AN NTU-APPROACH TO AUTOMATIC SENTENCE EXTRACTION FOR SUMMARY GENERATION

Kuang-hua Chen
Language & Information Processing System Lab. (LIPS)
Department of Library and Information Science
National Taiwan University
1, SEC. 4, Roosevelt RD., Taipei
TAIWAN, 10617, R.O.C.
E-mail: khchen@ccms.ntu.edu.tw
Fax: +886-2-23632859

Sheng-Jie Huang, Wen-Cheng Lin and Hsin-Hsi Chen
Natural Language Processing Laboratory (NLPL)
Department of Computer Science and Information Engineering
National Taiwan University
1, SEC. 4, Roosevelt RD., Taipei
TAIWAN, 10617, R.O.C.
E-mail: {sjhuang,denislin}@nlg.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw
Fax: +886-2-23638167

ABSTRACT

Automatic summarization and information extraction are two important Internet services. MUC and SUMMAC play their appropriate roles in the next generation Internet. This paper focuses on the automatic summarization and proposes two different models to extract sentences for summary generation under two tasks initiated by SUMMAC-1. For categorization task, positive feature vectors and negative feature vectors are used cooperatively to construct generic, indicative summaries. For adhoc task, a text model based on relationship between nouns and verbs is used to filter out irrelevant discourse segment, to rank relevant sentences, and to generate the user-directed summaries. The result shows that the NormF of the best summary and that of the fixed summary for adhoc tasks are 0.456 and 0.447. The NormF of the best summary and that of the fixed summary for categorization task are 0.4090 and 0.4023. Our system outperforms the average system in categorization task but does a common job in adhoc task.

1. INTRODUCTION

Towards the end of the 20th century, the Internet has become a part of life style. People enjoy Internet services from various providers and these ISPs (Internet Services Providers) do their best to fulfill users' information need. However, if we investigate the techniques used in these services, we will find out that they are not different from those used in traditional Information Retrieval or Natural Language Processing. However, the cyberspace provides us an environment to utilize these techniques to serve more persons than ever before.

The members under the leadership of Professor Hsin-Hsi Chen of Natural Language Processing Lab. (NLPL) in Department of Computer Science and Information Engineering, National Taiwan University have dedicated themselves in researches of NLP for many years. The research results have been reported in literature and received the reputation from colleagues of NLP field. Many systems for various NLP applications have been developed, especially for Chinese and English. Some systems could be accessed directly via WWW browsers. For example, an MT meta-server [1] provides an online English-to-Chinese translation service. (http://nlg3.csie. ntu.edu.tw/mtir/mtir.html)

Language & Information Processing System Lab. (LIPS) in Department of Library and Information Science, National Taiwan University also devotes itself in researches of language, information and library sciences. Chen and Chen [2] proposed hybrid model for noun extraction from running texts and provided an automatic evaluation method. Chen [3] proposed a corpus-based model to identify topics and used it to determine sub-topical structures.

Generally speaking, we are capable of dealing with numerous NLP applications or apply NLP techniques to other applications using our current research results. The two laboratories think that current Internet services are not enough for the people living in the next century. At least, two kinds of services are important and crucial in the 21st century: one is the information extraction; the other is automatic summarization.

Information Extraction (IE) [4] systems manage to extract predefined information from data or documents. What kind of information is appropriate is a domain-dependent problem. For example, the information conveyed by business news and by terrorism news is very different. As a result, the predefined information plays an important role in IE systems. In fact, the predefined information is the so-called metadata [5]. The joint efforts on IE and metadata will benefit both sides.

Automatic summarization is to use automatic mechanism to produce a finer version for the original document. Two possible methodologies could be applied to constructing summaries. The first is to extract sentences directly from texts; the second is to analyze the text, extract the conceptual representation of the text, and then generate summary based on the conceptual representation. No matter what methodology is adopted, the processing time should be as little as possible for Internet applications.

As we mentioned above, information extraction and automatic summarization are regarded as two important Internet services in the next century. Therefore, we take part in MET-2 and SUMMAC-1 for the respective purposes. In this paper, we will focus on the tasks of SUMMAC-1 and the details of MET-2 can be referred to the paper presented in MET-2 Conference [6].

