begin{cases} (1 + 0.5k)W(t_i, S), & t_i \in U \\ W(t_i, S), & \text{otherwise} \end{cases} \quad (5)$$

Step 2: Generate relevant keyword vector $\mathbf{V_K}$ consisting of $W'(t_i, S)$ $(i = 1, \ldots, k)$ assigned to each relevant keyword $(t_i \in K)$.

$$\mathbf{V_K} = (W'(t_1, S),\ W'(t_2, S), \ldots,\ W'(t_k, S))$$

Step 3: Generate sentence vector $\mathbf{V_s}$ consisting of $W'(t_j, S)$ $(j = 1, \ldots, m)$ assigned to each noun contained in sentence s $(t_j \in s)$.

$$\mathbf{V_s} = (W'(t_1, S),\ W'(t_2, S), \ldots,\ W'(t_m, S))$$

Step 4: Calculate a cosine metric between vector $\mathbf{V_K}$ and vector $\mathbf{V_s}$ as similarity $sim(s, K)$ by formula 6.

$$sim(s, K) = \frac{\mathbf{V_K} \cdot \mathbf{V_s}}{|\mathbf{V_K}||\mathbf{V_s}|} \quad (6)$$

Step 5: Extract the sentences with m largest similarity $sim(s, K)$ as important sentences and output these m sentences in chronological order. □

In document set S, the 1st sentence contained in the 1st document containing important sentences in chronological order is always adopted as an important sentence in order to improve the readability.

5 The method to delete redundant information

In the multiple-document summarization, it is necessary to measure the degree of closeness of contents in extracted sentences (or documents) and to delete redundant information. This is because, the documents including the same contents may exist in a document set to be summarized. Our multiple-document summarization system identifies close sentences in extracted important sentences set and close documents in the document set, and deletes redundant information contained therein.

First, redundant information contained in the sentences set is deleted as follows.

Step 1: Measure the difference $d(s_1, s_2)$ between cosine metric $sim(s_1, K)$ assigned to sentence s_1 and $sim(s_2, K)$ assigned to sentence s_2.

$$d(s_1, s_2) = |sim(s_1, K) - sim(s_2, K)| \quad (7)$$

Step 2: If $d(s_1, s_2)$ has a value smaller than a threshold value, delete sentence s_i having a smaller cosine metric $sim(s_i, K)$.

We determined the threshold value to be 0.0001 in Step 2. This is a sufficiently small value to regard contents of s_1 identical to contents of s_2. Next, redundant information contained in the document set is deleted as follows. Here, we define a set of important sentences contained in document d_i as sd_i. The method is as follows.

Step 1: Generate vector $\mathbf{V_{sd_1}}$, consisting of $W'(t_i, S)$ $(i = 1, \ldots, n)$ assigned to nouns contained in sd_1.

$$\mathbf{V_{sd_1}} = (W'(t_1, S), W'(t_2, S), \ldots, W'(t_n, S))$$

Step 2: Generate vector $\mathbf{V_{sd_2}}$, consisting of $W'(t_j, S)$ $(j = 1, \ldots, m)$ assigned to nouns contained in sd_2.

$$\mathbf{V_{sd_2}} = (W'(t_1, S), W'(t_2, S), \ldots, W'(t_m, S))$$

Step 3: Calculate a cosine metric between vector $\mathbf{V_{sd_1}}$ and vector $\mathbf{V_{sd_2}}$ as similarity $sim(sd_1, sd_2)$.

Step 4: If $sim(sd_1, sd_2)$ has a value larger than a threshold value, delete document d_i $(i = 1$ *or* $2)$ having a smaller score $W(sd_i)$ (sd_i is in d_i). Score $W(sd_i)$ is calculated by the following formula 8.

$$W(sd_i) = \sum_{s \in sd_i} sim(s, K) \tag{8}$$

Here, documents d_1 and d_2 are newspaper articles issued on the same day. We determined the threshold value to be 0.85 in Step 4 by trial and error using sample data provided by the organizer of TSC3. Note that this sample data has not used in the formal run as a document set to be summarized. Note that if d_i is deleted, sentences contained in document d_i are not extracted and the important sentences extracted by our system are changed. Hence, our system executes this algorithm to delete documents and the algorithm to extract important sentences iteratively until no document is deleted by this algorithm.

