

Multi-document Summarization Using Informative Words and Its Evaluation with a QA System

June-Jei Kuo, Hung-Chia Wung, Chuan-Jie Lin, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan, R.O.C.
{jjkuo, hjwong, cjlin}@nlg2.csie.ntu.edu.tw,
hh_chen@csie.ntu.edu.tw

Abstract. To reduce both the text size and the information loss during summarization, a multi-document summarization system using informative words is proposed. The procedure to extract informative words from multiple documents and generate summaries is described in this paper. At first, a small-scale experiment with 12 events and 60 questions was made. The results are evaluated by human assessors and a question answering (QA) system respectively. This QA system will help to prevent from drawbacks of human assessors. They show good performance of informative words. That encourages large-scale evaluation. An experiment is further conducted, which contains in total 140 questions out of 17,877 documents. Amongst these documents, 3,146 events were identified. The experimental results have also shown that the models using informative words outperform pure heuristic voting-only strategy when the metric of relative precision rate is used.

1 Introduction

The research of text summarization begins in the early 60s (Edmundson, 1964, 1969) and is one of the traditional topics in natural language processing. Recently, it attracts new attention due to the applications on the Internet. At this information explosion age, how to filter useless information, and to adsorb and apply information effectively become important issues to users. Many papers about document summarization have been proposed (Hovy and Marcu, 1998). Most of the previous works were done on single document summarization. Recently, the focus shifted to multiple documents summarization (Chen and Huang, 1999; Lin and Hovy, 2001; Mani and Bloedorn, 1997; Radev and McKeown, 1998; Radev, Blair-Goldensohn and Zhang, 2001) and even multilingual summarization (Chen and Lin, 2000). Of these, Chen and Huang (1999) employed named entities and other signatures to cluster documents; while as punctuation marks, linking elements, and topic chains to identify the meaningful units (MUs); employed nouns and verbs to find the similarity of MUs; and finally used a heuristic voting-only strategy¹ to generate summaries.

Although experimental results of Chen and Huang (1999) seemed promising, some issues had to be addressed as follows.

¹ The MUs that were reported by more than reporters were selected.

- (1) Goldstein, et al. (1999) mentioned that summary length depends on the document type, and fixed compression ratio is impractical. The summarization size of Chen and Huang's system is fixed and cannot be used to study the variance between the length and the precision rate on Chinese newswire documents.
- (2) The presentation order of sentences in a summary was based on the relative positions in the original documents instead of their importance. Thus, users might stop reading or miss the deferred appearing information.
- (3) The voting strategy gives a shorter summarization, which missed unique information reported only once.

This paper will follow the basic ideas of Chen and Huang (1999) on multi-document summarization and tackle the above problems. It is organized as follows: Section 2 presents a basic multi-document summarization system. Section 3 uses informative words to modify this system. The extraction of the related informative words and the sentence selection methodologies are described. Conventional evaluation model, i.e., human assessors, is adopted. Section 4 presents a QA system and introduces a new automatic evaluation model. Manual evaluation and automatic evaluation are compared. Section 5 shows a large-scale experiment. Two metrics, i.e., document reduction rate and QA precision rate, are considered. Finally, Section 6 is the conclusion.

2 A Basic Summarization System

Fig. 1 shows the architecture of a basic multi-document summarization system, which is used to summarize Chinese news from on-line newspapers. It is composed of two major components: a news clusterer and a news summarizer. The news clusterer receives a news stream from multiple on-line news sites, and directs them into several output news streams according to events. An event is denoted by five basic entities such as people, affairs, time, places and things. A news summarizer summarizes the news stories in each event cluster. All the tasks are listed below:

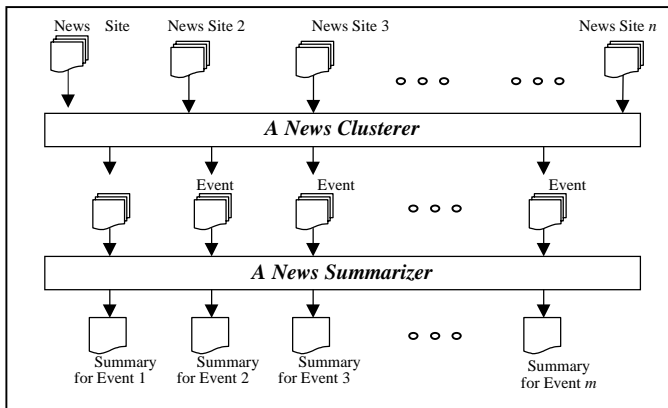


Fig. 1. System Architecture

- (1) Employing a segmentation system to identify Chinese words.
- (2) Extracting named entities like people, place, organization, time, date and monetary expressions.
- (3) Applying a tagger to determine the part of speech for each word.
- (4) Clustering the news stream based on the named entities and other signatures.
- (5) Partitioning a Chinese text into several meaningful units (MUs)².
- (6) Linking the meaningful units, denoting the same thing, from different news reports using the punctuation marks, linking elements, topic chains, etc.
- (7) Generating the summarization results using the longest sentence preference and voting strategy, which selects sentences reported more than once.