This paper is organized as follows. Section 2 discusses the types of summaries and their functions. In addition, the tasks of SUMMAC-1 and the corresponding functions to the traditional summaries are also described. Sections 3 and 4 propose the models to carry out the categorization task and adhoc task, respectively. The method for extracting feature vectors, calculating extraction strengths, and identifying discourse segments are illustrated in detail in the two sections. Section 5 shows our results in summary and compares with other systems. Section 6 gives a short conclusion.

2. SUMMARY AND SUMMAC-1 TASKS

In general, summarization is to create a short version for the original document. The functions of summaries are shown as follows [7]:

- Announcement: announce the existence of the original document
- Screening: determine the relativeness of the original document
- Substitution: replace the original document
- Retrospection: point to the original document

A summary can be one of four types, i.e., indicative summary, informative summary, critical summary, and extract. Indicative summaries are usually of functions of announcement and screening. By contrast, informative summaries are of function of substitution. It is very difficult to generate critical summaries in automatic ways. Extract can be of announcement, and replacement. In general, all of the four types of summaries are retrospective.

The most important summary types are indicative summary and informative summary in the Internet environment. However, for researchers devoting themselves in automatic summarization, the common type of summary is extract. This is because the extract is produced through extracting the sentences in the original document and this is an easier way to produce a summary. But, how to make extract possess the functionality of informative summary and that of indicative summary? A common way is to produce a fix-length extract for indicative summary and to produce a best extract for informative summary. That is the also two different summaries underlying the tasks of SUMMAC-1.

SUMMAC-1 announces three tasks for automatic summarization: the first is categorization task; the second is adhoc task; the third is Q&A task. These three tasks have their own designated purposes. As the SUMMAC-1 design, the tasks address the following types of summaries:

- Categorization: Generic, indicative summary
- Adhoc: Query-based, indicative summary
- Q&A: Query-based, informative summary

Although the definitions shown above are not the same as we talk about in previous paragraph, this will not interfere the development of an automatic summarization system.

Because we have many experiences in applying language techniques to dealing with the similar tasks [3, 8], we decide to take part in Categorization task and Adhoc task after long discussion. The reasons are described as follows. For an application in the Internet environment, to provide introductory information for naïve users is very important. It is very suitable to use generic indicative summaries to fulfill this function. However, the users have their own innate knowledge and they want that the generated summary is relative to the issued query at times. Therefore, the two different needs are fulfilled as the first and the second tasks initiated by SUMMAC-1. As to the third task, Q&A, we think that it is much more relative to the information

extraction. It can be resolved in association with IE as a part of MUC's tasks.

3.CATEGORIZATION TASK

As the call for paper of SUMMAC-1 says, the goal of the categorization task is to evaluate generic summaries to determine if the key concept in a given document is captured in the summary. The SUMMAC-1 documents fall into sets of topics and each topic contains approximately 100 documents. The task asks summarization systems to produce summary for each document. The assessor will read the summary and then assign the summary into one of five topics or the sixth topic, 'non-relevant' topic.

The testing set of documents consists of two general domains, environment and global economy. Each domain in turn consists of five topics and each topic contains 100 documents. As a result, these documents could be regarded as the positive cues for the corresponding topic. By contrast, documents of other topics could be treated as the negative cues for the topic under consideration. The training stage and the testing stage are described in the following paragraph.

For each topic, the following procedure is executed in the training stage.

(1) Screen out function words for each document
(2) Calculate word frequency for current topic as positive feature vector (PFV)
(3) Calculate word frequency for other topics as negative feature vector (NFV)

The testing stage is shown as follows.

(1) Exclude function words in test documents
(2) Identify the appropriate topic for testing documents
(3) Use PFV and NFV of the identified topic to rank sentences in test documents
(4) Select sentences to construct a best summary
(5) Select sentences to construct a fixed-length summary

Based on this line, the approach for summary generation under the categorization task could be depicted as Figure 1 shows.

Step (1) in training stage and testing stage are to exclude function words. A stop list is used as this purpose. A stop list widely distributed in the Internet and another list collected by us are combined. The resultant stop list consists of 744 words, such as abaft, aboard, about, above, across, afore, after, again, against, ain't, aint, albeit, all, almost, alone, along, alongside, already, also, although, always, am, amid, and so on.