6 The method to reduce the number of characters in a sentence

Our system deletes redundant adnominal verb phrases in sentences to reduce the number of characters in a sentence. We define adnominal verb phrases as phrases that modify a noun and include a verb modifying the noun. For example, in the case of "*SONY ga kaihatsu shita aibo*(ソニーが開発したアイボ: the AIBO developed by SONY", "*SONY ga kaihatsu shita*(ソニーが開発した: developed by SONY)" is an adnominal verb phrase, which modifies noun "*aibo*(アイボ: AIBO)". Here, the adnominal verb phrase "*SONY ga kaihatsu shita*(ソニーが開発した: developed by SONY)" may be deleted if a user has known that AIBO was developed by SONY. We define an adnominal verb phrase modifying a noun n as $VP(n)$. Redundant adnominal verb phrases are deleted by an improved method of (Sakai and Masuyama, 2002) proposed by us in order to apply to multiple documents summarization. For more details, please refer to reference (Sakai and Masuyama, 2002) [2]. The method is as follows.

Step 1: Calculate score $endf(n)$ to assign to noun n modified by adnominal verb phrase $VP(n)$ by formula 9.

Step 2: Calculate score $W(VP(n), s)$ for adnominal verb phrase $VP(n)$ by formula 12.

Step 3: Delete adnominal verb phrase $VP(n)$ if the score $endf(n)$ has a value smaller than threshold value $\theta(endf(n))$ and the score $W(VP(n), s)$ has a value smaller than threshold value $\theta(W(VP(n), s))$. □

We decided threshold value $\theta(endf(n))$ as 0.7 and threshold value $\theta(W(VP(n), s))$ as 8.7 in Step 3. These threshold values are decided by experiments with training corpus not to be summarized in the experiments. Score $endf(n)$ expresses the degree of modifier necessity of noun n and is calculated by formula 9.

$$endf(n) = \frac{1 + H(n)}{idf(n, \Delta)} \tag{9}$$

Here, $H(n)$ is an entropy based on a probability that verbs modify noun n. It reflects "frequency of modification of noun n by adnominal verb phrases", "variety of adnominal verb phrases modifying noun n". $H(n)$ is calculated by formula 10:

$$H(n) = -\sum_{v \in V(n)} P(v, n) \log_2(P(v, n)) \tag{10}$$

[2] The method of deleting adnominal verb phrases proposed in (Sakai and Masuyama, 2002) attained precision 79.3%.

$$P(v,n) = \frac{f(v,n)}{\sum_{v \in V(n)} f(v,n)} \quad (11)$$

where,

$V(n)$**:** set of verbs contained in adnominal verb phrases modifying noun n in entire document set Δ,

$f(v,n)$**:** frequency of verb v modifying noun n in entire document set Δ.

Next, $W(VP(n),s)$ is calculated by formula 12.

$$W(VP(n),s) = \frac{NM(n)IM(VP(n),s)}{0.5 + 0.5CV(n,s)} \quad (12)$$

$$NM(n) = 0.5 + \frac{endf(n)}{J(n)} \quad (13)$$

where,

$IM(VP(n),s)$**:** a factor to reflect rating of context in adnominal verb phrase $VP(n)$ contained in sentence s.

$CV(n,s)$**:** the number of occurrences of noun n modified by adnominal verb phrases from the 1st sentence in the 1st document to sentence s in document $d \in S$ in document set S in chronological order.

$J(n)$**:** the number of common nouns contained in noun n if noun n is a compound noun.

The $IM(VP(n),s)$ is calculated by formula 14.

$$IM(VP(n),s) = 0.5 + R \sum_{c \in VP(n)} I(c,s) \quad (14)$$

$$I(c,s) = \frac{W'(c,S)}{0.5 + 0.5CT(c,s)} \quad (15)$$

where,

R**:** the number of segments composing adnominal verb phrase $VP(n)$,

$W'(c,S)$**:** the score calculated by formula 5 to noun c contained in adnominal verb phrase $VP(n)$.

$CT(c,s)$**:** the number of occurrences of noun c contained in adnominal verb phrases from the 1st sentence in the 1st document to sentence s in document $d \in S$ in document set S in chronological order.

We introduced $CV(n,s)$ in formula 12 and $CT(c,s)$ in formula 15 in order to recognize more deletable adnominal verb phrases than our previous method applied directly to multiple-document summarization.

