國科會研究計畫結案報告

多語言資訊檢索與擷取—子計畫四：自動摘要方法之研究

單一中文文本之摘要(NSC89-2213-E-002-064)

多中文文本之摘要(NSC89-2218-E-002-041-)

多語言文本之摘要(NSC90-2213-E-002-045-)

1999.08.01 至 2002.07.31

主持人

陳信希 教授

台灣大學資訊工程學系

研究人員

陳光華	黃聖傑	林紋正	林偉豪	薛沛芸
陳偉珊	郭俊桔	林川傑	翁鴻加	蘇哲君

民國 91 年 8 月

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.

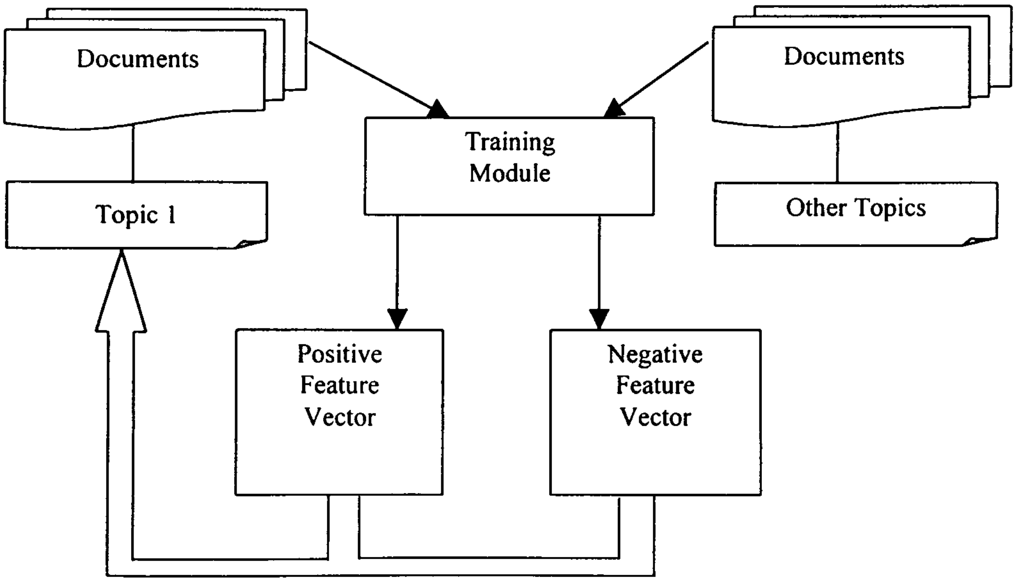


Figure 1. The Training Procedure for Categorization Task

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_i and NFV_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.

$$RS(S) = sim(PFV_i, S)\text{-}sim(NFV_i, S)$$

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.

$$D(w_1,w_2) = abs(C(w_1)\text{-}C(w_2)),$$

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.

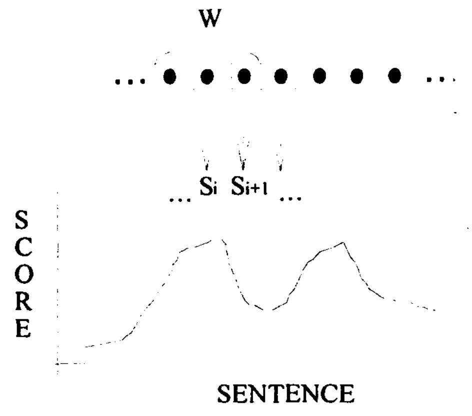


Figure 2. Determination of discourse segments

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

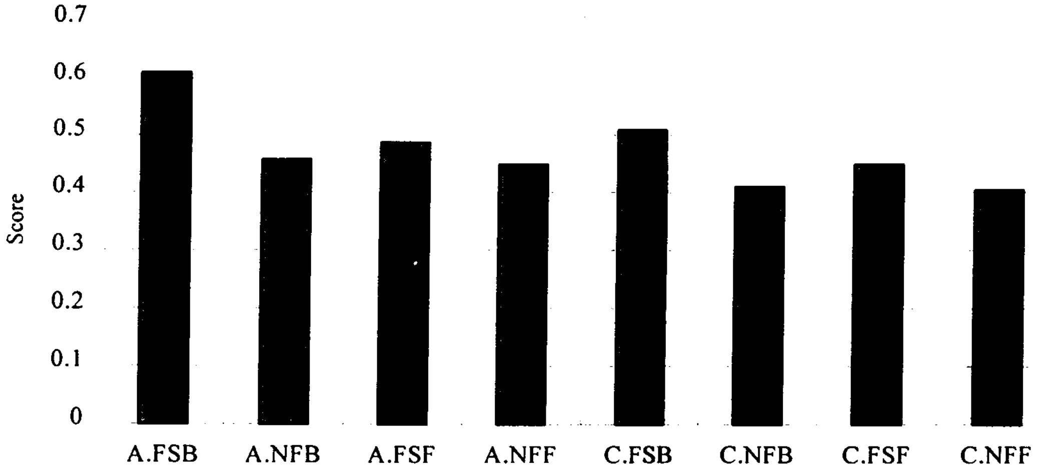


Figure 3. The performance of our system

- 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

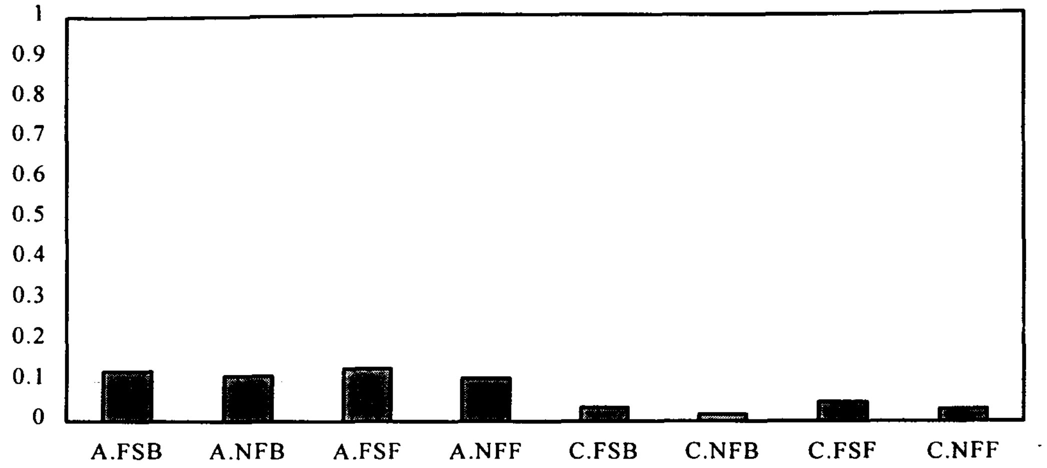


Figure 4. Comparison with the best participant

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.

論文題目：單一中文文本自動摘要問題之研究

(A Study on Automatic Summarization of a Chinese Text)

指導老師：陳信希

學生：林偉豪、薛沛芸、陳韋珊

科系：資訊工程學系

【摘要】

由於網際網路興起帶來的大量流通資訊，如何有效率地擷取文本中的重要資訊成為重要課題，自動摘要系統提出一種解決方法，但是目前中文世界卻一直沒有相關議題的討論。本文擬提出一套中文摘要自動化的流程分析，從最基本的統計模型出發，分析一般性文件中句子和主題的相似程度、以及句子中(名詞、名詞)與(名詞、動詞)的關聯性，提出兩種適用於中文上的統計性模型，由這種基於語料庫研究建立的語言模型，對其建構適當的評估方式，並實驗討論不同策略在中文上的適用程度。本系統希望能完成基本的單一中文文本自動摘要系統，奠定文件分類的基礎，未來推廣到多語系、多文件自動摘要系統上。

關鍵詞 Keywords:

自動摘要、資訊檢索、網際網路、分類任務、相似性、關聯性

一、 緒論

1-1 前言

由於資訊化時代的來臨，人們所需要處理的資訊量大增，又由於網際網路的興起使得文件流通率大幅提高，讀者能接觸到的層面比以往更加廣，要如何用最「有效率」的方法提供「最必要的」資訊，讓使用者從廣泛的各式文件中檢索到所需要的資訊，就成為網路應用上方興未艾的問題。也因此自動摘要的問題從第一次提出、到 1993 年在 Dagstuhl Seminar 中的討論、到去年的 SUMMAC(Firmin and Sundheim,1998)，一直是自然語言處理和資訊檢索領域中的熱門課題。

正由於人類做摘要的過程，牽涉到複雜的認知機制，所以要研製一套自動摘要系統不只要綜合人類在語言學上的知識，如言談結構 (discourse structure)、文法(grammar)、語法(syntax)、語意(semantics)，也要配合心理學、人工智慧、自然語言處理等領域的知識，以不斷更新提出更佳的解決方法；也正因為這種複雜度，所以至目前為止，學者專家仍未在這個問題的解決方式上達成一致的結論。

至於有關中文自動摘要的相關討論，導因於中文文章缺乏結構的本質，和相關文型分析的付之闕如，中文的自動摘要處理一直討論得並不多。其實以今日 BBS 系統在中文網路世界裡的盛行程度，網路文件分類可以想見為一個很重的工作，若是有一套中文摘要系統，可以增加許多資料查詢檢索的效率。但是，截至目前，並沒有見到一套可行的、針對全球第一大語系中文使用者提出的自動摘要系統。

本文的目標與焦點即放在提出一套適用在分類的中文自動摘要系統。從最基本的統計模型出發，分析一般性文件中句子和主題的相似程度、以及句子中(名詞、名詞)與(名詞、動詞)的關聯性，提出兩種在中文上適用的統計性模型，藉由這種基於語料庫研究建立的語言模型，對其建構適當的評分方式，並由所得的實驗數據，了解在英文上所採行的策略，在中文上應用的結果，討論其適用程度。

以下各節將逐次討論摘要特性與應用、相關研究的文獻探討、決定採用的模型流程、評分方式、解釋加入各式策略後的實驗結果。

1-2 摘要的特性分析

一般摘要可依照摘要動機概分為「指示性」(indicative)、「資訊提供性」(informative)；或依照摘要形成方法概分為「節選」(extraction)、「摘錄」(abstraction)；或依照摘要形成中使用者偏好介入程度，概分為「一般性」和「使用者偏好性」的。

當需要指出新的文件與已瀏覽過文件是否相關時，吾人可採用「指示性摘要」，但是當摘要需要提供給一個尚未閱讀過相關文章、也沒有相關背景知識的使用者有關該文件的完整資訊時，就需要提供有衍繹性的「資訊提供性摘要」；另一方面，若摘要採用揀取本文句子以忠實呈現原文意涵時，稱之為「節選」，但當句子經過摘要者合併修飾精簡後，所表現的是摘要者認定的重要資訊，這種摘要就稱為「摘錄」；當摘要選取概念時有加入參考使用者需要調整權重，稱為「使用者偏好性」的概要，對一般文件可以普遍適用的規則的摘要則為「一般性」摘要(Hovy, 1998)。

摘要分析又分為兩個層次，一個是語言學上的層次(linguistic)、一個是概念性的層次(conceptual)；根據 Hovey 和 Lin, 1998 提出的完整的摘要形成過程則經歷三階段：主題辨識、主題闡釋、摘要產生<Hovey and Lin, 98>。第一階段會用到諸如位置重要性、線索辭彙(cue phrases)、單字出現頻率、言談結構等在文本中、屬於語言學層次的資訊；第三階段所需要

的則是關鍵字、樣版產生器、句子組織等由外部提供、屬於概念性層次的資訊；第二階段則介於兩者之間，包含建立相關主題辭典、相關字彙連索鏈等等在概念融合、濃縮敘述時所需用到的資訊，並不純然是本文文本語句中可以取得的資訊，須有外界加入的概念以做判斷哪些是多餘的、可以被濾掉的敘述。

上文提到的「節選」到「摘錄」間所需要的就是第二階段融合觀念、闡釋主題的這個過程。

做到這三階段後的理想是一套能辨識概念符號(concept symbol)、自動學習的系統，能在概念性層次上做資訊闡釋的工作，且根據事先定義的概念規則、改版規則、言談結構計畫原則等在語言學層次上做濃縮工作。

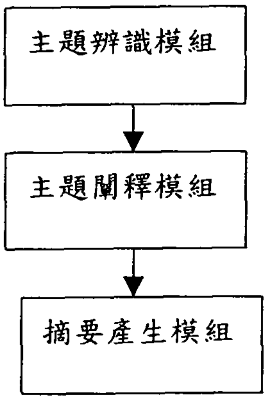


Fig 1-1 摘要產生三階段

目前學者專家的研究成果集中在第一階段的討論，第二、第三階段的討論還多停留於對特定領域的分析討論和概念性的介紹，如針對不同文化的文型(genre)特徵發展系統功能性語言學(System-Functional Linguistics，簡稱SFL)、修辭結構理論(Rhetorical Structure Theory，簡稱RST)、及針對某一種文類(如自傳的文型結構重視時間軸發展和人物動作)或某一種特殊文型（如固定規格的修繕報告）的摘要分析等等(Jones, 1995)。

由於這些模型仍尚屬概念層次，離在系統上實作還有差距，或者是過於限定適用領域，不具有一般性。所以本文考慮實用性與效率，如網路上的文件就不可能都在事先選定的領域或文型中，現階段又不可能專門針對各個領域建立字典與模型，而又因為各種文章在語言學層次上的特徵較概念性具有一般性原則，也就是說，實作起來效率會較佳；故選擇以又適用於「一般性」文件、又可透過自動學習修正改進的第一階段「主題辨識」方法為優先考慮，建立具通用性的統計模型來完成「節選」的工作，達成以「指示性」為動機的摘要。