Steps (2) and (3) in training stage regard the document collection of a topic as a whole to extract the *PFV* and *NFV*. Firstly, the document collection of a topic is thought as the pool of words. Step (2) calculates the frequency of each word in this pool and screens out those words with frequency lower than 3. Step (3) repeats the same procedure. However, this time the pool consists of words from document collections of other topics. After normalization, two feature vectors $PFV = (pw_1, pw_2, pw_3, \ldots, pw_n)$ and $NFV = (nw_1, nw_2, nw_3, \ldots, nw_n)$ are constructed to be unit vectors. The *PFV* and *NFV* are used to extract sentences of document and those extracted sentences consist of the summary. The idea behind this approach is that we use documents to retrieve the strongly related sentences in parallel to IR system use query sentence to retrieve the related documents.

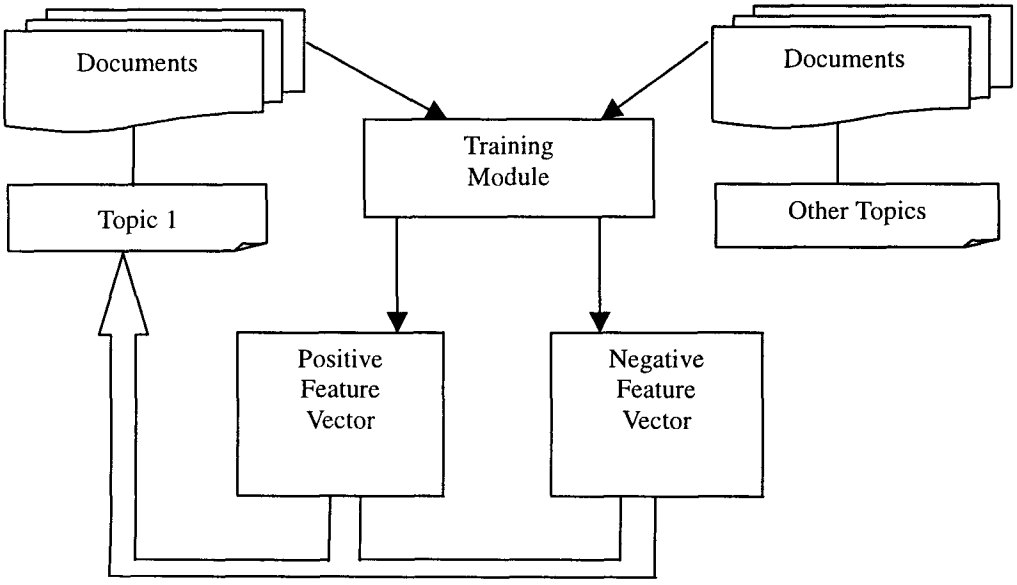

Figure 1. The Training Procedure for Categorization Task

Step (2) in testing stage is to identify which topic the testing document belongs to. The *PFV*s and the *NFV*s are used to compare with testing documents. Assume that the testing document D consists of dw_1, dw_2, dw_3,..., and dw_n words, i.e., $D = (dw_1, dw_2, dw_3, \ldots, dw_n)$ and there are m pairs of *PFV* and *NFV*. The following equation is used to determine that the $\hat{i}$'th topic is best for the document under consideration.

$$\hat{i} = \arg\max_{1 \le i \le m}(sim(PFV_i, D) - sim(NFV_i, D))$$

The similarity shown in the following is measured by inner product.

$$sim(PFV, D) = \sum_{j=1}^{n}(pw_j \times dw_j)$$

While the topic is determined, Step (3) uses the corresponding $PFV_{\hat{i}}$ and $NFV_{\hat{i}}$ to select sentences in the document. Whether a sentence $S = (sw_1, sw_2, sw_3, \ldots, sw_n)$ is selected as part of a summary depends on the relative score shown as follows. The similarity is also measured by inner product.

$$RS(S) = sim(PFV_{\hat{i}}, S) - sim(NFV_{\hat{i}}, S)$$

In Step (4), the ranked list of RSes is examined and the maximal score gap between two immediate RSes is identified. If the number of sentences above the identified gap is between 10% to 50% of that of all sentences, these sentences are extracted as the best summary. Otherwise, the next maximal gap is examined whether it is a suitable gap or not. Step (5) just uses the best summary generated in Step (4) and makes a fixed-length summary according to the SUMMAC-1 rule.