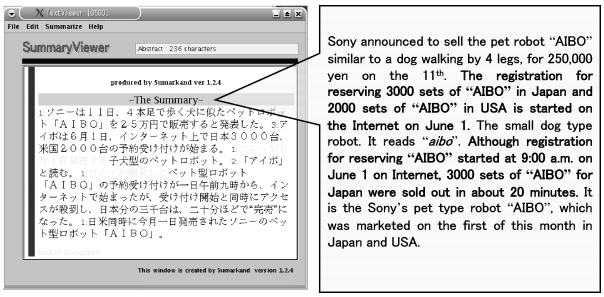

Figure 1: A summary produced by our system

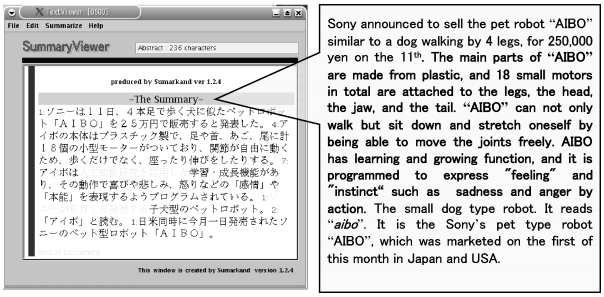

Figure 2: A summary produced by changing relevant keywords

7 Implementation

We implemented our method and developed a multiple-document summarization system. We employed JUMAN [3] as a morphological analyzer, and KNP[4] as a parser. We show a summary produced by our system in Figure 1. The document set to be summarized contains 9 documents relevant to "releasing AIBO" and the summary consists of less than 236 characters. Moreover, we show a summary in Figure 2 when a user selects keywords relevant to the movement and performance of AIBO (e.g., "人工知能 (artificial intelligence)") and deletes keywords relevant to the way to sell (e.g., "予約 (Reservation)"). Comparing Figure 1 with Figure 2, we can make sure that summaries have been changed by keywords selected by a user.

8 Evaluations of our system in TSC3

We participate in TSC3 (Text Summarization Challenge - 3) of NTCIR4 workshop for evaluation of information access technologies. The purpose of TSC3 is to evaluate performance of automatic multiple-document summarization that summarizes newspaper articles from two sources (Mainichi newspaper and Yomiuri newspaper from January 1 to December 31, 1998,

[3] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[4] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

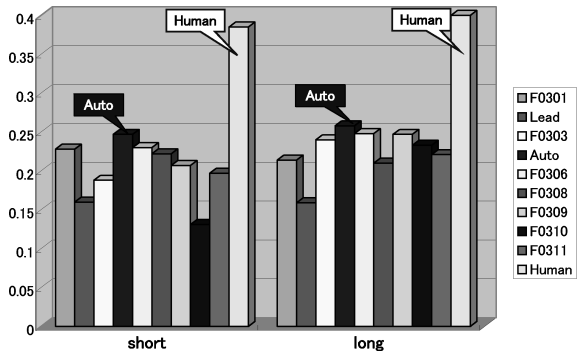

Figure 3: Content evaluation

1999.) Our system participated in TSC3 is not a system with user interaction for realizing automatic multiple documents summarization. Hence, we define the following execution of our system to be "Auto" for realizing an automatic multiple-document summarization system without user interaction.

Auto: The execution of our system where 12 best keywords regarding scoring by the system are selected.

The number of keywords selected by the system is determined by trial and error using sample data provided by the organizer of TSC3. The main evaluation method of TSC3 is :

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 at the final evaluation [5].

8.1 Evaluation results of TSC3

The result of content evaluation is shown in Figure 3 [6]. Here, "AUTO" shows our system that participated in TSC3. "Lead" is the lead method, a baseline method. In TSC3, we are given the sets of questions about important information of the document sets by the organizer of TSC3. Note that these sets of questions are produced from summaries made by human as correct data. (For example: when will AIBO be released ? etc.) Here, we exclude evaluation results of a system that uses the sets of questions for producing summaries of multiple documents [7]. The reason is as follows. As mentioned above, the sets of questions are produced from summaries made by human as correct data. Hence, we consider that using the sets of questions as machine-readable information for producing summaries is not realistic. Moreover, we consider that comparing systems using the sets of questions with systems not using them by ranking is unfair.

By the result shown in figure 3, our system that implemented "AUTO" has attained the best performance among the systems not using the sets of questions.

8.2 Evaluation of user interaction

Our system is essentially a multiple-document summarization system with user interaction. Hence, we evaluate effectiveness of user interaction of our system in this subsection. For evaluating it, we consider the following execution of our system:

Interaction: Execution of our system where relevant keywords contained in the set of questions are selected, and relevant keywords not contained in the set of questions are deleted.