二、 模型討論

承接以上討論，這裡討論自動摘要可採行的策略，提出兩種模型做為討論的基礎，並分析模型中跟語言有關的影響因素，以做為下一階段實驗的準備。

2-1 模型中所採用的方法

一般產生摘要方法可概分為”top-down”和”bottom-up”兩類，分別適用於不同的用途，看是要用在「節選重要資訊」的，還是為了「資訊檢索」方便所產生的摘要，適用的方法都有分別。

本文擬處理跟檢索有關係的主題導向摘要，所以採bottom-up的統計資料方法，由詞彙與詞彙的關係（特別是「名詞」與「動詞」）去了解個別詞彙重要程度、出現頻率（見Fig 2-1）、共同出現次數、互相參考次數、詞與詞間的距離會對彼此之間的相關性造成多大的影響、和由這些影響所衍申出來的共容語意代表的意義。

至於為什麼選擇「名詞」和「動詞」做為模型中討論的焦點，因為根據Fig2-1(Luhn, 98)在文章中扮演關鍵角色的並不是出現頻率最高

或最低的，而比較名詞、動詞、和其它轉折連接詞對各個句子的鑑別性，就會發現出現頻率過高的連接詞反而失去了對個別句子的代表性，只有功能性沒有鑑別性，如英文中所謂的"stop words"；反而是出現頻率在各種詞性裡不算最高的名詞、動詞，具有左右單句句意的功能，且對於上下文最具有承接情境的功能。

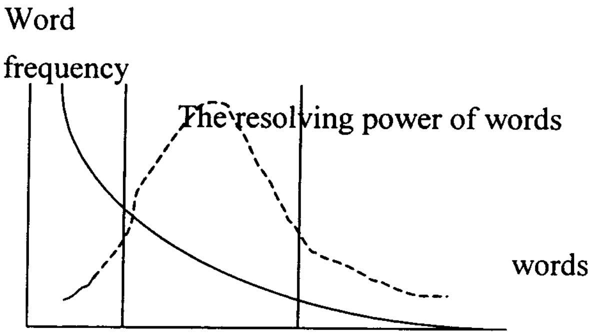


Fig 2-1 單字出現頻率與重要性相關圖 (Luhn, 98)

以下就把這些關係用兩種模型來檢視之：

1. 主題相似性模型

由於我們可以認定出現頻率和主題有其相關性。例如在某一主題文章中越常出現，而在其他種主題中越不常看到的詞彙，就可能是該主題的代表。

所以嘗試由訓練階段學習得各名詞、動詞和主題的相關性及非相關性，即在訓練階段提供的該主題文章中詞彙出現的頻率(PFV)、及在非該主題文章中詞彙出現的頻率(NFV)，來決定各詞彙和某一主題有關係或沒關係的程度，藉以找出特別相關的字作為主題關鍵字。

測試階段中，首先需要辨識出的是該篇文章的主題為何，接著就可以就在各句子中出現的各詞彙和主題的相似性決定句子的分數，選取分數高的作為節選對象。

其中辨識主題的方法為:

$$i = \arg\max_{1 \le i \le m}(Sim(PFVi, D) - Sim(NFVi、D))$$

選文件中計入某主題相關字和非相關字的影響後，得到的相似性最大者。

和主題的相似性(Similarity)定義為PFV和出現頻率的內積:

$$Sim(PFV, D) = \sum_{j=1}^{n}(Pwj \times dwj)$$

Pwj: Wj 的 PFV

Dwj: Wj 在文件中的頻率

句子分數的算法則定義為:

$$RS(S) = Sim(PFVi, S) - Sim(NFVi, S)$$

2. 名詞連接強度模型

由於不只名詞和動詞的個別出現頻率對單句重要性有影響，由名詞及動詞之間關係也代表了重要的句意資訊，所以以下這個模型計算詞彙的相互參考性，再加計考慮辭彙重複性、距離的影響，定義出名詞與名詞、或名詞與動詞之間的作用力，成為計算連結強度成為很重要的依據。

1. 距離因子：
   標計每一個名詞、動詞的出現順序，以相距的字數定義為距離。

$$D(w1, w2) = |C(w1) - C(w2)|$$

2. 共現強度：以逆文件頻率代表字的重要性，考慮距離因子，再將其對 N 和 V 的共現頻率作 normalization，分別算出名詞與動詞、名詞與名詞之間的共現強度。

$$SNV(Ni) = \sum\left(\frac{idf(Ni) \times idf(Vj) \times f(Ni, Vj)}{f(Ni) \times f(Vj) \times D(Ni, Vj)}\right)$$

$$SNN(Ni)=\sum(\frac{idf(Ni)\times idf(Vj)\times f(Ni,Vj)}{f(Ni)\times f(Vj)\times D(Ni,Vj)})$$

3. 關聯強度: 加總共現強度為各自的關聯性。

$$ANV(Ni,Vj)=\sum SNV(Ni,Vj)$$

$$ANN(Ni,Vj)=\sum SNV(Ni,Vj)$$

4. 考慮關聯性和距離的關係，標準化後加總

$$CSNN(Ni)=\sum\frac{ANN(Ni,Nk)}{D(Ni,Nk)}$$

$$CSNV(Ni)=\sum\frac{ANN(Ni,Vk)}{D(Ni,Vk)}$$

$$CS(Ni)=pn\times CSNN(Ni)+PV\times CSNV(Ni)$$

1. pn, pv 起始值設為 0,5

2. 算出

$$SN=\sum\frac{pn\times CSNN(Ni)}{pn\times CSNN(Ni)+pv\times CSNV(Ni)}$$

$$SV=\sum\frac{pv\times CSNN(Ni)}{pn\times CSNN(Ni)+pv\times CSNV(Ni)}$$

代回求

$$\text{pn}=\frac{SN}{SN+SV}\quad \text{pv}=\frac{SV}{SN+SV}$$

重複此一步驟，直到 pn,pv 收斂。

5. 以 Sliding window=3 算出

$$\text{Score(DS)}=\sum(\frac{CS(Ni)}{n})$$

由分數可以看出在這個 window 裡面的內容是否互相相關或是有良好組織的，若是分數低，則表示相關性較低，就可能是段落所在位置。

6. 計算

NCS(No(i))=

NCS(No(i-1)+(1-NCS(No(i-1))*CS(No(i))

不過要說明的是，我們採用這個模型配合學長以前完成的部份，採用的方法是直接算 $CS(N)=pn\times SNN(N)+pv*SNV(N)$

(令 pn,pv=0.5)

之後實驗的模型二也就以這個算式為準。

2-2 模型中的語言因素

有了基本的模型之後，要考慮模型中的語言因素，針對中文的特性對模型作些修正，討論有哪些經驗法則是可以再被加入的。

例如，有一些在外文摘要上被普遍應用的經驗法則是否合乎中文使用就需要被再驗證，如位置資訊、先導資訊、線索詞彙資訊、緊密度資訊、區域重要性資訊、句意連結鏈資訊等等，經判斷前三種較具有一般性，和中文的關係討論如下。

1. 位置資訊：

出現在標題和頭段的詞是對本文內容較具指示性的詞彙，所以可加重在標題中出現的詞彙權重。在中文文章也常發現這種特性。

2. 先導資訊：

每段中第一句通常都包含整段結論的重要資訊。所以摘要要收錄第一句。在中文文章中，這種特性並沒有特別明顯，但是在新聞類文章中常見

3. 線索詞彙資訊：

意指含有「線索詞彙」的句子即是主題句。這裡的線索詞彙是指常在代表結論的句子中出現的詞彙，如英文中的” the most important”, “importantly”等。

這個部份要用在中文中，要收集易於在結論句出現的線索詞彙，由於斷詞後，我們擁有各種詞性的資訊，我們之前把除了 N 和 V 開頭的其它詞性都濾掉，現在，可以靠經驗法則再多選取入其它的常見線索詞彙，如關聯連接詞(cbb)的「總之」、「是故」、「因為」、「因此」、「所以」、「只要」，副詞(D)的「首先」、「重要」、句副詞(Dk)的「簡言之」。(這裡的詞性以中研院平衡語料庫的分類為主)(黃居仁等, 1995)

4.專有名詞資訊：

推測比較少出現的專有名詞的出現通常都代表了某種特定意思。但是由於這種詞彙也有可能是出現在代表舉例的部份，若是這種只有在特例中出現的詞彙，就不具主題代表性。由於這種策略難已評估效果暫不考慮使用。

2-3 完整的系統流程

配合語言因素，列出整個流程如下：

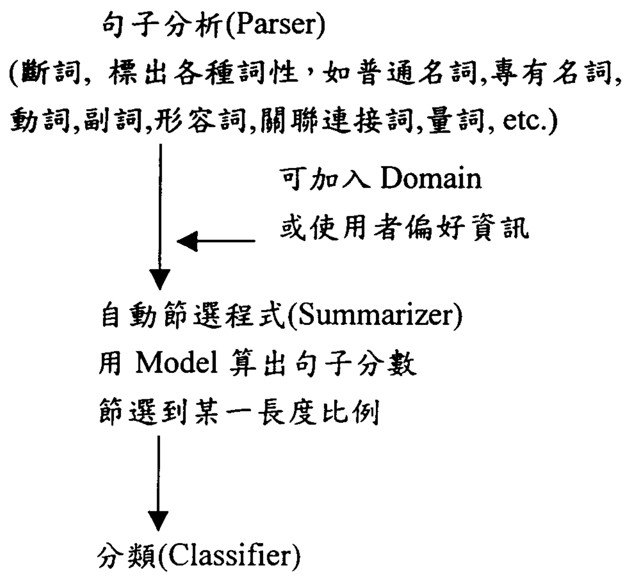


Fig 2-2 流程圖

(台大資訊系自然語言處理實驗室)

其中中文的剖析，實驗室已可做到九成六以上的正確辨識率，若我們使用平衡語料庫，則語料庫本身就已具備標記完成的詞性。

目前則還沒有專門針對中文文件考慮的分類程式，不過一般的分類程式稍加修改也有一定的效能，往後若是可以加入中文 stopwords 的討論，應該可以做得更好，不過這部份並不是本文的重點。

中文的節選程式，模型與語言因素探討如上節。

三、 實驗規畫

3-1 分類任務說明（categorization task）

有了模型之後，我們應該要整個模型和評估包裝起來，成為一個任務(task)，其目的在了解自動摘要系統是否有把分類上會用到的關鍵概念從文本中抓出來。

本文根據 1998 年在 SUMMAC-1(Firmin and Sundheim, 1998)採用的三種任務所適用的摘要型態：

- Categorization: 通用性、指示性摘要
- Adhoc: 使用者偏好、指示性摘要
- Q&A: 使用者偏好、資訊提供性摘要

採取第一類分類任務(Categorization task)作為本文的模型評估方法，以下就大略列出實驗步驟:

實驗一__模型一：

1. 訓練階段(training)

步驟一

使用中研院平衡語料庫(見附表一註釋)，取得如附表一列 67 類涵概各種語料型態的文字資料。從中選取兩大領域各四種主題做為實驗文本（見 Table2-1)。每種主題各 40 篇，其中有長有短，文體、語式並不固定。另外再加選一種非相關主題做為 noise，用來做為評估依據。每 40 篇裡取 32 篇作為訓練基準。(即訓練：測試=4：1)

Table 3-1 分類表

科學類
醫學 生物 天文 資訊
非相關主題
體育、建築、歷史

步驟二

去掉 function words

步驟三

計算各主題的各項 PFV, NFV，去掉其中小於 3 者。

2. 測試階段(testing)

步驟一

由各主題的關鍵字，算各主題與該文章的相似性(similarity)。

步驟二

考慮和各主題的不相似性，算出與各主題相似程度的分數

步驟三

選分數最高的作為辨識出的主題

步驟四

求出每個句子與主題相似程度的分數。由分數高者，依分數排行依次擷取入節選。

步驟五 輸出節選結果

fixed 摘要是取前 10%的句子，best-length 摘要則是從最好的取到 10%-50%。

實驗一__模型二：

步驟一 根據語料庫訓練出逆文件頻率(idf)、名詞動詞的關聯性(ANV、ANN)

步驟二 計算加總後各動詞、名詞之間的連結強度，算出每一名詞的強度。由是為基礎，計算 sliding window 中的分數。

步驟三 根據分數切份段落。

步驟四 計算段落分數。由高者開始選取。

步驟五 輸出節選結果如模型一。

實驗二：

重複上面的兩個階段，先訓練再測試。

只是訓練的文章數改為 20 篇，即訓練：測試＝5：5。

實驗三：

除了計算主題相似性，加入 lead method、title-based method、cue phrase-method 等經驗法則，再重複上面的流程，分別產生出 fixed length 和 best length 的摘要。

四、 評估流程

評估分成內在本質(intrinsic)和外來資訊(extrinsic)兩種，intrinsic 者意在評估摘要本身的性質，請評估者幫忙判斷流暢度，流暢度包括句子長度、結構的保留度、文法樣式的保留度，或是定好關鍵必要的資訊，然後看涵括度，或是觀察自動摘要跟理想摘要的相似性；extrinsic 者則不是從摘要本身出發，而是由外加評量具作為判斷摘要優劣標準，如問卷評量。

這裡採分類任務(categorization task)配合問卷調查，一方面是因為吾人所想要達到的目的是一個能在分類上有效能的摘要系統，所以採用分類方法；另一方面則是因為內在資訊較難有較公允的評量標準，所以採用這種可以量化的標準。

以下將每個實驗結果交由十個評估者和機器分類程式作分類的工作，以了解每個實驗對摘要系統的影響程度，並對每個實驗加做兩組「對照組」，一組是全文，一組是亂數檢取 10%的文本，詳列步驟如下:

步驟一 出示一些原文範例讓評估員了解原分類依據。

步驟二 分類。記錄 evaluator 在該項的 performance metrics 值。(evaluator 為人和機器分類程式)

(TP: True Positive、TN: True Negative、FP: False Positive、FN: False Negative)

	X	Y	非相關
X 類為真	TP	FN	FN
非相關為真	FP	FP	TN

步驟三 填回饋問卷（feedback）

1. 你在回答分類時的信心程度 low medium high(l m h)?