4. ADHOC TASK

Adhoc Task is designed to evaluate user-directed summaries, that is to say, the generated summary should be closely related to the user's query. This kind of summary is much more important for Internet applications. We have devoted ourselves in related researches for a long time. A text model based on the interaction of nouns and verbs was proposed in [3], which is used to identify topics of documents. Chen and Chen [8] extended the text model to partition texts into discourse segments.

The following shows the process of NTU's approach to adhoc task in SUMMAC-1 formal run.

(1) Assign a part of speech to each word in texts.
(2) Calculate the extraction strength (ES) for each sentence.
(3) Partition the text into meaningful segments.
(4) Filter out irrelevant segments according to the user's query.
(5) Filter out irrelevant sentences based on ES.
(6) Generate the best summary.
(7) Generate the fixed-length summary from the best summary.

Step (1) is used to identify the nouns and the verbs in texts, which are regarded as the core words in texts and will be used in Step (2). Step (2) is the major stage in our approach and will be discussed in detail.

Generally speaking, each word in a sentence has its role. Some words convey ideas, suggestions, and concepts; some words are functional rather than meaningful. Therefore, it is much more reasonable to strip out these function words, while we manage to model information flow in texts. Nouns and verbs are two parts of speech under consideration. In addition, a measure for word importance should be worked out to treat each noun or verb in an appropriate scale. In tradition, term frequency (TF) is widely used in researches of information retrieval. The idea is that after excluding the functional words, the words occur frequently would carry the meaning underlying a text. However, if these words appear in many documents, the discriminative power of words will decrease. Spack Jones [9] proposed inverse document frequency (IDF) to rectify the aforementioned shortcoming. The IDF is shown as follows:

$$IDF(w) = \log(P\text{-}O(w))/O(w),$$

where P is the number of documents in a collection, $O(w)$ is the number of documents with word w.

Nouns and verbs in well-organized texts are coherent in general. In order to automatically summarize texts, it is necessary to analyze the factors of composing texts. That is, the writing process of human beings. We use four distributional parameters to construct a text model:

- Word importance
- Word frequency
- Word co-occurrence
- Word distance

The following will discuss each factor in sequence.

The word importance means that when a word appears in texts, how strong it is to be the core word of texts. In other words, it represents the possibility of selecting this word as an index term. The IDF is chosen to measure the word importance in this paper. In addition, the frequency of a word itself does also play an important role in texts. For example, the word with high frequency usually makes readers impressive. The proposed model combines the two factors as the predecessors did.

If a text discusses a special subject, there should be many relative words together to support this subject. That is to say, these relative words will co-occur frequently. From the viewpoint of statistics, some kind of distributional parameters like mutual information [10] could be used to capture this phenomenon.

Including the distance factor is motivated by the fact that related events are usually located in the same texthood. The distance is measured by the difference between cardinal numbers of two words. We assign a cardinal number to each verb and noun in sentences. The cardinal numbers are kept continuous across sentences in the same paragraph. As a result, the distance between two words, w_1 and w_2, is calculated using the following equation.

$$D(w_1,w_2) = abs(C(w_1)-C(w_2)),$$

where the D denotes the distance and C the cardinal number.

Consider the four factors together, the proposed model for adhoc task is shown as follows:

$$CS(n) = pn \times SNN(n) + pv \times SNV(n)$$

CS is the connective strength for a noun *n*, where *SNN* denotes the strength of a noun with other nouns, *SNV* the strength of a noun with other verbs, and *pn* and *pv* are the weights for *SNN* and *SNV*, respectively. The determination of *pn* and *pv* is via deleted interpolation [11] (Jelinek, 1985). The equations for *SNV* and *SNN* are shown as follows.

$$SNV(n_i) = \sum_j \frac{IDF(n_i) \times IDF(v_j) \times f(n_i, v_j)}{f(n_i) \times f(v_j) \times D(n_i, v_j)}$$

$$SNN(n_i) = \sum_j \frac{IDF(n_i) \times IDF(n_j) \times f(n_i, n_j)}{f(n_i) \times f(n_j) \times D(n_i, n_j)}$$

$f(w_i,w_j)$ is the co-occurrence of words w_i and w_j, and $f(w)$ is the frequency of word w. In fact, $f(w_i,w_j)/f(w_i)\times f(w_j)$ is a normalized co-occurrence measure with the same form as the mutual information.