3 Generating Summaries with Informative Words

The concepts of topic words and event words were applied to topic tracking successfully (Fukumoto and Suzuki, 2000). The basic hypothesis is that an event word associated with a story appears across paragraphs, but a topic word does not. In contrast to event word, the topic word frequently appears across all documents. Thus, the document frequency of each word becomes an important factor in searching for the appropriate sentences ready for making summaries. As to the event words, that have higher term frequency in a document, will be more distinctive for the document. Therefore, we defined the words that have both high document frequency and high term frequency as informative words, and used them to improve the performance of step (7) of the basic system, which is specified in Section 2.

3.1 Informative Words and Sentence Selection for Summarization

The score function (IW) of an informative word W_{id} is defined as (3). $Ntf(W_{id})$ is normalized term frequency of term W_{id} . $tf(W_{id})$ and $mtf(d)$ are term frequency of W_{id} , and mean term frequency in document d , respectively. $D(W_{id})$ denotes document frequency of W_{id} , and N is total number of documents in an event. In formula (3), λ denotes a weighted number that can be learned from a corpus. λ was set to 1/2 and 1 in the later experiments.

$$Ntf(W_{id}) = \frac{tf(W_{id}) - mtf(d)}{tf(W_{id}) + mtf(d)} \quad (1)$$

$$DF(W_{id}) = D(W_{id}) / N \quad (2)$$

$$IW(W_{id}) = \lambda * (1) + (1 - \lambda) * (2) \quad (3)$$

In summarization, the more informative words a MU contains, the more possible the MU is used for generating summaries. In this paper, only the top 10 terms with

² Because Chinese writers often assign punctuation marks at random (Chen, 1994), the sentence boundary is not clear. Meaning units (MUs) are used for clustering instead of sentences. Here, a MU that is composed of several sentence segments denotes a complete meaning.

the higher IW scores will be chosen as informative words for a document. The score of each MU symbolizes the total number of informative words in it. The MUs with the highest score will be selected. Moreover, the selected MUs in a summary will be arranged in the descending order. In other words, the sentences which have more important MUs will appear before the less ones in a summary. In this case, even if the readers unfortunately stop reading the summaries half way, they would not miss out much important information.

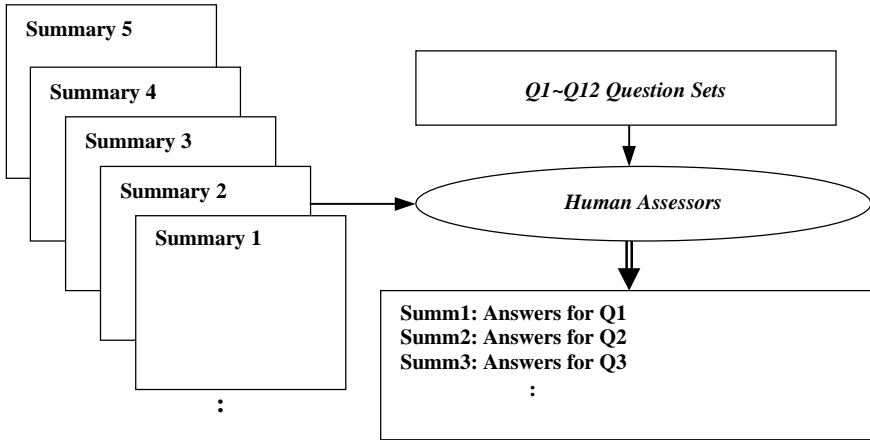


Fig. 2. Example of QA Task

3.2 Experiment Result

Fig. 2 shows our block diagram of the intrinsic evaluation task (Tsutomo, Sasaki and Isozaki, 2001) on text summarization by referring the SUMMAC Q&A evaluation (SUMMAC, 1998). For simplicity, we call it *QA task*. First, the question sets (query sets) are collected under the document collection. While as, the corresponding answer sets are made after reading all the documents. After various kinds of document summaries are completed, the assessors will be involved in the evaluation. Each assessor will be assigned for summary texts and their related question sets. During the evaluation, the reading and the answering time will be recorded. When assessors finish the question and answering task, we review their answers responding to its respective answer sets and compute the precision rate of each question. Besides, the average document reduction rate and the average Q&A precision of various types of summary text are computed, respectively.