(在此給一篇原文)

2. 看過原文之後，你覺得這個 summary 的聰明程度(l m h)?

3. 長度適中程度(l m h)?

4. 你覺得一個讓你足以辨識主題的”好”摘要大致應該具備什麼條件?

5. 其他建議與改進?

<u>步驟四</u> 算召回率、精確率、NormF

1. 召回率(Recall)$=\dfrac{TP}{TP+FP}$

2. 精確率(Precision)$=\dfrac{TP}{TP+FN}$

3. Fallout$=\dfrac{FP}{FP+TN}$

4. Fscore$=\dfrac{2\times \mathrm{Pr}\mathit{ecision}\times \mathrm{Re}\mathit{call}}{\mathrm{Pr}\mathit{ecision}+\mathrm{Re}\mathit{call}}$

5. Compression(c) $=\dfrac{summarylength}{fulltextlength}$

6. NormF$=(1-\mathrm{c})\times$Fscore

五、 實驗結果討論

5-1 實驗數據

1. 實驗一—Model 一：

人做評估的部份：

	TP	FP	FN	TN	P	R	Fa	Fs
Full	.754	.018	.023	.206	.970	.977	.080	.973
Best	.733	.018	.044	.206	.976	.943	.080	.959
Fixed	.728	.018	.054	.200	.976	.931	.083	.953
Base	.667	.051	.077	.205	.897	.929	.200	.913

(P: precision R: recall F: fallout
Fa: Fallout Fs:F-score)

(Base: baseline 是 random 揀取的 10%文本)

問卷部份：

(1) 回答分類的信心程度

	Low	l-m	Med	m-h	High
Best			5		5
Fixed			4		6

(2)摘要的聰明程度：

	Low	l-m	med	m-h	High
Best			3	2	5
Fixed			4	1	5

(3)摘要長度的適中程度：

	Low	l-m	Med	m-h	High
Best			6		4
Fixed			4		6

用機器分類程式模擬人的行為做評估的部份：

	TP	FP	FN	TN
Full	.675	.025	.175	.175
Best	.6	.025	.2	.175
Fixed	.625	.025	.175.	.175
Base	.2	.025	.575	.175

實驗一—Model 二：

用機器分類程式模擬人的行為做評估的部份：

	TP	FP	FN	TN
Full	.675	.025	.175	.175
Best	.675	.025	.175	.175
Fixed	.6	.025	.2	.175
Base	.2	.025	.575	.175

5-2 結果討論

1. 召回率的意涵是這個系統實際上正確填對相關類別的文章數除以總相關文章數的比例；即這個系統對整體分類的效能。

   而精確度的意涵則是正確填對相關類別的文章數除以被認定為相關的總數；即這個系統能多正確地填入正確的類別。

   Fallout 則代表誤判為非相關的比例。

   F-score 則把這些效果合併顯示。

   這些數據可以幫我們了解系統分類的效能。可觀察到的整體趨勢是:
   FULLTEXT > BEST > FIXED>BASELINE

(不論是機器或是人來做的趨勢皆是如此)

而在做問卷的過程中，我們調出評估者分錯類別的文章，讓他們看一次原文決定應屬哪一類，有四成左右的評估者仍然會把原文放到錯誤的類別去，這顯示了受到每個評估者不同的背景知識影響，他們對類別的定義各自不同，也影響到分類的效果，若去除掉這種也許是因為文章類別原就不清楚、或是因為受測者本身認定的類別效應造成的誤差，系統的效能表現應該比數據表現得還要更好。

至於模型一和模型二比較起來差別並不大，但是在人眼判讀時，模型二因為是分段選取，彼此之間的句意連結性有被保留下來，所以可讀性較佳，對人來說比較親切。

舉例來看，模型一會選跟關鍵字較有關係的句子，模型二著重彼此之間的關聯性。

<例一>

模型一：

「ＵｎｉＳＱＬ目前取得國立漢城醫院的標案，以ＵｎｉＳＱＬＸ及ＵｎｉＳＱＬＭ多元資料庫管理系統（ＭｕｌｔｉｄａｔａｂａｓｅＳｙｓｔｅｍ，ＭＤＢＳ）來建構完成「健康中心資料管理系統」（ＨｅａｌｔｈＡｓｓｅｓｓｍｅｎｔＩｎｆｏｒｍａｔｉｏｎＳｙｓｔｅｍ，ＨＡＩＳ）此系統是以物件導向的最新技術完成醫院企業處理重構（ＨｏｓｐｉｔａｌＢｕｓｓｉｎｅｓｓＰｒｏｃｅｓｓＲｅｅｎｇｉｎｅｅｒｉｎｇ）的目標，它並可支援目前最尖端醫療技術的資訊服務業務。」

模型二：

「此系統解決了傳統健康診斷中心普遍存在的缺點，例如：候診時間長、診療結果取得耗時、對病患服務不佳、無法對病患資料做永久性管理等等問題，其對於診斷中心內所有的重要資訊（如：健康檢查、X光照射、胃腸透視、臨床病理等），都能在最短的時間內做出正確的處理，進而改善病患健康善和預防疾病防，達到整體工作效率的提升目標。」

<例二>

模型一：

「四度空間。再來就可以想像四度空間是怎樣的形狀了，依照上面的邏輯可推知四度空間的球體（可稱之爲超球），它的外表會是三度空間的彎曲，而超球的剖面將會是三度空間的球。因爲當在平面中遇有災害時唯有跳離此二度空間才有辦法得救，可是平面是無ｚ軸的，所以平面人是不可能逃離二度空間的；相同地，我們也無法從三度空間跳脫出來，因此當宇宙毀滅時，人類也是註定要毀滅的。」

模型二：

「以空間觀點來看，黑洞是宇宙吸入口，白洞則是噴出口，兩者以蟲洞相連，黑洞、白洞連接的宇宙可能是同一宇宙的兩個不同時間，也可能是兩個完全不同的宇宙藉著黑洞與白洞相連接，科學家大都承認黑洞的存在，因其可由巨大星球收縮而成，但對白洞的爭議可多了，因爲白洞只是我們在討論黑洞時，由方程式所推論出來的副產品而已，不過有的科學家認爲似星體是白洞，這只有待未來的科技來証明其對錯了。」

2. 觀察整體我們發現整體效能表現甚好，可能原因有幾個：

   (1) 我們選取的類別雖然都是科學類，可是由於分屬不同學門，之間的歧異性可以預期並不小，所以節選之後仍可以維持一定的鑒別特徵。

   (2) 人本身有自己長久累積的背景知識，配合這些知識，在看過幾篇文章後，可以很快地掌握類別的重點關鍵字，對往後的分類打下很有力的基礎。

   (3) 我們的模型中以單詞的重要性與關聯性為基礎抽句子，所以很容易保留下重要

的關鍵字，而分類任務是以分類為動機的評估方法，有沒有取到關鍵字會影響很大，所以使用者種模型在這種 task 中很有利。

3. 觀察我們選取的幾類文章，我們發現醫學和生物類的答對率較低，很明顯地我們可以發現類別較近，模糊性越高。
4. 再觀察各類文章的平均長度，生物類明顯偏低，其實 CKIP 在選材時特別強調過平衡特性，兼有或長或短的文章，但我們選擇的生物類剛好都是短篇的圖片描述，常在 100 字以內，訓練時期能學習到的資訊相對變少，影響抽取關鍵字時的準確性；另外一方面，節選出的句子相對少，「節選」摘要句子間缺乏連結的缺點易於突顯，讀者能接觸到的資訊也變少，以致於分類效能變差，也是在可以理解的範圍之內。
5. best-length(10%~50%)的摘要給予評估者的印象的確有比 fixed-length(10%)稍好，但在是否有助於幫助評估者做分類判斷上似乎並無幫助；可能的原因在於 10% 的摘要已經足夠於做分類判斷，這點可由評估者對於 best-length 摘要的長度適中程度較不滿意可以看出來。
6. 為了判斷同樣是 10%文章內容 的結果，我們加做了 random 取句子的部份做為 baseline 評估，可以看出有沒有經過揀選的抽取句子還是有差別的，並不是任何 10%的文本都可以包含到足夠判斷分類的重要資訊。
7. 機器分類程式與人的差別在於人有深厚的背景知識，機器之前只受過不到一百篇文章的的訓練，所以我們可以想像之前判斷力的差別，在調整過程中我們也發現訓練篇數的增加的確是有助於分類的準確性的，有些語料庫中資料較少的類別，如天文，被分類的效能就明顯低於其他類別。

5-3 改進建議

1. 若要採取機器分類方式來評估，要找更多不同模型、不同偏好的分類程式來模擬實際上多元化的人類想法。
2. 由使用者回饋的意見我們了解一個好的摘要，應該要重視簡明扼要，切中命題，抓到關鍵字、沒有多餘句子及流暢度五點，我們的摘要系統在最後一點流暢度上還有可以加強的地方。所以未來如要再改進，可以考慮分段之外，再處理 anaphora 首句重複的問題，以減少不必要重複的資訊。
3. 下一階段的自動摘要可以考慮加入主題闡釋或摘要產生階段的方法來增加摘要的可讀性。

六、 結論與未來工作

由於自動摘要牽涉到許多不同學門的知識，所以要解決這個問題必須以了解人類做摘要的心智歷程為出發點，探討各類摘要的特性、摘要系統採取的複雜技巧以及人類評估摘要的優劣的著眼點，才得以獲得可以採行於機器自動摘要系統的策略。

本文只是就目前可以做到的基本部份，如統計模型、經驗法則等等，做中文化的處理與改進，並嘗試不同的評估方法來檢驗效能。

未來的應用希望能針對網際網路和 BBS 電子布告欄上的文件做訓練，收集每日資訊做訓練工作，補強對各領域學習得的資訊，對網際網路的文件分類上有所助益；另一方面也希望能進一步推展到多文件的自動摘要，將一系列的文章做濃縮，更有效地達成中文資料的分類。

參考文獻

1. Chen, Kuang-Hua and Chen, Hsin-Hsi (1996) “Content Analysis-A Corpus-Based Model.” *Journal of Library Science*, **11**, 1996, 95-112.
2. Chen, Kuang-Hua; Huang, Sheng-Jie; Lin, Wen-Cheng and Chen, Hsin-Hsi (1998). "An NTU-Approach to Automatic Sentence Extraction for Summary Generation." *Proceedings of TIPSTER Text Phase III 18-Month Workshop*, Fairfax, VA, 4-6 May 1998.
3. Firmin Hand, T. and Sundheim, B.(1998) “TIPSTER-SUMMAC Summarization Evaluation.” *Proceedings of TIPSTER Text Phase III 18-Month Workshop*, Washington, 1998.
4. Hovy, Eduard and Marcu, Danial (1998) “A Guideline for Summary,” USC Information Sciences Institute.
5. Jones, Karen Sparck and Endres-Niggemeyer, Brigitte (1995) “Introduction: Automatic Summarizing.” *Information Processing & Management*, Vol.31, 1995, 625-630.
6. Luhn, H.P(1959) “The Automatic Creation of Literature Abstracts.” *IBM Journal of Research and Development*, 159-165.
7. Endres-Niggemeyer, Brigitte; Maier, Elisabeth and Sigel, Alexander (1995) “How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor.” *Information Processing & Management*, Vol.31, 1995, 631-674.
8. Brandow, Ronald; Mitze, Karl and Rau, Lissa F. (1995) “Automatic Condensation of Electronic Publications by Sentence Selection.” *Information Processing & Management*, Vol. 31, 1995, 675-685.
9. 黃居仁、陳克健、張莉萍、許蕙麗 (1995)“中央研究院平衡語料庫簡介”,第九屆計算語言學研討會論文集, 中華民國八十四年, 255-279。

[附表一]

中研院平衡語料庫(Sinica Corpus)(CKIP)由中研院詞庫小組在一九九五年發表推出，是世界上第一個有完整詞類標記的漢語平衡語料庫(黃居仁等，1995，詞庫小組，1995)，提供豐富的語言事實，涵括各種文型、文長和主題， 提供各種語法行為的統計資料。

包含主題分為六十七類如下：

16. 兒童文學
17. 經濟
18. 醫學
19. 消費
20. 社會現象
21. 內政
22. 生物
23. 衛生保健
24. 食物
25. 公益
26. 商管
27. 財政
28. 衣飾
29. 福利
30. 教育
31. 交通運輸
32. 體育
33. 批評與鑑賞
34. 政治現象
35. 犯罪
36. 國家政策
37. 家庭
38. 俠義文學
39. 政黨

40. 統計調查
41. 災禍
42. 國際關係
43. 政治學
44. 地理
45. 天文
46. 軍事
47. 傳播
48. 美術
49. 環保
50. 影藝
51. 行銷
52. 旅遊
53. 考古
54. 農漁牧業
55. 舞蹈
56. 戲劇
57. 技藝
58. 建築
59. 攝影
60. 鄉土文學
61. 雕塑
62. 數學
63. 文物
64. 大氣科學
65. 礦冶
66. 言勗學
67. 藝術總論

To appear in *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*. November 11-12, 1999.