When the connectivity score for each noun in a sentence is available, the chance for a sentence to be extracted as a part of summary can be expressed as follows. We call it extraction strength (ES).

$$ES(S_i) = \sum_{j=1}^{m} CS(n_{ij}) / m,$$

where m is the number of nouns in sentence Si.

Because texts are well organized and coherent, it is necessary to take the paragraph into consideration for summary generation. However, the number of sentences in paragraphs may be one or two, especially in newswire. It is indispensable to group sentences into meaningful segments or discourse segments before carrying out the summarization task. Step (3) is for this purpose. A sliding window with size W is moved from the first sentence to the last sentence and the score for sentences within the window is calculated. Accordingly, a series of scores is generated. The score-sentence relation determines the boundaries of discourse segments. Figure 2 shows aforementioned process and how to calculate the scores. The window size W is 3 in this experiment.

While discourse segments are determined, the user's query is used to filter out less relevant segments. This is fulfilled in Step (4). The nouns of a query are compared to the nouns in each segment and the same technique for calculating *SNN* mentioned above is used [8]. As a result, the precedence of segments to the query is calculated and then the medium score is identified. The medium is used to normalize the calculated score for each segment. The segments with normalized score lower than 0.5 are filtered out.

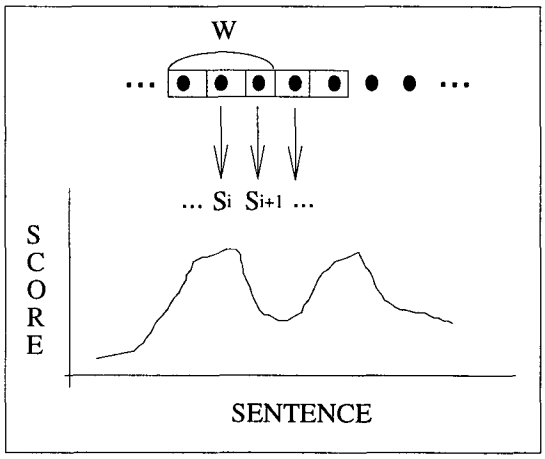

Figure 2. Determination of discourse segments

Step (5) is to filter out the irrelevant sentences in the selected segments in Step (4). The ES of each sentence calculated in Step (2) is used as the ranking basis, but the ES of first sentence and that of the last sentence are doubled. Again, the medium of these ESes is chosen to normalize these score. The sentences with normalized score higher than 0.5 are selected as the best summary in Step (6). Because the length of fixed-length summary cannot exceed the 10% of the original text, Step (7) selects the top sentences that do not break this rule to form the fixed-length summary.

5. EXPERIMENT RESULTS

In general, the results are evaluated by assessors, and then measured by recall (R), precision (P), F-score (F) and the normalized F-score (NormF). Table 1 shows the contingence table of the real answer against the assessors.

	Given Answer by Assessors	
Real Answer	TP	FN
	FP	TN

Table 1. Contingence Table

The meanings of TP, FP, FN, and TN are shown in the following:

- TP : Decides relevant, relevant is correct = true positive
- FP : Decides relevant, relevant is incorrect = false positive