The "Interaction" simulates user interaction on our system. (i.e., we regard the set of questions mentioned at the beginning of Sec.8.1 as user's summarization need. Since the set of questions produced from summaries by human (i.e., user), we will be able to regard the questions as user's summarization need.) The coverage and precision of "Interaction" is shown in Figure 4. Moreover, the coverage and precision of "Auto" and "Lead" are shown for comparison. Here, the coverage and precisions which take redundancy into account are obtained by using the scoring tool provided for the subtask in TSC3.

9 Discussion

From the result shown in figure 3, our system participating in TSC3 as "Auto" attained the best performance among the systems not using the sets of questions. We think the reason why the good performance was attained is that the first 12 keywords extracted from a document set to be summarized by scoring by our method

[5] http://www.lr.pi.titech.ac.jp/tsc/cfp3/task_description_e.html

[6] About 383 characters are involved in a summary of "short" and about 742 characters are involved in a summary of "long".

[7] In TSC3, systems not using the sets of questions and systems using them were evaluated together.

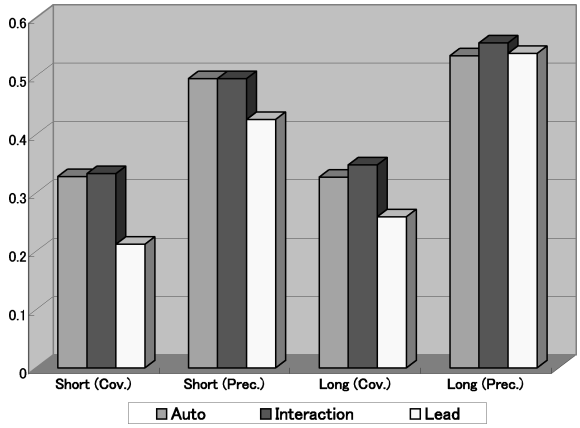

Figure 4: Evaluation of user interaction

were appropriate. Sentences extracted by using keywords irrelevant to the document set may not probably be important.

From the result shown in figure 4, we conclude that the "Interaction" is more effective than the "AUTO". Moreover, the effectiveness of user interaction in the case of "long" is more remarkable than that of "short". The reason why the effectiveness of user interaction in the case of "long" is more remarkable is as follows. In the case of "short", our system has to extract sentences fewer than that of "long". Even if a user had changed relevant keywords to use for sentence extraction, the sentences extracted by our system are not necessarily changed in the case of "short". However, the extracted sentences are greatly changed in the case of "long" when a user had changed relevant keywords. Hence, we consider that sentences are extracted well by changing relevant keywords in the case of "long".

Acknowledgment

This work was supported in part by The 21st Century COE Program "Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology of Japan and The Grant-in-Aid from the Japan Society for the Promotion of Science.

References

R. Ando, B. Boguraev, R. Byrd, and M. Neff. 2000. Multi-document summarization by visualizing topical content. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 79–88.

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

R. Barzilay, K. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computatonal Linguistics*, pages 550–557.

J. Goldstein, V. Mittal, J. Carbonel, and M. Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 40–48.

T. Hirao, K. Takeushi, H. Isozaki, Y. Sasaki, and E. Maeda. 2003. Svm-based multi-document summarization integrating sentence extraction with *bunsetsu* eliminate. *IEICE Trans. on Information and Systems*, E86-D(9):1702–1709.

C-Y. Lin and E. Hovy. 2002. From single to multi-document summarizaton: A prototype system and its evaluation. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, pages 457–464.

I. Mani and E. Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67.

I. Mani and M. T.Maybury. 1999. *Advances in Automatic Text Summarization*. the MIT Press.

I. Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Company.

C. Nobata and S. Sekine. 2002. A summarization system with categorization of document sets. In *Working Notes of the Third NTCIR Workshop Meeting*, pages 33–38.

T. Nomoto and Y. Matsumoto. 2001. An experimental comparison of supervised and unsupervised approaches to text summarization. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 630–632.

H. Saggion and G. Lapalme. 2002. Generating indicative-informative summaries with sumum. *Computational Linguistics*, 28(4):497 – 526.

H. Sakai and S. Masuyama. 2002. Unsupervised knowledge acquisition about the deletion possibility of adnominal verb phrases. In *Proceedings of Workshop on Multilingual Summarization and Question Answering 2002 (post-conference workshop to be held in conjunction with COLING-2002)*, pages 49–56.