A Summarization System for Chinese News from Multiple Sources

Hsin-Hsi Chen and Sheng-Jie Huang
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
E-mail: hh_chen@csie.ntu.edu.tw

Abstract

This paper will propose a personal news secretariat that helps on-line readers absorb news information from multiple sources. Such a news secretariat eliminates the redundant information in the news, and reorganizes the news for readers. This multiple document summarization employs named entities and other signatures to cluster news stream; employs punctuation marks, linking elements, and topic chains to identify the meaningful units (MUs); employs nouns and verbs to find the similarity of MUs; and finally employs focusing and browsing models to display the summarization results. On the average, the document reduction rates are 70.77% and 42.26% for focusing and browsing summarization, respectively. The reading-time reduction rate is 30.86%, and the correct rate of question-and-answering task is 88.46% for browsing summarization.

1. Introduction

Due to the development of Internet, information dissemination enters into a new era. Large scale multicultural, multimedia, and multilingual information is generated quickly, and crosses the geographic barrier to disseminate to different users. At such an information explosion age, how to filter useless information, and to absorb and employ information effectively becomes an important issue for users. Take news broadcast as an example. Many newspapers provide online services. Readers can access the online news sites quickly, but it takes much time for people to read all the news. This paper will propose a personal news secretariat that helps on-line readers absorb news information from multiple sources. Such a news secretariat eliminates the redundant information in the news, and reorganizes the news for readers.

Reorganization of news is some sort of document summarization, which tries to extract important information for users to save the reading time. Besides this application, document summarization can also help users decide document relevance. That will eliminate some degree of bottlenecks on information highway. The research of document summarization is very early (Edmundson, 1964; Edmundson, 1969), and is one of the traditional topics in research of natural language processing. Recently, it attracts new attentions due to the applications on Internet. Many papers about documentation summarization have been proposed (Hovy and Marcu, 1998). A special summarization evaluation *Summac-1* (Mani, *et al.*, 1998) organized by DARPA Tipster Text Program was held to deal with three kinds of evaluation tasks, i.e., categorization, ad hoc, and question-and-answering. However, most of the previous works are done on single document summarization. Only a few touch on multiple document summarization (Mani and Bloedorn, 1997; Radev and McKeown, 1998).

This paper extends our results on single document summarization in Summac-1 (Chen, *et al.*, 1998) to multiple Chinese news summarization. This paper is organized as follows. Section 2 presents the architecture of our summarization system. Section 3 specifies a news clusterer. Sections 4 and 5 deal with a news summarizer. Similarity analysis and presentation models are discussed respectively. Section 6 introduces our experiments and analyzes summarization performance. Finally, Section 7 concludes the remarks.

2. Architecture of a Summarization System

Figure 1 shows the architecture of our 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 by using events. An event is denoted by five basic entities such as people, affairs, time, places and things. The news articles for each event are summarized by a news summarizer. The tasks for the clusterer are listed below:

(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.

The tasks for the news summarizer are shown as follows:

(1) Partitioning a Chinese text into several meaningful units (MUs).
(2) Linking the meaningful units, denoting the same thing, from different news reports.
(3) Displaying the summarization results by two kinds of modes: a sequence of news by information decay, and a summarization by voting from reporters.

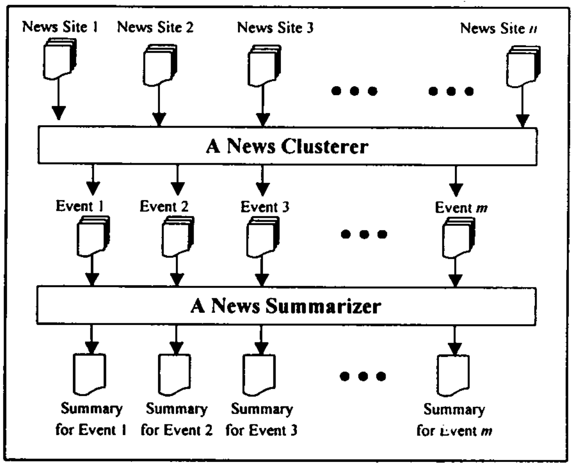


Figure 1. Architecture of a Summarization System

3. A News Clusterer

Because a Chinese sentence is composed of characters without boundaries, segmentation is indispensable. We employ a dictionary, some morphological rules and an ambiguity resolution mechanism for segmentation. Besides, we also extract named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions (Chen, *et al.*, 1998). Different types of information from different levels of text are employed, including character conditions, statistic information, titles, punctuation marks, organization and location keywords, speech-act and locative verbs, cache and *n*-gram model. The recall rates and the precision rates for the extraction of person names, organization names, and location names in the Chinese named entity extraction task of a famous message understanding competition (MUC, 1998) are (87.33%, 82.33%), (76.67%, 79.33%) and (77.00%, 82.00%), respectively.

When we apply the segmentation system in the summarization experiments, several errors shown as follows may affect the performance of summarization:

(1) Two sentences denoting the similar meaning may be segmented differently due to the segmentation strategies. Consider the following examples.

(E1) 但法務部長城仲模內定升任司法院副院長 ... is segmented into
但　法務部(Nc) 長城(Nc) 仲模(Nb) 內定(VC) 升任(VG)　司法院(Nc)　副院長(Na) ...

(E2) 而城仲模轉任司法院副院長之後的法務部長遺缺 ... is segmented into
而　城仲模(Nb)　轉任(VG)　司法院(Nc)　副院長(Na)　之後(Ng)　的　法務(Na)　部長(Na) 遺缺(Na)　...

The major title "法務部長" and the major person "城仲模" are segmented in different ways in the above examples. That will introduce errors in similarity analysis.

(2) Unknown words generate many single-character words. For example, "土(Na)　石(Na)　流(VC)", "園(Nc) 山(Na)　村(Nc)", "芭(Nb)　比(VC)　絲(Na)", "老(VH)　丙(Neu)　建(VC)", and so on. After tagging, these words tend to be nouns and verbs, which are used in computing the scores for similarity measure. Thus errors may be introduced.

We adopt a two-level approach to cluster the news from multiple sources. At first, news is classified on the basis of a predefined topic set. Then, the news articles in the same topic set are partitioned into several clusters according to named entities. Clustering is necessary. On the one hand, a famous person may appear in many kinds of news stories. For example, President Clinton may make a public speech (political news), join an international meeting (international news), or even just show up in the opening of a baseball game (sports news). On the other hand, a common name is frequently seen but denotes different persons. Clustering helps reducing the ambiguity introduced by famous persons and/or common names.

4. Similarity Analysis

4.1 Meaning Units

The basic idea in our study is to tell the similarity of the news articles in the same event. The basic unit for similarity checking may be a paragraph or a sentence. For the former, text segmentation is necessary for documents without paragraph markers (Chen and Chen, 1995). For the latter, text segmentation is necessary for languages like Chinese. Because Chinese writers often assign punctuation marks at random (Chen, 1994), the sentence boundary is not clear. Consider Example (E3):

(E3) 西班牙裔　是　美國　少數　族裔　人口　成長　最快　的　一　支　，這 股　支持　力量　將　使　喬治　未來　在　與　共和黨　內　的　提名　競爭者　相較　之下　，　別　具　優勢　。

It is composed of three sentence segments separated by commas. We will find two meaningful units (MUs) shown as Examples (E4) and (E5):

(E4) 西班牙裔　是　美國　少數　族裔　人口　成長　最快　的　一　支

(E5) 這 股　支持　力量　將　使　喬治　未來　在　與　共和黨　內　的　提名　競爭者　相較　之下　，　別　具　優勢

Here a MU that is composed of several sentence segments denotes a complete meaning.

Three kinds of linguistic knowledge – punctuation marks, linking elements and topic chains, are used to identify the MUs.

(1) Punctuation marks
There are fourteen marks in Mandarin Chinese (Yang, 1981). Only period, question mark, exclamation mark, comma, semicolon and caesura mark are employed. The former three are sentence terminators, and the latter three are segment separators.

(2) Linking elements
There are three kinds of linking elements (Li and Thompson, 1981): forward-linking elements, backward-linking elements, and couple-linking elements. A segment with a forward-linking (backward-linking) element is linked with its next (previous) segment. A couple-linking element is a pair of words that exist in two segments. Apparently, these two segments are joined together. Examples (E6)-(E8) show each kind of linkings.

(E6) forward linking
因為天氣不好，飛機改在明天起飛。

(E7) backward linking
你先把資料準備好，以便開會用。

(E8) couple linking
他一邊走路，一邊唱歌。

(3) Topic chains
The topic of a clausal segment is deleted under the identity with a topic in its preceding segment. The result of such a deleting process is a *topic chain*. Thus we have the following postulation (Chen, 1994): given two VP segments, or one S and one VP segments, if their expected subjects are unifiable, then the two segments can be linked. This paper applies the postulation to parse each segment independently, and compose a parsing tree from the trees of segments. That will reduce the complexity of parsing very long Chinese sentences. In our summarization system, we do not parse the Chinese sentences actually. We employ part of speech information only to predict if a subject of a verb is missing. If it does, we postulate that it must appear in the previous segment and the two segments are connected to form a larger unit. Consider Example (E9).

(E9) 國民黨是靠組織起家的政黨，現在的組織體質卻很虛弱，所以選戰最後也要仰仗文宣，實在很可惜。

There are four segments in this example. The words "卻" (but) and "所以" (therefore) in the second and the third segments are backward linking

noun-sim(MU1,MU2)=0.89, and verb-sim(MU1,MU2)=0.82. The basic ideas for the other models are dealt with in the

affected.

(E10) 為了 宣示(VE) 穩定(VHC) 股市(Nc) 進行(VC)

elements. Thus the three segments are connected together. The last segment does not have any subject, so that it is connected to the previous one. In summary, these four segments form a MU. Here we do not touch on the associative problem mentioned in parsing (Chen, 1994). We just care about which segments can be formed a MU.

4.2 Similarity Models

The next step is to find the similarity among MUs in the news articles reporting the same event, and to link the similar MUs together. Predicate-argument structure forms the kernel of a sentence, thus verbs and nouns are important clues for similarity measures. We consider several strategies shown as follows:

(S1) Nouns in one MU are matched to nouns in another MU, so are verbs.

(S2) The operations in (1) are exact matches.

(S3) A Chinese thesaurus, i.e., 同義詞詞林 (tong2yi4ci2ci2lin2), (Mei, *et al.*, 1982), is employed during the matching. That is, the operations in (S1) may be relaxed to inexact matches.

(S4) Each term specified in (S1) is matched only once.

(S5) The order of nouns and verbs in MU is not considered.

(S6) The order of nouns and verbs in MU is critical, but it is relaxed within a window.

(S7) When continuous terms are matched, an extra score is added.

(S8) When the object of transitive verbs are not matched, a score is subtracted.

(S9) When date/time expressions and monetary and percentage expressions are matched, an extra score is added.

The similarity of two MUs is in terms of noun-similarity and verb-similarity:

$$noun\text{ - }sim(A,B) = \frac{m}{\sqrt{ab}}$$

$$verb\text{ - }sim(A,B) = \frac{n}{\sqrt{cd}}$$

where m (n) denotes the number of matched nouns (verbs),
a and b denote total number of nouns in MUs A and B respectively,
c and d denote total number of verbs in MUs A and B respectively.

Five models shown below are constructed under different combinations of the strategies specified in the above.

(M1) strategies (S1)+(S3)+(S4)+(S5)
(M2) strategies (S1)+(S3)+(S4)+(S6)
(M3) strategies (S1)+(S3)+(S4)+(S5)+(S7)+(S8)
(M4) strategies (S1)+(S3)+(S4)+(S5)+(S7)+(S8)+(S9)
(M5) strategies (S1)+(S2)+(S4)+(S5)+(S7)+(S8)+(S9)

Consider the following two MUs as an example for M1 model:

MU1: 國父紀念館(Nc) 展出(VC) 的 國父(Na) 書畫(Na) 、 墨寶(Na)， 傳出(VC) 了 失蹤(VH) 的 消息(Na)

MU2: 國父紀念館(Nc) 傳出(VC) 了 國父(Na) 書畫(Na) 失蹤(VH) 的 消息(Na)

In this example, m=4, n=2, a=5, b=4, c=3, and d=2. Thus, noun-sim(MU1,MU2)=0.89, and verb-sim(MU1,MU2)=0.82. The basic ideas for the other models are dealt with in the following section.

4.3 Preparation of Testing Corpus

Nine events selected from Central Daily News, China Daily Newspaper, China Times Interactive, and FTV News Online in Taiwan are used to measure the performance of each model. They are shown as follows:

(1) 社會役的實施 (military service): 6 articles
(2) 老丙建建築 (construction permit): 4 articles
(3) 三芝鄉土石流 (landslide in Shan Jr): 6 articles
(4) 總統布希之子 (Bush's sons): 4 articles
(5) 芭比絲颱風侵台 (Typhoon Babis): 3 articles
(6) 股市穩定基金 (stabilization fund): 5 articles
(7) 國父墨寶失竊案 (theft of Dr Sun Yat-sen's calligraphy): 3 articles
(8) 央行調降利率 (interest rate of the Central Bank): 3 articles
(9) 內閣總辭問題 (the resignation issue of the Cabinet): 4 articles

The news events are selected from different boards. An annotator reads all the news articles, and connects the MUs that discuss the same story. Because each MU is assigned a unique ID, the links among MUs form the answer keys for the performance evaluation.