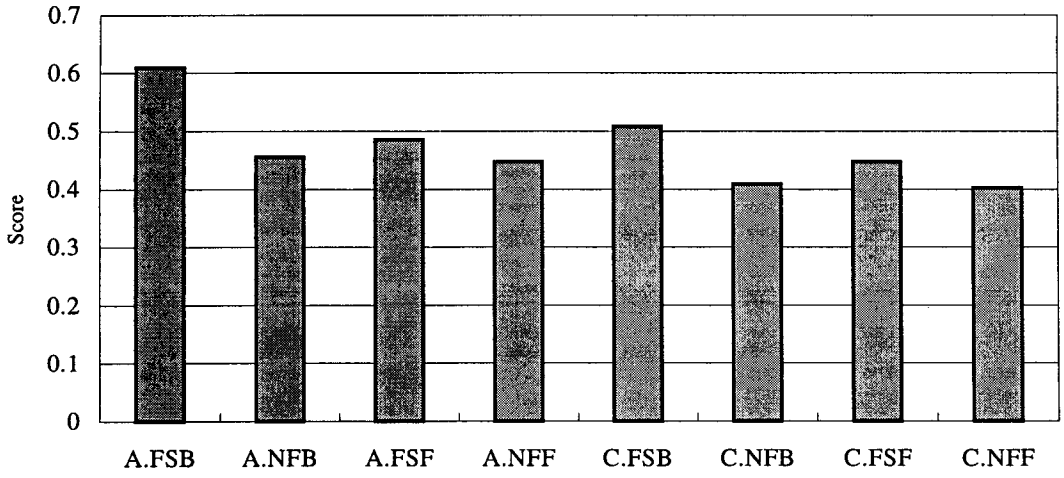

Figure 3. The performance of our system

- FN : Decides irrelevant, relevant is correct = false negative
- TN : Decides irrelevant, irrelevant is correct = true negative

The aforementioned measures for evaluation based on Table 1 are shown in the following:

- Precision (P) = (TP/(TP+FP))
- Recall (R) = (TP/TP+FN)
- F-score (F) = (2*P*R/(P+R))

Each group could provide up to two kinds of summary. One is the fixed-length summary and the other is the best summary. In order to level off the effect of length of summary, compression factor is introduced to normalize the F-score.

- Compression (C) = (Summary Length/Full Text Length)
- NormF = ((1-C)*F)

Table 2 shows the result of our adhoc summary task. Table 3 shows the result of our categorization summary task. The NormF of the best summary and that of the fixed summary for adhoc tasks are 0.456 and 0.447, respectively. In comparison to other systems, the performance of our system is not good. One reason is that we have not developed an appropriate method to determine the threshold for selection of sentence. Besides, we are the only one team not from Indo-European language family. This maybe has some impacts on the performance. However, considering the time factor, our system perform much better than many systems.

The NormF of the best summary and that of the fixed summary for categorization task are 0.4090 and 0.4023, respectively. Basically, this task is like the traditional categorization problem. Our system performs much well. However, there is no significant difference among all participating systems.

Table 4 shows our system's performance against average performance of all systems. Although some measures of our performance are worse than that those of the average performance, the difference is not very significant. In categorization task, we outperform the average performance of all systems. Table 5 is the standard deviation of all systems. Essentially, the difference of all systems is not significant. Figure 3 shows each measure of performance for our system. Figure 4 shows our system against the best system.

A.FSB	F-Score Best summary	0.6090
A.NFB	NormF Best summary	0.4560
A.FSF	F-Score Fixed summary	0.4850
A.NFF	NormF Fixed summary	0.4470

Table 2. Result of Adhoc

C.FSB	F-Score Best summary	0.5085
C.NFB	NormF Best summary	0.4090
C.FSF	F-Score Fixed summary	0.4470
C.NFF	NormF Fixed summary	0.4023