In our experiment, the test data is collected from 6 news sites in Taiwan, they are: China Times, Commercial Times, China Times Express, United Daily News, Tomorrow Times, and China Daily News, through the Internet. There are in total 17,877 documents (near 13MB) from January 1, 2001 to January 5, 2001. The total number of MUs is 189,774. After clustering, there are 3,146 events. Because of assessor cost, only 12 events were selected randomly in the first stage. 60 questionnaires (5 questions of each event) are made manually with answers to their related documents. Moreover, 12 members of our laboratory who are all graduate

students majoring in computer science are selected to conduct these following experiments: (1) full text (FULL), (2) Chen and Huang's system (1999) as the base line system (BASIC) (3) term frequency only with vote strategy (TFWV, i.e., $\lambda=1$), (4) informative words with vote strategy (PSWV, i.e., $\lambda=1/2$) (5) term frequency without vote strategy (TFNV, i.e., $\lambda=1$), and (6) informative words only without vote strategy (PSNV, i.e., $\lambda=1/2$). The above "proposed system" denotes our text summarization system using informative words. Each assessor evaluates a summarization method twice, using different question sets (i.e., answer only once per event) shown as Table 1. The characters A, B, C, ..., L in the first column denote the assessors A, B, C, ..., L. The names in the first row are the types of summary text. Symbol Q_n in the cell denotes the question set for event n . To evaluate objectively, each assessor does not know the text types what he (she) assesses. The experimental results are shown in Table 2. R&A time means the summation of reading time and answering time. On the one hand, Reduction Rate-S and Reduction Rate-T mean the relative reduction rate of size and R&A time, respectively. The definition of Relative Reduction Rate of size is (Size of a specified system)/(Size of FULL). The average precision and its relative variance of each text type are also given to show the statistical information.

Table 1. Assessor Assignments

	FULL	BASIC	TFWV	PSWV	TFNV	PSNV
A	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12
B	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7
C	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8
D	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9
E	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10
F	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11
G	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12
H	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7
I	Q3, Q9	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8
J	Q4, Q10	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9
K	Q5, Q11	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10
L	Q6, Q12	Q1, Q7	Q2, Q8	Q3, Q9	Q4, Q10	Q5, Q11

3.3 Discussion

Several observations from Table 2 are shown below.

- (1) The size of TFNV and PSNV is larger than that of BASIC (near 15%), but the precision rate of TFNV and PSNV is lower than that of BASIC.
- (2) The size of TFWV and PSWV is smaller than that of BASIC, and their precision rate is still smaller than that of BASIC.
- (3) The precision rates of both TFWV and PSWV are larger than those of TFNV and PSNV.

The above observations are out of our expectation. From observations (1) and (2), the informative words seem not to be useful in MU selection. From observation (3), the vote strategy seems to be useful in improving the precision. In other words,

neglecting the news story reported by only one reporter seems to have no problems in Q&A. However, due to limitations and drawbacks of human assessment, evaluation shown below in the QA task may mislead.

- (1) Due to different background among human assessors, the evaluation is unable to be objective. We have to conduct several evaluations in order to obtain correct and objective results. Nevertheless, this will be cost-effective.
- (2) Fatigue and limited of time scale to work may effect the assessor to of the assessors to quit reading or read too fast so as to miss the information that will be useful to answer the questions. This will cause the low precision of summarizing the text.
- (3) Due to the high cost of the assessors, the large-scale evaluation is nearly impossible.