4.4 Experiment Results

Traditional precision and recall are computed. Table 1 lists the performance of these five models. The thresholds for noun-similarity and verb-similarity are set to 0.3. M1 is regarded as a baseline model. M2 is different from M1 in that the matching order of nouns and verbs are kept conditionally. It tries to consider the subject-verb-object sequence. The experiment shows that the performance is worse. The major reason is the syntax of Chinese sentences is not so restricted. We can express the same meaning using different syntactic structures. Movement transformation like topicalization, relativization, ba-construction and bei-construction affects the order of subject-verb-object. Thus in M3 we give up the order criterion, but we add an extra score when continuous terms are matched, and subtract some score when the object of a transitive verb is not matched. Compared with M1, the precision is a little higher, and the recall is improved about 4.5%. If we further consider some special named entities such as date/time expressions and monetary and percentage expressions in M4, the recall is increased about 7.6% at no expense of precision. M5 tries to estimate the function of the Chinese thesaurus. It uses exact matching. The precision is a little higher, but the recall is decreased about 6% compared with M4.

Table 1. Performance of Similarity of MUs

Model	Precision	Recall
M1	0.5000	0.5434
M2	0.4871	0.3905
M3	0.5080	0.5888
M4	0.5164	0.6198
M5	0.5243	0.5579

Several major errors affect the overall performance. Using nouns and verbs to find the similar MUs is not always workable. The same meaning may not be expressed in terms of the same words or synonymous words. Examples (E10) and (E11) talk about the same event, but use different verbs. The similarity contributed from verb is 0, thus the recall is affected.

(E10) 為了 宣示(VE) 穩定(VHC) 股市(Nc) 進行(VC) 第(Neu) 一(Neu) 次 的 穩定(VHC) 股市(Nc)

專案(Na) 小組(Na) 會議(Na)
(E11) 專案(Na) 小組(Na) 昨日(Nd) 召開(VC) 第(Neu)
一(Neu) 次 會議(Na)

Besides. we can use different format to express monetary and percentage expressions. For example. "二千八百三十億元" (two hundreds and eighty-three billions) in Chinese can be written as "二八三〇億元" or "2830 億". Similarly. the percentage expression "百分之七點二五" (seven point two five percent) can be expressed in terms of "七・二五%" or "7.25%". Segmentation is another source of errors. On the one hand, although our segmentation system takes care of named entities, there are still many new words in news articles. The new-invented words are segmented into a sequence of single-character words. On the other hand, the dictionary used in segmentation and the thesaurus used in the inexact matching are not integrated together in our experiments. Total 40% of nouns and 21% of verbs are not found in the thesaurus.

5. Presentation Models

Two models, i.e., focusing model and browsing model, are proposed to display the summarization results. In the focusing model, a summarization is presented by voting from reporters. For each event, a reporter records a news story from his own viewpoint. Recall that a news article is composed of several MUs. Those MUs that are similar in a specific event are common focuses of different reporters. In other words, they are worthy of reading. In the current implementation, the MUs that are reported more than twice are our target. For readability, the original sentences that cover the MUs are selected. For each set of similar MUs, only the longest sentence is displayed. The display order of the selected sentences is determined by relative position in the original news articles. That is. if a sentence appears in the introduction part of a news article. it tends to be displayed in the front of focusing summarization. Appendix A lists the focusing summarization for a typhoon news (event 5).

In the browsing model. the news articles are listed by information decay. The first news article is shown to the user in its whole content. In the latter shown news articles. the MUs denoting the information mentioned before are shadowed (or eliminated), so that the reader can focus on the new information. The amount of information in a news article is measured in terms of the number of MUs, so that the article that contains more MUs is displayed before the others. For readability, a sentence is a display unit. Appendix A also demonstrates the browsing summarization. In this model, users can read both the common views and different views of reporters. It saves the reading time by listing the common view only once.

6. Experiments and Evaluations

The same nine events specified in Section 4 are used to measure the performance of the two summarization models. Three kinds of measure are considered – say, the document reduction rate, the reading-time reduction rate, and the information carried. The higher the document reduction rate is, the more time the reader may save, but the higher possibility the important information may be lost. Tables 2 and 3 list the document reduction rates for focusing and browsing summarization, respectively. Only focuses are displayed in focusing summarization. so that the average document reduction rate is higher than that of browsing summarization. Table 4 further lists the ratio of summary to full article in each news. In most cases, we have good reduction rates after the second news article is read. The only exception is event 4. After analysis. we find that the recall rate of similarity measure of MUs is only 30%, but the precision rate is 84% in this case.

Table 2. Reduction Rates for Focusing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction%
1. military service	7658	2402	0.3137	68.63%
2. construction permit	4182	1226	0.2932	70.68%
3. landslide in Shan Jr	5491	1823	0.3320	66.80%
4. Bush's sons	6186	924	0.1494	85.06%
5. Typhoon Babis	4068	1460	0.3589	64.11%
6. stabilization fund	8434	2243	0.2659	73.41%
7. theft of Dr Sun Yat-sen's calligraphy	4576	1524	0.3330	66.70%
8. interest rate of the Central Bank	4578	1690	0.3692	63.08%
9. the resignation issue of the Cabinet	4980	1368	0.2747	72.53%
Average	50153	14660	0.2923	70.77%

Table 3. Reduction Rates for Browsing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction%
1. military service	7658	2716	0.3547	64.53%
2. construction permit	4182	2916	0.6973	30.27%
3. landslide in Shan Jr	5491	2946	0.5365	46.35%
4. Bush's sons	6186	5098	0.8241	17.59%
5. Typhoon Babis	4068	2270	0.5580	44.20%
6. stabilization fund	8434	4299	0.5097	49.03%
7. theft of Dr Sun Yat-sen's calligraphy	4576	2840	0.6206	37.94%
8. interest rate of the Central Bank	4578	2682	0.5858	41.42%
9. the resignation issue of the Cabinet	4980	3190	0.6406	35.94%
Average	50153	28957	0.5774	42.26%

Table 4. Ratio of Summary to Full Article in Browsing Summarization

Article\ Event	1	2	3	4	5	6	7	8	9
1	100%	100%	100%	100%	100%	100%	100%	100%	100%
2	12%	84%	68%	71%	56%	39%	27%	29%	67%
3	32%	36%	68%	77%	0%	7%	49%	24%	50%
4	24%	47%	10%	79%		51%			24%
5	12%		0%			17%			
6	0%		9%						

Besides the document reduction rate, we also measure the correct rate of question-answering, and reading-time reduction rate. Assessors read the highlight parts only in the browsing summarization, and answer 3 to 5 questions. Appendix B demonstrates the evaluation procedure of accessors. Table 5 lists the evaluation results of the first six events. The average document reduction rate is 43.79%. On the average. the summary saves 30.86% of reading time. While reading the summary only, the correct rate of question-answering task is 88.46%.

Table 5. Assessors' Evaluation

Event Name	Document Reduction Rate	Question-Answering Correct Rate	Reading-Time Reduction Rate
1. military service	64.53%	100%	45.24%
2. construction permit	30.27%	33.33%	33.54%
3. landslide in Shan Jr	46.35%	80%	10.28%
4. Bush's sons	17.59%	100%	36.49%
5. Typhoon Babis	44.20%	100%	35.10%
6. stabilization fund	49.03%	100%	18.49%

Average	43.79%	88.46%	30.86%

7. Concluding Remarks

This paper presents a multiple Chinese news summarization system. A two-level clusterer is employed to collect news articles for the same event. A summarizer identifies the similar MUs from news articles belonging to the same event, and displays the summarization results based on two strategies. The information decay strategy helps reducing the redundancy, and the user can get all the information provided by the news. But it may still read too much. Besides, the order of the sequence is not according to the importance. The user may quit reading and miss the information not shown yet. The voting strategy gives a shorter summarization, on the other hand, also misses some unique information reported by some news sites. A hybrid strategy should be studied to meet all the requirements. Besides, an important event is always reported in different languages. How to extend the summarization system to absorb news from multilingual news sites is a new challenge on the Internet.

References

Chen, H.H. (1994) "The Contextual Analysis of Chinese Sentences with Punctuation Marks," *Literal and Linguistic Computing*, Oxford University Press, 9(4), 1994, pp. 281-289.

Chen, H.H., *et al.* (1998) "Named Entity Extraction for Information Retrieval," *Computer Processing of Oriental Languages*, Special Issue on Information Retrieval on Oriental Languages, 12(1), 1998, pp. 75-85.

Chen, K.H, *et al.* (1998) "An NTU-Approach to Automatic Sentence Extraction for Summary Generation," *Proceedings of TIPSTER Text Phase III 18-Month Workshop*, Fairfax, VA, 4-6 May 1998.

Chen, K.H. and Chen, H.H. (1995) "A Corpus-Based Approach to Text Partition," *Proceedings of International Conference of Recent Advances on Natural Language Processing*, Tzigov Chark, Bulgaria, 1995, pp. 152-160.

Edmundson H.P. (1964) "Problems in Automatic Extracting," *Communications of the ACM*, 7, 1964, pp. 259-263.

Edmundson H.P. (1969) "New Methods in Automatic Extracting." *Journal of the ACM*, 16, 1969, pp. 264-285.

Hovy, E. and Marcu, D. (1998) "Automated Text Summarization," Tutorial in *COLING/ACL98*, 1998.

Li, C.N. and Thompson, S.A. (1981) *Mandarin Chinese: A Functional Reference Grammar*, University of California Press, 1981.

Mani, I. and Bloedorn, E. (1997) "Multi-document Summarization by Graph Search and Matching," *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, RI, pp. 622-628.

Mani, I., *et al.* (1998) *The TIPSTER SUMMAC Text Summarization Evaluation: Final Report*, Technique Report, Automatic Text Summarization Conference, 1998.

Mei, *et al.* (1982) *tong2yi4ci2ci2lin2*, Shanghai Dictionary Press, 1982.

MUC (1998) *Proceedings of 7th Message Understanding Competition*, http://www.muc.saic.com /proceedings/proceedings_index.html.

Radev, D.R. and McKeown, K.R. (1998) "Generating Natural Language Summaries from Multiple On-Line Sources," *Computational Linguistics*, Vol. 24, No. 3, pp. 469-500.

Yang, Y. (1981) *The Research on Punctuation Marks*, Tian-jian Publishing Company, Hong Kong, 1981.

Appendix A Summarization Results

There are two kinds of summarization results, i.e., focusing summarization and browsing summarization. In the former, only the focuses are listed. The complete news articles can refer to the browsing summarization part. In the latter, the summarization is highlighted by the symbol □.

(1) Focusing Summarization

1. 受芭比絲颱風路徑偏東影響，中央氣象局昨（廿六）日晚間九時發布全台灣地區陸上颱風警報，臺灣地區全部為颱風警戒區域，預估今天上午起南部地區將進入芭比絲颱風暴風圈內，颱風中心距本島一百五十公里沿臺灣海峽北上，明（廿八）天上午以後暴風圈才會逐步脫離本島。
2. 中央氣象局昨日晚間針對臺灣本島、澎湖、金門、馬祖等地區及各海域發出輕度颱風芭比絲的海上、陸上颱風警報，由於芭比絲颱風外圍環流挾帶著豐沛的水氣，直撲臺灣西南部，在沒有中央山脈屏障的情況下，原先僅出現於臺灣東半部、北部地區的豪大雨將擴及全省，氣象局特別呼籲西南部沿海低漥地區應嚴防海水倒灌。
3. 中央氣象局指出，芭比絲颱風暴風圈接觸廣東陸地後，強度已在昨天下午二時減弱為輕度颱風，芭比絲颱風昨晚位於澎湖西南方二百六十公里海面上，預測以時速十三公里朝向北北東轉東北移動，中心附近最大風速每秒三十公尺，相當於十一級風，瞬間最大陣風每秒四十公尺，相當於十三級風，七級風暴風半徑二五Ｏ公里，預測今天晚間八時的中心位置在金門東方一ＯＯ公里之海面上。
4. 首當其衝的澎湖和金門，今日清晨即已進入暴風圈內，並出現十二級的最大陣風。
5. 根據天氣資料顯示，臺中以南及花蓮以南地區的風雨自昨日晚間起已經愈來愈強，特別是因為西南部地區沒有山脈屏障，氣象局預測今日午後平池地區一般風力可達五到六級，陣風可達九到十級，濱海地區風力則有六到七級，陣風可達十一級以上，各地的風力及雨勢都將相當驚人。

(2) Browsing Summarization

(A) The first news article

受芭比絲颱風路徑偏東影響，中央氣象局昨（廿六）日晚間
九時發布全台灣地區陸上颱風警報，臺灣地區全部為
颱風警戒區域，預估今天上午起南部地區將進入芭比
絲颱風暴風圈內，颱風中心距本島一百五十公里沿臺
灣海峽北上，明（廿八）天上午以後暴風圈才會逐步
脫離本島。