Table 3. Result of Categorization

A.FSB	-0.040	C.FSB	+0.0045
A.NFB	-0.064	C.NFB	+0.0140
A.FSF	-0.054	C.FSF	+0.0120
A.NFF	-0.067	C.NFF	-0.0057

Table 4. Performance against Average

A.FSB	0.0451
A.NFB	0.0420
A.FSF	0.0438
A.NFF	0.0379
C.FSB	0.0203
C.NFB	0.0202
C.FSF	0.0211
C.NFF	0.0182

Table 5. Standard Deviation of All systems

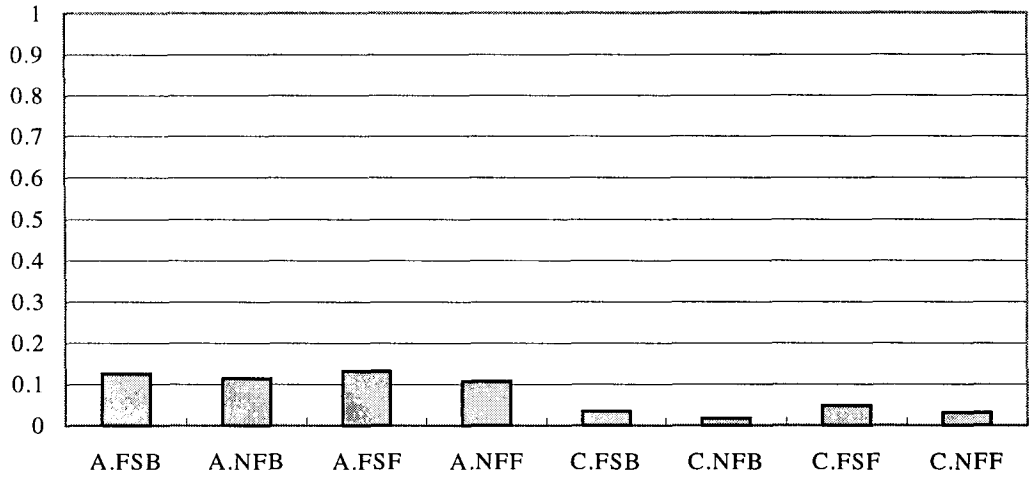

Figure 4. Comparison with the best participant

SUMMAC also conducts a series of baseline experiments to compare the system performance. From the report of these experiments, we find that for categorization task, the fixed-length summary is pretty good enough. For adhoc task, the best summary will do the better job. Another important finding is that the assessors are highly inconsistent. How to find out a fair and consistent evaluation methodology is worth further investigating.

6. CONCLUDING REMARKS

This paper proposes models to generate summary for two different applications. The first is to produce generic summaries, which do not take the user's information need into account. The second is to produce summaries, while the user's information need is an important issue. That is to say, the automatic summarization system interacts with users and takes user's query as a clue to produce user-oriented summaries. In addition, our approach is extract-based, which generates summaries using the sentences extracted from original texts. For the categorization task, the positive feature vector and the negative feature vector trained from the SUMMAC-1 texts are used as the comparative basis for sentence selection to produce generic summaries. As to adhoc task, the ES of each sentence is calculated based on the interaction of nouns and verbs. Then, the nouns of a query are compared with nouns in sentences and the closely related sentences are selected to form the summary. The result shows that the NormF of the best summary and that of the fixed summary for adhoc tasks are 0.456 and 0.447, respectively. The NormF of the best summary and that of the fixed summary for categorization task are 0.4090 and 0.4023, respectively. Our system outperforms the average system in categorization task but does a common job in adhoc task. We think that there are many further works to be studied in the future, e.g., extending the proposed approach to other languages, optimizing parameters of the proposed model, investigating the impact of errors introduced in tagging step, and developing a appropriate method to setup the threshold for sentence selection.

REFERENCES

[1] Bian, Guo-Wei and Chen, Hsin-Hsi (1997) "An MT Meta-Server for Information Retrieval on WWW." *Natural Language Processing for the World Wide Web*, AAAI-97 Spring Symposium, 10-16.

[2] Chen, Kuang-hua and Chen, Hsin-Hsi (1994) "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, New Mexico, USA, June 27-July 1, 234-241.

[3] Chen, Kuang-hua (1995) "Topic Identification in Discourse." *Proceedings of the 7th Conference of the European Chapter of ACL*, 267-271.

[4] Appelt, D.E. and Israel, D. (1997) *Tutorial on Building Information Extraction Systems*, Washington, DC.

[5] Weibel, S.; Godby, J. and Miller, E. (1995) *OCLC/NCSA Metadata Workshop Report*, (http://gopher.sil.org/sgml/metadata.html).

[6] Chen, Hsin-Hsi *et al.* (1998) "Description of the NTU System Used for MET 2." *Proceedings of the MUC-7 Conference*, forthcoming.

[7] Rush, J.E.; Salvador, R. and Zamora, A. (1971) "Automatic Abstracting and Indexing. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria." Journal of American Society for Information Sciences, 22(4), 260-274.

[8] Chen, Kuang-hua and Chen, Hsin-Hsi. (1995) "A Corpus-Based Approach to Text Partition."

Proceedings of the Workshop of Recent Advances in Natural Language Processing, Sofia, Bulgaria, 152-161.

[9] Sparck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation*, 28(1), 11-21.

[10] Church, K.W. and Hanks, P. (1990) "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics*, 16(1), 22-29.

[11] Jelinek, F. (1985) "Markov Source Modeling of Text Generation." In J.K. Skwirzynski (ed.), *The Impact of Processing Techniques on Communication*, Nijhoff, Dordrecht, The Netherlands.