Table 2. Results Using Question-Answering Task

	<i>FULL</i>	<i>BASIC</i>	<i>TFWV</i>	<i>PSWV</i>	<i>TFNV</i>	<i>PSNV</i>
Size (Byte)	59637	12974	12002	12348	15192	15267
Reduction Rate-S	1	0.22	0.20	0.21	0.25	0.26
Reading Time (sec)	2224	780	744	660	816	804
Answering Time (sec)	1752	1236	1200	1128	1356	1260
R&A Time (sec)	3976	2016	1944	1788	2172	2064
Reduction Rate-T	1	0.51	0.49	0.45	0.55	0.52
Precision	0.923	0.525	0.513	0.519	0.502	0.513
Variance	0.010	0.047	0.095	0.054	0.712	0.061

4 An Evaluation Model Using Q&A Systems

4.1 Model Using Q&A System

In order to improve the QA task and verify the experimental results, a QA system is used to substitute the human assessors in Fig. 2 and the flow of the revised evaluation model is shown in Fig. 3. Both full texts and summaries are read by QA systems, and QA systems find the answers from full texts and summaries. Although the efficiency of a QA system may affect the evaluation results, that is fair for all summarization models under the same evaluation environment.

The QA system we adopted was borrowed from Lin and Chen (1999), whose main strategies are keyword matching and question-focus identifying. This system has been used in open domain question and answering on heterogeneous data (Lin, *et al.*, 2001). It is composed of three major modules shown as follows:

- (1) Preprocessing the Question Sentences

At first, the parts-of-speech are assigned to the words in question sentences. Then, the stop-words are removed. The remaining words are transformed into the canonical forms and considered as the keywords of question sentences. For each keyword, they find all synonyms from the related thesaurus, e.g. WordNet (Fellbaum, 1998). Those terms are the expansion set of the keywords. Moreover, no matter whether the keyword is a noun, a verb, an adjective or an adverb, all the possible morphological forms of the word are also added into this set.

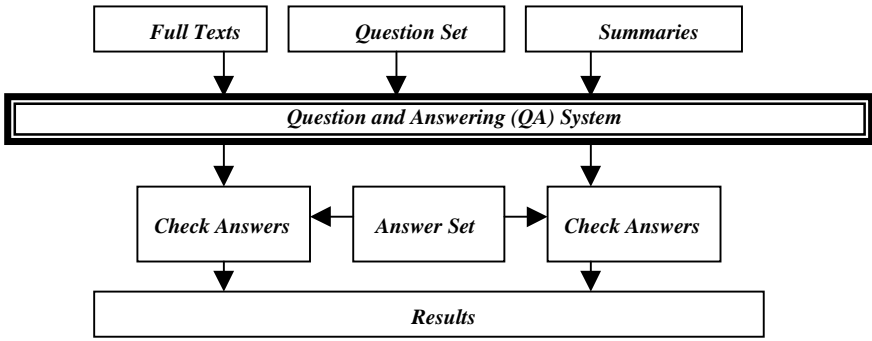


Fig. 3. Revised Evaluation Model

(2) Retrieving the Documents Containing Answers

A full text retrieval system is implemented to decrease the number of documents to be searched for the answering sentences. Each keyword of a expanded question sentence is assigned a weight. Especially, those words tagged with proper-noun markers have been assigned higher weights. This is because they may be presented in the answer. The score of a document D is computed as follows:

$$score(D) = \sum_{t \text{ in } D} weight(t) \quad (4)$$

where t is one of the keywords in expanded question sentence.

Those documents that score more than a threshold are selected as the answering documents. Threshold is set to the sum of weights of the words in the original question sentences. If documents do not have scores bigger than the threshold, we assume that there is no answer to the question.

(3) Retrieving the Sentences Containing Answers

Finally, each sentence in the retrieved documents is examined. Those sentences that contain most words in the expanded question sentence are retrieved. The top five sentences are regarded as the answers. The answers are sorted according to the number of matched words and the retrieving scores computed at step (2).

4.2 Evaluation

The experimental results using the same data in Section 3.2 are shown in Table 3. The precision from Table 1 is reproduced here for comparison. After the QA system reads all documents of 12 events, it will propose five plausible answers for each question. The metric is MRR (Mean Reciprocal Rank) (Voorhees, 2000):

$$MRR = \sum_{i=1}^N r_i / N \quad (5)$$

where $r_i = 1/\text{rank}_i$ if $\text{rank}_i > 0$, or 0 if $\text{rank}_i = 0$. rank_i is the rank of the first correct answer of the i^{th} question, and N is total number of questions. That is, if the first correct answer is at rank 1, the score is $1/1=1$; if it is at rank 2, the score is $1/2=0.5$, and so on. If no answer is found, score is 0. In this way, the evaluation time can be

reduced significantly. That makes large-scale evaluation feasible. Meanwhile, to compare with the precision of QA task in Table 2, we also use five strategies (e.g. Best-1, Best-2, and so on) to compute the precision of the QA system. With Best-1 strategy, the answer must exist in ranked one answer of QA system. With Best-2 strategy, the answer exists in either ranked 1 or 2, or both. Furthermore, to show the feasibility of the proposed evaluation method, we also perform a large-scale experiment that will be discussed in the next section, which human assessment is in question.