因無地形屏障本島西半部風雨將非常劇烈，東石、安平沿海
居民應嚴防海水倒灌。

中央氣象局指出，芭比絲颱風暴風圈接觸廣東陸地後，強度
已在昨天下午二時減弱為輕度颱風，芭比絲颱風昨晚
位於澎湖西南方二百六十公里海面上，預測以時速十
三公里朝向北北東轉東北移動，中心附近最大風速每
秒三十公尺，相當於十一級風，瞬間最大陣風每秒四
十公尺，相當於十三級風，七級風暴風半徑二五Ｏ公
里，預測今天晚間八時的中心位置在金門東方一ＯＯ
公里之海面上。

氣象局表示，芭比絲颱風將沿著臺灣海峽北上，在左側有福
建武夷山、右側為本島中央山脈的挾擠下，未來颱風
威力可能繼續減弱，但民眾仍不可掉以輕心，芭比絲

是近年來少數由臺灣海峽方向行進的颱風，本島西半部人口密集且無地形可以屏障，西部地區民眾可明顯感受到芭比絲颱風之風、雨威力。

氣象局指出，芭比絲颱風暴風圈將在今天清晨接觸金門澎湖，上午隨著颱風北移高雄臺南將逐漸進入暴風圈，南部地區最大陣風將達十級，西南沿海之東石、安平等地應嚴防海水倒灌，北部地區預計下午起進入芭比絲颱風暴風圈內，風雨最大時間應在入夜以後，預估芭比絲颱風暴風圈要到二十八日上午才會完全脫離臺灣地區。

氣象局統計，在東北季風與芭比絲颱風雙重影響下，不到三天時間內宜蘭新寮降雨量已達一千二百五十七公釐、花蓮太安九百四十公釐、蘇澳七百四十二公釐、火燒寮六百六十八八釐、花蓮六百五十九公釐、五堵六百二十六公釐、基隆五百六十三公釐、臺北市南港三百九十六公釐。

(B) The second news article

中央氣象局表示，芭比絲颱風昨晚接近北緯二十三度時，受到強盛西風的吹拂，行進方向從北北東轉向東北，暴風圈今日沿著台灣海峽南部北上，橫掃台澎金馬。

中央氣象局已在昨晚對全台灣地區發布陸上警報，呼籲各地民眾今日都應嚴防芭比絲的強風豪雨侵襲。

氣象局預報中心課長王世堅指出，芭比絲昨晚以每小時十三公里的速度，向北北東前進，進入台灣海峽南端，逐漸逼近北緯二十三度西風帶後，隨即受到強盛西風的吹拂，向東偏轉，穩定地向東北進行，沿著台灣海峽北上。

根據昨晚資料，芭比絲已在昨天夜間登陸廣東汕頭沿海地區。

王世堅表示，芭比絲颱風中心雖不致直接登陸台灣，但因暴風半徑達二百五十公里，在其沿著台灣海峽北上的過程中，全台各地先後被籠罩在暴風圈範圍內。

首當其衝的澎湖和金門，今日清晨即已進入暴風圈內，並出現十二級的最大陣風。

高屏地區也將從中午起逐漸進入暴風圈，風雨明顯增強。

雲嘉南下午開始風雨會轉大，北部要到傍晚才會真正起風，東半部因處於背風面，風力較小。

王世堅分析說，芭比絲暴風圈將在今日下午到晚上之間，將全台灣地區團團圍住。

這段期間也是颱風中心最接近台灣陸地的時刻。

颱風眼將從嘉南外海一百五十公里處通過，此時也是各地風雨最強的時刻，除須防範豪雨成災，西南部沿海也要提防海水倒灌。

根據氣象局統計，從二十四日到昨晚八時為止，芭比絲外圍環流已經為台灣地區帶來超過一千公厘的豪雨，降雨量最多的地方是宜蘭縣新寮山區的一二五七公厘，花蓮縣太安山區雨量也高達九四〇公厘，東半部平地的雨量都在五百到八百公厘之間，北部山區和基隆山區雨量也都超過六百公厘。

氣象局預測，各地雨勢將隨著暴風圈從今天傍晚起，從南而北逐漸脫離台灣陸地後漸趨緩和，已經豪雨成災的北部和東北部則要等到明天上午雨勢才會轉小。

(C) The third news article

中央氣象局昨日晚間針對臺灣本島、澎湖、金門、馬祖等地區及各海域發出輕度颱風芭比絲的海上、陸上颱風警報，由於芭比絲颱風外圍環流挾帶著豐沛的水氣，直撲臺灣西南部，在沒有中央山脈屏障的情況下，原先僅出現於臺灣東半部、北部地區的豪大雨將擴及全省，氣象局特別呼籲西南部沿海低漥地區應嚴防海水倒灌。

氣象局指出，芭比絲颱風強度雖已由中度減弱為輕度，但外圍環流仍非常強，且芭比絲的中心逐漸靠近臺灣海峽南部，自昨晚起全省陸續進入颱風警戒區，預計今日中午至傍晚將是風雨影響臺灣地區最嚴重的時段。

根據天氣資料顯示，臺中以南及花蓮以南地區的風雨自昨日晚間起已經愈來愈強，特別是因為西南部地區沒有山脈屏障，氣象局預測今日午後平地地區一般風力可達五到六級，陣風可達九到十級，濱海地區風力則有六到七級，陣風可達十一級以上，各地的風力及雨勢都將相當驚人。

氣象局預報中心主任陳來發指出，雖然芭比絲有逐漸減弱成為熱帶性低氣壓的可能，但對臺灣地區引進的豪雨量仍相當可觀。

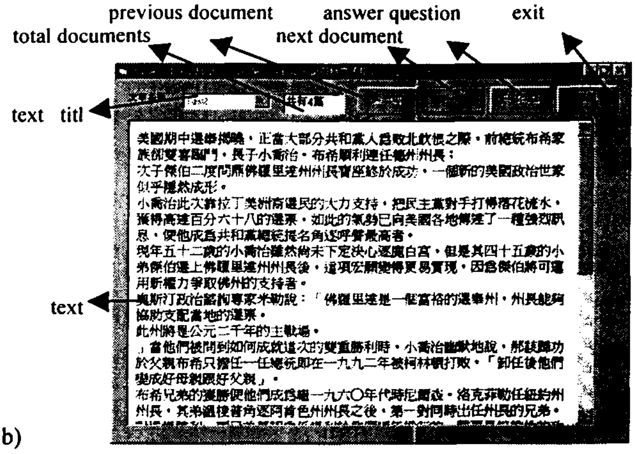
根據氣象局的資料顯示，芭比絲颱風昨日晚間中心位置在澎湖西南西方約二百六十公里海面上，以每小時十三公里速度向北北東轉東北進行，近中心最大風速每秒三十公尺，相當於十一級風，瞬間最大陣風每秒四十公尺，相當於十三級風，暴風半徑二百五十公里，預測今日晚間八時的位置在北緯二十四點四度，東經一百一十九點四度，即在金門東方約一百公里的海面上。

To appear in *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*. November 11-12, 1999.

Appendix B Evaluation Procedure of Assessors

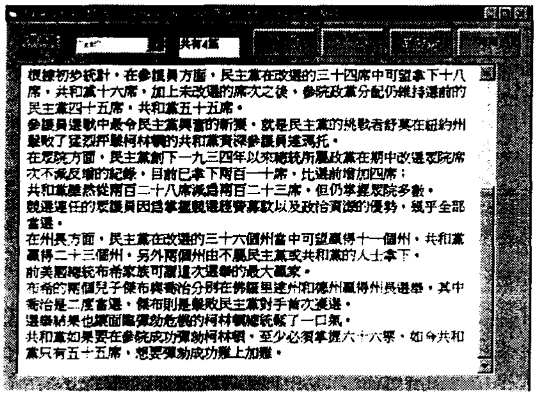
The assessors are divided into two groups. One read the full documents, and the other read summaries only. The following shows an example for event 4 (Bush's sons). We count how much time assessors read the document/summary, and answer questions.

(a)



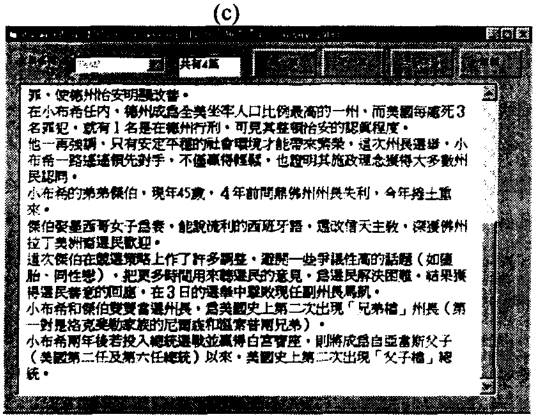
美國期中選舉揭曉，正當大部分共和黨人爲敗北飲恨之際，前總統布希家族卻雙喜臨門，長子小喬治．布希順利連任德州州長；
次子傑伯二度問鼎佛羅里達州州長寶座終於成功，一個新的美國政治世家似乎隱然成形。
小喬治此次靠拉丁美洲裔選民的大力支持，把民主黨對手打得落花流水，獲得高達百分六十八的選票，如此的氣勢已向美國各地傳達了一種強烈訊息，使他成爲共和黨總統提名角逐呼聲最高者。
現年五十二歲的小喬治雖然尚未下定決心逐鹿白宮，但是其四十五歲的小弟傑伯選上佛羅里達州州長後，這項宏願變得更易實現，因爲傑伯將可運用新權力爭取佛州的支持者。
奧斯汀政治諮詢專家米勒說：「佛羅里達是一個富裕的選舉州，州長能夠協助支配當地的選票。
此州將是公元二千年的主戰場。
」當他們被問到如何成就這次的雙重勝利時，小喬治幽默地說，那該歸功於父親布希只擔任一任總統即在一九九二年被柯林頓打敗，「卸任後他們變成好母親跟好父親」。
布希兄弟的獲勝使他們成爲繼一九六○年代時尼爾森．洛克菲勒任紐約州州長，其弟溫梭普角逐阿肯色州州長之後，第一對同時出任州長的兄弟。

(b)



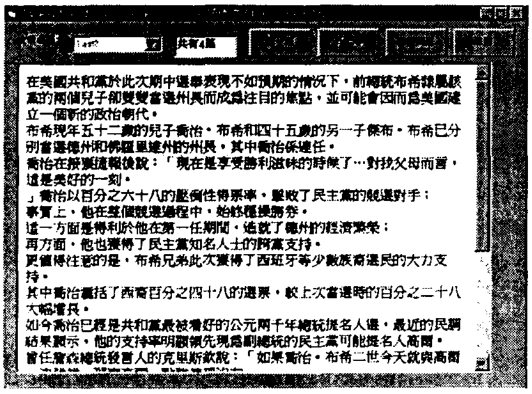
根據初步統計，在參議員方面，民主黨在改選的三十四席中可望拿下十八席，共和黨十六席，加上未改選的席次之後，參院政黨分配仍維持選前的民主黨四十五席，共和黨五十五席。
參議員選戰中最令民主黨興奮的斬獲，就是民主黨的挑戰者舒莫在紐約州擊敗了猛烈抨擊柯林頓的共和黨資深參議員達瑪托。
在眾院方面，民主黨創下一九三四年以來總統所屬政黨在期中改選眾院席次不減反增的紀錄，目前已拿下兩百一十席，比選前增加四席；
共和黨雖然從兩百二十八席減爲兩百二十三席，但仍掌握眾院多數。
競選連任的眾議員因爲掌握競選經費募款以及政治資源的優勢，幾乎全部當選。
在州長方面，民主黨在改選的三十六個州當中可望贏得十一個州，共和黨贏得二十三個州，另外兩個州由不屬民主黨或共和黨的人士拿下。
前美國總統布希家族可謂這次選舉的最大贏家。
布希的兩個兒子傑布與喬治分別在佛羅里達州和德州贏得州長選舉，其中喬治是二度當選，傑布則是擊敗民主黨對手首次獲選。
選舉結果也讓面臨彈劾危機的柯林頓總統鬆了一口氣。
共和黨如果要在參院成功彈劾柯林頓，至少必須掌握六十六票，如今共和黨只有五十五席，想要彈劾成功難上加難。

(c)



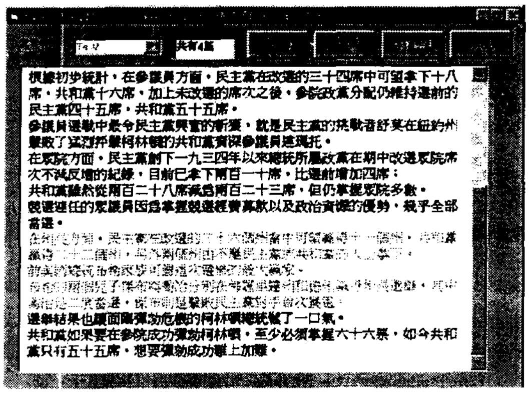
罪，使德州治安明顯改善。
在小布希任內，德州成爲全美坐牢人口比例最高的一州，而美國每處死3名罪犯，就有1名是在德州行刑，可見其整頓治安的認眞程度。
他一再強調，只有安定平穩的社會環境才能帶來繁榮，這次州長選舉，小布希一路遙遙領先對手，不僅贏得輕鬆，也證明其施政理念獲得大多數州民認同。
小布希的弟弟傑伯，現年45歲，4年前問鼎佛州州長失利，今年捲土重來。
傑伯娶墨西哥女子爲妻，能說流利的西班牙語，還改信天主教，深獲佛州拉丁美洲裔選民歡迎。
這次傑伯在競選策略上作了許多調整，避開一些爭議性高的話題（如墮胎、同性戀），把更多時間用來聽選民的意見，爲選民解決困難，結果獲得選民善意的回應，在3日的選舉中擊敗現任副州長馬凱。
小布希和傑伯雙雙當選州長，爲美國史上第二次出現「兄弟檔」州長（第一對是洛克斐勒家族的尼爾森和溫索普兩兄弟）。
小布希兩年後若投入總統選戰並贏得白宮寶座，則將成爲自亞當斯父子（美國第二任及第六任總統）以來，美國史上第二次出現「父子檔」總統。

(d)



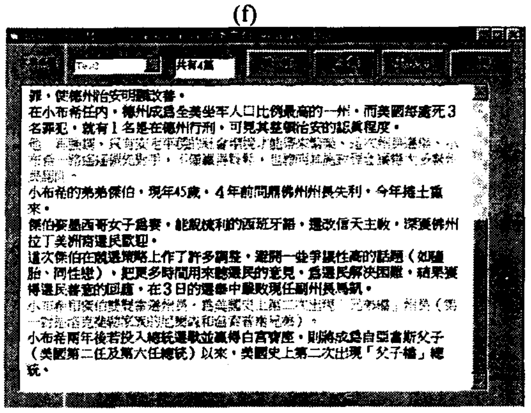
在美國共和黨於此次期中選舉表現不如預期的情況下，前總統布希隸屬該黨的兩個兒子卻雙雙當選州長而成爲注目的焦點，並可能會因而爲美國建立一個新的政治朝代。
布希現年五十二歲的兒子喬治．布希和四十五歲的另一子傑布．布希已分別當選德州和佛羅里達州的州長，其中喬治係連任。
喬治在接獲捷報後說：「現在是享受勝利滋味的時候了…對我父母而言，這是美好的一刻。
」喬治以百分之六十八的壓倒性得票率，擊敗了民主黨的競選對手；
事實上，他在整個競選過程中，始終穩操勝券。
這一方面是得利於他在第一任期間，造就了德州的經濟繁榮；
再方面，他也獲得了民主黨知名人士的跨黨支持。
更值得注意的是，布希兄弟此次獲得了西班牙等少數族裔選民的大力支持。
其中喬治囊括了西裔百分之四十八的選票，較上次當選時的百分之二十八大幅增長。
如今喬治已經是共和黨最被看好的公元兩千年總統提名人選，最近的民調結果顯示，他的支持率明顯領先現爲副總統的民主黨可能提名人高爾。
曾任詹森總統發言人的克里斯欽說：「如果喬治．布希二世今天就與高爾

(e)



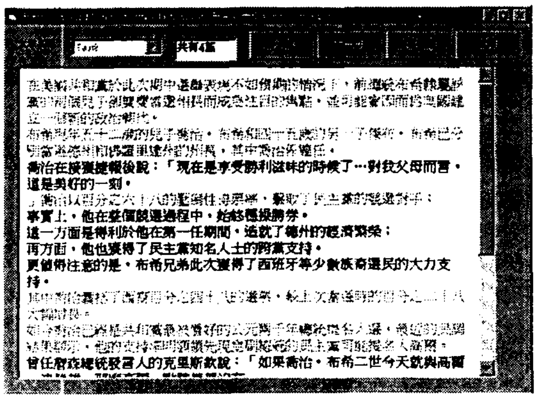
根據初步統計，在參議員方面，民主黨在改選的三十四席中可望拿下十八席，共和黨十六席，加上未改選的席次之後，參院政黨分配仍維持選前的民主黨四十五席，共和黨五十五席。
參議員選戰中最令民主黨興奮的斬獲，就是民主黨的挑戰者舒莫在紐約州擊敗了猛烈抨擊柯林頓的共和黨資深參議員達瑪托。
在眾院方面，民主黨創下一九三四年以來總統所屬政黨在期中改選眾院席次不減反增的紀錄，目前已拿下兩百一十席，比選前增加四席；
共和黨雖然從兩百二十八席減爲兩百二十三席，但仍掌握眾院多數。
競選連任的眾議員因爲掌握競選經費募款以及政治資源的優勢，幾乎全部當選。
在州長方面，民主黨在改選的三十六個州當中可望贏得十一個州，共和黨贏得二十三個州，另外兩個州由不屬民主黨或共和黨的人士拿下。
前美國總統布希家族可謂這次選舉的最大贏家。
布希的兩個兒子傑布與喬治分別在佛羅里達州和德州贏得州長選舉，其中喬治是二度當選，傑布則是擊敗民主黨對手首次獲選。
選舉結果也讓面臨彈劾危機的柯林頓總統鬆了一口氣。
共和黨如果要在參院成功彈劾柯林頓，至少必須掌握六十六票，如今共和黨只有五十五席，想要彈劾成功難上加難。

(f)



罪，使德州治安明顯改善。
在小布希任內，德州成爲全美坐牢人口比例最高的一州，而美國每處死3名罪犯，就有1名是在德州行刑，可見其整頓治安的認眞程度。
他一再強調，只有安定平穩的社會環境才能帶來繁榮，這次州長選舉，小布希一路遙遙領先對手，不僅贏得輕鬆，也證明其施政理念獲得大多數州民認同。
小布希的弟弟傑伯，現年45歲，4年前問鼎佛州州長失利，今年捲土重來。
傑伯娶墨西哥女子爲妻，能說流利的西班牙語，還改信天主教，深獲佛州拉丁美洲裔選民歡迎。
這次傑伯在競選策略上作了許多調整，避開一些爭議性高的話題（如墮胎、同性戀），把更多時間用來聽選民的意見，爲選民解決困難，結果獲得選民善意的回應，在3日的選舉中擊敗現任副州長馬凱。
小布希和傑伯雙雙當選州長，爲美國史上第二次出現「兄弟檔」州長（第一對是洛克斐勒家族的尼爾森和溫索普兩兄弟）。
小布希兩年後若投入總統選戰並贏得白宮寶座，則將成爲自亞當斯父子（美國第二任及第六任總統）以來，美國史上第二次出現「父子檔」總統。

(g)



在美國共和黨於此次期中選舉表現不如預期的情況下，前總統布希隸屬該黨的兩個兒子卻雙雙當選州長而成爲注目的焦點，並可能會因而爲美國建立一個新的政治朝代。
布希現年五十二歲的兒子喬治．布希和四十五歲的另一子傑布．布希已分別當選德州和佛羅里達州的州長，其中喬治係連任。
喬治在接獲捷報後說：「現在是享受勝利滋味的時候了…對我父母而言，這是美好的一刻。
」喬治以百分之六十八的壓倒性得票率，擊敗了民主黨的競選對手；
事實上，他在整個競選過程中，始終穩操勝券。
這一方面是得利於他在第一任期間，造就了德州的經濟繁榮；
再方面，他也獲得了民主黨知名人士的跨黨支持。
更值得注意的是，布希兄弟此次獲得了西班牙等少數族裔選民的大力支持。
其中喬治囊括了西裔百分之四十八的選票，較上次當選時的百分之二十八大幅增長。
如今喬治已經是共和黨最被看好的公元兩千年總統提名人選，最近的民調結果顯示，他的支持率明顯領先現爲副總統的民主黨可能提名人高爾。
曾任詹森總統發言人的克里斯欽說：「如果喬治．布希二世今天就與高爾

evaluation : full docuemnt a -> b -> c -> d , summary a -> e -> f -> g

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

June-Jei Kuo[1], Hung-Chia Wung[1], Chuan-Jie Lin[1] and Hsin-Hsi Chen[2]

Department of Computer Science and Information Engineering

National Taiwan University, Taipei, Taiwan, R.O.C.

[1] { jjkuo, hjwong, cjlin}@nlg2.csie.ntu.edu.tw, [2] 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 both human assessors and a question and answering (QA) system. 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 outperforms 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 an 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 strategy[1] to generate summaries.

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

(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 defered 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

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

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:

(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

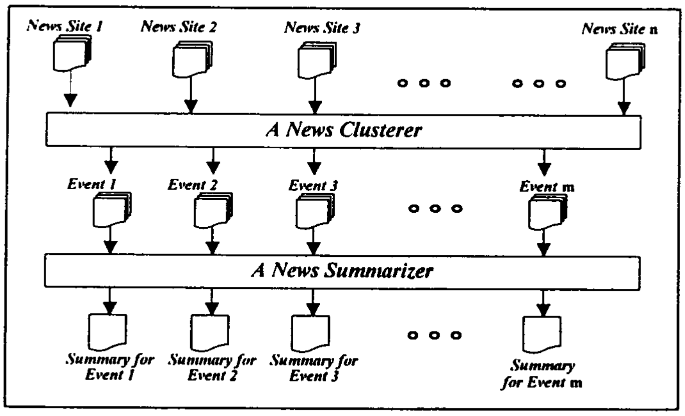


Fig. 1. System Architecture

[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.

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.

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

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

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

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 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, the readers unfortunately stop reading the summaries half way, they would not miss out much important information.

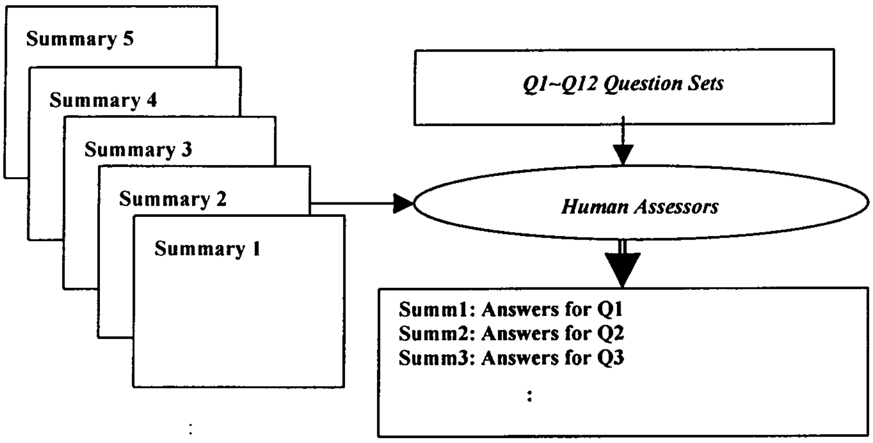


Fig. 2. Example of QA Task

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 questionares (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	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}
B	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7
C	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8
D	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9
E	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}
F	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}
G	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}
H	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7
I	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8
J	Q_4, Q_{10}	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9
K	Q_5, Q_{11}	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}
L	Q_6, Q_{12}	Q_1, Q_7	Q_2, Q_8	Q_3, Q_9	Q_4, Q_{10}	Q_5, Q_{11}

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

	FULL	*BASIC*	*TFWV*	*PSWV*	*TFNV*	*PSNV*
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:

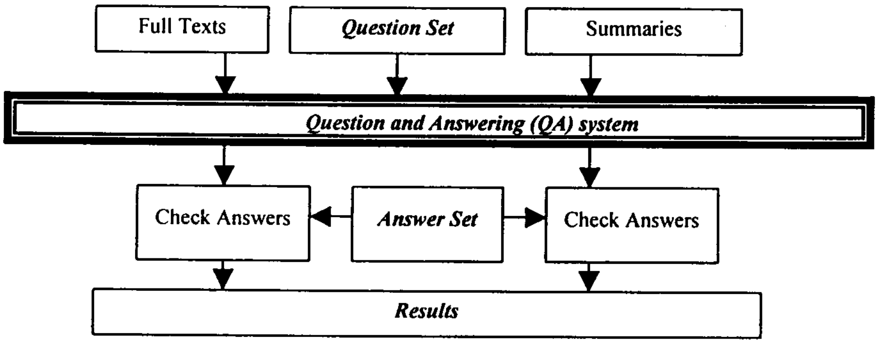


Fig. 3. Revised Evaluation Model

(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.

(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:

$$score(\mathrm{D}) = \sum_{t\ in\ D} \boldsymbol{weight(t)},$$

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):

$$MRR = \sum_{i=1}^{N} r_i / N \ , \qquad (4)$$

where r_i = 1/rank$_i$ if rank$_i$ > 0, or 0 if rank$_i$ = 0. rank$_i$ is the rank of the first correct answer of the ith 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 a 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 questionaires 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 summaried 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 presents 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 frequencey 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 5. Results with Large-Scale Data

	FULL	*BASIC*	*TFWV*	*PSWV*	*TFNV*	*PSNV*
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 Elhada: Using Lexical Chains for Text Summarization. Proceedings of The Intelligent Scalable Text Summarization Workshop, ACL/EACL (1997) 10-17.
18. Tsutomo, 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)

Event Clustering and Visualization in a Chinese-English Multi-Document Summarization System

Abstract

This paper propose a Chinese-English Multi-document summarization system that helps on-line readers absorb news information from multiple sources. The similarity measures of words, sentences and documents to cluster the sentences or documents which denote the same contents or events are proposed respectively. Meanwhile, several strategies, e.g. position free, first-match-first-occupy etc., are described for the similarity computation of multilingual documents. Furthermore, several strategies, e.g. subsumption-based clustering, using complete link for events clustering of multilingual documents are also proposed and the related experimental results are also reported. Finally, in order to tackle the readability issue of multilingual multi-documents summarization, focusing and browsing models considering the readers' preference are proposed.