Table 3. Results with Small-Scale Data using a QA system

	FULL	BASIC	TFWV	PSWV	TFNV	PSNV
Precision of QA Task	0.923	0.525	0.513	0.519	0.502	0.513
Precision of Best-1	0.881	0.441	0.407	0.457	0.475	0.475
Precision of Best-2	0.915	0.475	0.475	0.508	0.576	0.559
Precision of Best-3	0.949	0.491	0.475	0.508	0.576	0.559
Precision of Best-4	0.966	0.508	0.491	0.525	0.576	0.559
Precision of Best-5	0.966	0.541	0.517	0.525	0.576	0.559
QA_MRR	0.914	0.493	0.476	0.487	0.508	0.517
Relative MRR	1	0.576	0.521	0.533	0.556	0.566

4.3 Discussion

Because the QA system avoids the above limitation and drawback of human assessments, the precisions of some types of summarization text are different from the results shown in Table 2. Observing Table 2 and Table 3, there are some differences shown below:

- (1) QA_MRR values of TFNV and PSNV are larger than those of the corresponding TFWV and PSWV. Thus, we can conclude that the vote strategy will lose some useful information.
- (2) QA_MRR values of PSWV and PSNV are larger than those of the corresponding TFWV and TFNV. We can draw to the conclusion that using both term frequency and document frequency of informative words will select more important MUs than only using term frequency of informative words.
- (3) Comparing the precisions of QA task with the corresponding precisions of best-5 strategy, QA system is better than QA task. Thus, we can say that the QA system can find the answers more effective than human assessors.

In order to show the feasibility of large-scale evaluation using Q&A system, we continue to perform an even greater scale of experiment in the next section, which is impossible to be performed using QA task.

5 Experiments Using Large Documents and Results

5.1 Data Set

From the above analysis, we can conclude that a high performance QA system can be used to play the role of human assessors. Besides the evaluation time and scale, it can

obtain more objective and precise results. In the next experiment, the complete data set as described in Section 3.2 was used. Under the data set, 140 new questionnaires are made and 93 questions have been answered. Thus, using these practical questions we can further observe the performance of QA system in text summarization evaluation. Some samples of questions are shown below.

- Q68. 英特爾最新發表產品為何?
What is the newest product of Intel Company?
- Q95. 歐拉朱萬何時受傷?
When was Mr. Olajuwon wounded?

5.2 Experimental Results and Discussion

Table 5 shows the experimental results using large documents. According the data obtained from the QA system using a large scale of documents, the results are summarized as follows:

- (1) Due to the increase of document size, the QA_MRR of all models decreased.
- (2) Due to increasing noise of FULL, the QA_MRR of FULL drops drastically. The relative MRRs of the other models increased when comparing with Table 3.
- (3) The QA_MRR values of TFWV, PSWV, TFNV and PSNV are also larger than the value of BASIC. This is consistent with the above results in small-scale evaluation using QA system. Thus, informative words in MU's selection present good performance.
- (4) The QA_MRR values of PSWV and PSNV are also larger than those of TFWV and TFNV, respectively. To achieve better result, it is recommended to use combination of term frequency and document frequency in MU's selection.
- (5) Since the performance of each model has the similar results to those shown in Table 4, it is feasible to use the QA system in evaluating the performance of large-scale multiple document summarization.

Table 4. Results with Large-Scale Data

	<i>FULL</i>	<i>BASIC</i>	<i>TFWV</i>	<i>PSWV</i>	<i>TFNV</i>	<i>PSNV</i>
Size (Kbyte)	13,137	1,786	1,771	1,773	2,226	2,218
QA_MRR	0.515	0.314	0.342	0.346	0.359	0.380
Relative MRR	1	0.610	0.664	0.672	0.697	0.738

6 Conclusion

This paper presents a multi-document summarization system using informative words and an automatic evaluation method for summaries using a QA system. Using the normalized term frequency and document frequency, the informative words can be extracted effectively. The informative words are shown to be more useful to select sentences for generating summaries than the heuristic rule. Moreover, the sentences in the summaries can be put in order according to the total number of informative words. In this way, the important sentences are generated in the early part. The

summaries can be compressed easily by deleting sentences from the end without losing much important information, and the length of summary can be adjusted robustly. On the other hand, the evaluation processes show that QA system can play an important role in conducting large-scale evaluation of multi-document summarization and make the results more objective than the human assessors. There are still some issues that need further research:

- (1) Investigating to what extent the errors of QA system may affect the reliability of the evaluation results
- (2) Using other QA systems to justify the feasibility of the above evaluation model.
- (3) Introducing the machine learning method to obtain λ value and its possible size of summary for various kinds of documents.
- (4) Using some statistical model and null hypothesis test to study the results' relationship between QA task and QA systems.
- (5) Introducing the statistical methods, such as the dispersion values of words among document (Fukumoto and Suzuki, 2000) to find the informative words more effectively for the purpose of improving the performance of the summarization system.

Acknowledgements

The research in this paper was partially supported by National Science Council under the contracts NSC89-2213-E-002-064 and NSC90-2213-E-002-013.

References

1. Chen, H.H.: The Contextual Analysis of Chinese Sentences with Punctuation Marks. *Literal and Linguistic Computing*, Oxford University Press, 9(4) (1994) 281-289
2. Chen, H.H. and Huang, S.J.: A Summarization System for Chinese News from Multiple Sources. *Proceeding of 4th International Workshop on Information Retrieval with Asia Language* (1999) 1-7.
3. Chen, H.H. and Lin, C.J.: A Multilingual News Summarizer. *Proceeding of 18th International Conference on Computational Linguistics*, (2000) 159-165.
4. Edmundson, H.P.: Problems in Automatic Extracting. *Communications of the ACM*, 7, (1964) 259-263.
5. Edmundson, H.P.: New Methods in Automatic Extracting. *Journal of the ACM*, 16, (1969) 264-285.
6. Firmin Hand, T. and B. Sundheim (eds): TIPSTER-SUMMAC Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*, Washington. (1998)
7. Fukumoto, F. and Suzuki, Y.: Event Tracking based on Domain Dependency. *Proceedings of SIGIR 2000* (2000) 57-64.
8. Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J.: Summarizing Text Documents: Sentences Selection and Evaluation Metrics. *Proceedings of SIGIR 1999* (1999) 121-128.
9. Hovy, E. and Marcu, D.: Automated Text Summarization. *Tutorial in COLING/ACL98* (1998)
10. Lin, C.J. and Chen, H.H.: Description of Preliminary Results to TREC-8 QA Task. *Proceedings of The Eighth Text Retrieval Conference* (1999) 363-368.

11. Lin, C.J., Chen, H.H., Liu, C.J., Tsai, C.H. and Wung, H.C.: Open Domain Question Answering on Heterogeneous Data. Proceedings of ACL Workshop on Human Language Technology and Knowledge Management, July 6-7 2001, Toulouse France, (2001) 79-85.
12. Lin, C.Y. and Hovy E.: NEATS: A Multidocument Summarizer. Workshop of DUC 2001 (2001) [on-line] Available:
http://www-nlpir.nist.gov/projects/duc/duc2001/agenda_duc2001.html
13. Mani, I. and Bloedorn, E.: Multi-document Summarization by Graph Search and Matching. Proceedings of the 10th National Conference on Artificial Intelligence, Providence, RI, (1997) 623-628.
14. Mani, I. et al.: The TIPSPER SUMMAC Text Summarization Evaluation: Final Report, Technique Report. Automatic Text Summarization Conference, (1998)
15. Radev, D.R. and McKeown, K.R.: Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics, Vol. 24 No. 3 (1998) 469-500.
16. Radev, D.R., Blair-Goldensohn and Zhang, Z.: Experiment in Single and Multi-Document Summarization Using MEAD. Workshop of DUC 2001 (2001) [on-line] Available:
http://www-nlpir.nist.gov/projects/duc/duc2001/agenda_duc2001.html
17. Regina Barzilay and Michael Elhadad: Using Lexical Chains for Text Summarization. Proceedings of The Intelligent Scalable Text Summarization Workshop, ACL/EACL (1997) 10-17.
18. Tsutomu, H., Sasaki, T. and Isozaki H.: An Extrinsic Evaluation for Question-Biased Text Summarization on QA Tasks. Proceedings of workshop on Automatic Summarization (2001) 61-68.
19. Voorhees: QA Track Overview (TREC) 9, (2000) [on-line] Available:
<http://trec.nist.gov/presentations/TREC9/qa/index.htm>
20. Fellbaum, C.: WordNet. The MIT Press, Cambridge Masschusettes (1998)