Introduction

In a basic multi-document summarization system (Mckeown, K., et al.,1999: Chen and Huang, 1999; Goldstein, J., et al., 2000; Hatzivassiloglou, V. Et al., 2001) how to decide which documents deal with the same topic and which sentences (and later, meaningful units) touch on the same event are indispensable. Because a document is composed of some sentences and a sentence consists of several words, how to measure the similarity of words, sentences and documents, respectively is a basic operation for clustering.(Barzilay, R. Et al., 1997; Goldstein, J. Et al, 2000; Radev. D.R., 2000; Mani, I. Et al., 1999) Besides, we have to face the multilinguality problem (Chen and Lin, 2000), i.e., how to know if a Chinese word (sentence, document) and an English word (sentence, document) denote the same concept. Section 2 discusses the issues for similarity measures. Section 3 focuses on the clustering issues. There are three kinds of models for multilingual clustering, i.e., (1) merge the documents from different language sources, do the document and sentence clustering; (2) do the document clustering for each language source, merge the documents clusters denoting the same topic in different languages, and do the sentence clustering; (3) do the document and sentence clustering for each language source, and merge the sentence clusters denoting the same event in different languages. After linking the sentences denoting the same event, Section 4 addresses the visualization issue, e.g., which sentence in which language will be selected, and the preference.

1 Similarity Measure

1.1 Similarity of Words

Word is a basic lexical unit in a sentence. For Chinese, word segmentation is needed to identify the word boundary. For English, a word is converted into the corresponding root form. Besides the word exact matching, we consult thesauri, e.g., Cilin (Mei, et al., 1991) for Chinese and WordNet (Fellbaum, 1998) for English, to find the possible synonym matching. To measure the similarity between Chinese and English words, a bilingual dictionary integrated from four resources, including the LDC dictionary [1], Denisowski's DICT[2], the BDC dictionary v2.2 and a dictionary used in query translation in the MTIR project (Bian and Chen, 2000), is employed. We say two words w_1 and w_2 are similar if w_1 appears in the translation entry of w_2.

2.1 Similarity of Meaningful Units

An English sentence is ended by a full stop. Comparatively, the period (。) in a Chinese

[1] The LDC term list is available at http://morph.ldc.upenn.edu/Projects/Chinese/.

[2] The Denisowski's CEDICT is available at http://ftp.cc.monash.edu.au/pub/nihongo/.

sentence only has the loose function because Chinese writers often assign punctuation marks at random. In most cases, a Chinese sentence ended by a period is composed of sentence fragments separated by comma (，) and other punctuation marks. Several sentence fragments may form a complete meaningful unit even though they are terminated by commas (Yang, 1981). Chen (1994) proposed a method to find the meaningful units from a very long Chinese sentence. For the uniformity, both an English sentence and a Chinese meaningful unit are called an MU later.

Predicate and the surrounding arguments form the basic skeleton in an MU, so that verbs and nouns are considered as basic features for similarity measurement. The similarity of two monolingual MUs is defined as follows:

$$mu\_sim(M_i, M_j) = \frac{|M_i \cap M_j|}{\sqrt{|M_i|}\sqrt{|M_j|}} \quad \cdots\cdots\cdots\cdots (1)$$

where M_i and M_j are two sets denoting two MUs, $M_i \cap M_j$ denotes the common occurrences of two MUs3, and| M_i |, | M_j | and | $M_i \cap M_j$ | denotes the number of elements in the the sets M_i, M_j, and $M_i \cap M_j$, respectively

For computing the similarity of Chinese and English MUs, we have to deal with the ambiguity problem. That is, an English word may correspond to more than one Chinese word in a bilingual dictionary. Five possible strategies are proposed.

(1) position-free

This strategy is similar to the model used in the above method. Let the English MU and the Chinese MU be M_i and M_j, respectively. For each word in M_i, find its Chinese translation. We merge the translation of all the English words and let M_i' be the translation of M_i. The formula is modified as below:

$$mu\_sim(M_i, M_j) = \frac{|M_i' \cap M_j|}{\sqrt{|M_i|}\sqrt{|M_j|}} \quad \cdots\cdots\cdots (2)$$

In this method, an English word may link to more than one Chinese word and a Chinese word may be linked by more than one English word.

(2) first-match-first-occupy

We compare the Chinese translation words of each word in English MU with the words in Chinese MU. As a word in Chinese MU is matched first, the word will be deleted from the Chinese MU and the similarity score (SC) will be added 1. The formula is further modified as shown below.

$$mu\_sim(M_i, M_j) = \frac{SC}{\sqrt{|M_i|}\sqrt{|M_j|}} \quad \cdots\cdots\cdots (3)$$

For example, in figure 1 the C3 has been matched by E1, thus the C3 will be deleted from the candidate words of Chinese MU when processing E2.

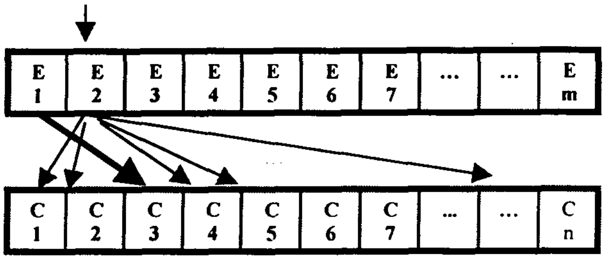


Figure1 Example of first-match-first-occupy strategy

(3) strategy (2) and position relationship within a window

This method is similar with strategy (2), but the Chinese candidates words is limited within a window size. For example, as the Chinese word C2 has been matched by English word E1 shown in Figure 2 and the window size is 3, thus the Chinese candidates words for E2 are C1 and C2.

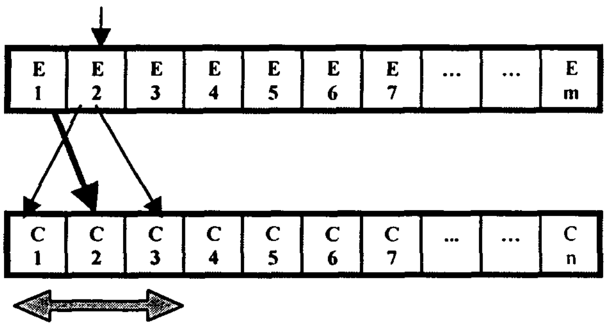


Figure 2 Example of position relationship within a window (=3)

(4) ambigus-word-first and position relationship within a window

[3] The synonym matching specified in Section 11 may be adopted.

This strategy decides those pairs without ambiguity first, thus perform the similar operation described in strategy (3). For example, as the (E2,C5) and (E6,C6) are unambiguous during word matching, they will be decided first shown in Figure 3. Then, the Chinese candidate words for E5 are C3, C4 and C7 when window size is 5.

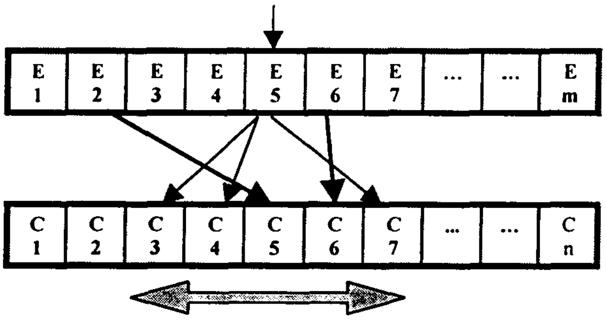


Figure 3 Example of position relationship within a window (=5)

(4) unambiguous-word-first and position relationship within a range

This strategy does not decide the window size beforehand. On the other hand, the Chinese candidate words are decided by the interval between decided pairs. For example, as shown in Figure 4 the (E1,C1) and (E3,C6) are paired unambiguously in advance, thus the Chinese candidates for E2 are C2,C3,C4 and C5.

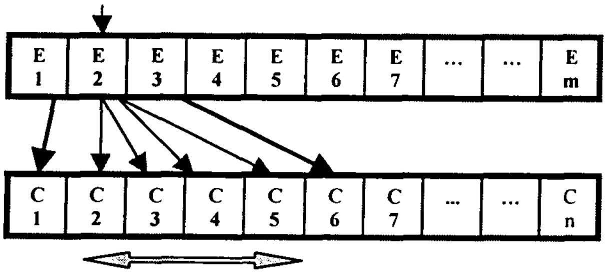


Figure 4 Example of position relationship with a range

2.2.1 Experiment

In order to evaluate the performance of each strategy, we extracted 81 pair news articles (Chinese and English) from the website of United Daily News in Taiwan. On the one hand, we extracted 43 sentence pairs from the above pair news articles to be the test corpus of sentence similarity. Furthermore, we extracted 43 Chinese articles and 43 English news articles randomly from the website respectively and mix them with the above 81 pair news article to be the test corpus of document similarity. In *document similarity* shown as formula (4), we also use the same five strategies described above, but the candidate is no longer words other than a whole MU. The document which has the largest similarity score will be deemed as the corresponding document. The experiment results using precision rate are show in Table 1 and Table 2 respectively.

$$doc\_sim(D_i,D_j)=\frac{|D_i \cap D_j|}{\sqrt{|D_i|}\sqrt{|D_j|}} \quad \text{.................} (4)$$

where D_i and D_j are two sets denoting two documents, $D_i \cap D_j$ denotes the common MUs of two documents, and| D_i |, | D_j | and | $D_i \cap D_j$ | denotes the number of MUs in the sets D_i, D_j, and $D_i \cap D_j$, respectively

Table 1 Result of sentence similarity

	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Best 1	0.883	0.767	0.441	0.255	0.255
Best 2	0.930	0.813	0.674	0.279	0.279
Best 3	0.976	0.860	0.697	0.325	0.325
Best 4	1.000	0.930	0.790	0.372	0.372
Best 5	1.000	0.930	0.790	0.372	0.372

Table 2 Results of document similarity

	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Best 1	0.951	0.839	0.506	0.320	0.320
Best 2	0.987	0.925	0.604	0.432	0.444
Best 3	1.000	0.925	0.666	0.464	0.469
Best 4	1.000	0.950	0.740	0.518	0.518
Best 5	1.000	0.975	0.740	0.530	0.530

2.2.2 Discussion

By observing the above experimental results, we found that strategy 1 and strategy 2 are superior to other there strategies. Thus, we can conclude that position relationship or unambiguous words are not useful cues in deciding sentence or document similarities. Moreover, strategy 1 is also superior to strategy 2. We conclude two reasons as the following:

(1) Due to the difference of word order between Chinese and English, first-match-first-occupy strategy seems not suitable.

(2) The Relationships between lexical item and its translation of the dictionary described in section 1 and above test corpus are as shown in Table 3 and Table 4.respectively. Thus, due to the different feature in our test corpus the first-match-first-occupy has little influence in deciding similarity.

Table 3 Relationship between lexical item and its translation in the Dictionary

Number of lexical item	**Average number of translation word per lexical item**	**Number of lexical item which has one translation**
172,734	2.168658	111,120

Table 4 Relationship between lexical item and its translation in test corpus

Number of lexical item	**Average number of translation word per lexical item**	**Number of lexical item which has one translation**
9,636	10.841012	841

2 Clustering Events

2.1 Clustering Documents

In order to be able to generate the summary from monolingual multiple documents, we use the *complete link* (Salton, 1989) to cluster the documents which describe the same topic. In other word, only all the document similarities, described in section 2.2.1, between any two documents of the two events are larger than the threshold, the two events are clustered into one event.

For computing the document similarities of English and Chinese documents, we proposed two strategies.

(1) One phase multilingual document clustering

As shown in Figure 5, the English and Chinese documents are clustered using the document similarity and complete link directly.

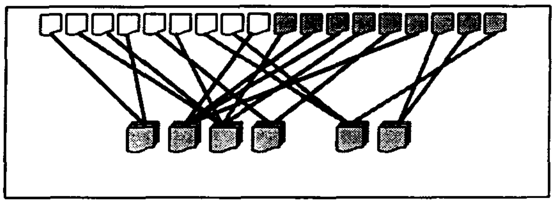


Figure 5 One phase document clustering

(2) Two phases multilingual document clustering

As shown in Figure 6, first the English and Chinese documents will be clustered respectively, then the respective clusters in each language are further clustered using document similarity and complete link.

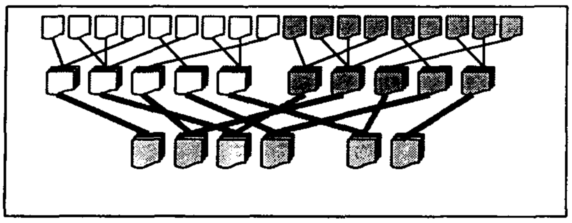


Figure 6 Two phases document clustering

2.1.1 Experiment

We collected the Chinese and English news articles (2001/5/8) from the following news sites in Taiwan.

Mandarin Chinese: Central News agency, Central Daily News, China Times and United Daily News

English: Central News agency, China Post, China Times and United Daily News

First we cluster above Chinese and English documents manually and the results are shown as in Table 5. Furthermore, the experiment results of document clustering using the above two strategies are shown as in Table 6 and Table 7 respectively.

Table 5 Experiment data and manual clustering results

	Number of article	**Number of cluster**
Mandarin Chinese	369	265
English	91	75
Chinese-English	460	318

Table 6 Experimental result using one phase document clustering

Threshold	**Article Number of Clusters (N)**			**Perfect Matching**	**Precision**	**Recall**
	1	5<N<1	N>5			
0.1	156	37	50	154	0.633	0.484
0.2	250	16	36	223	0.738	0.701
0.3	430	0	1	264	0.612	0.830

Table 7 Experiment result using two phases document clustering

	Threshold	**Article Number of Clusters (N)**			**Perfect Matching**	**Precision**	**Recall**
		1	5<N<1	N>5			
C	0.3	240	33	21	253	0.860	0.954
E	0.5	52	13	6	60	0.845	0.800
CE	0.2	281	29	44	296	0.841	0.931

2.1.2 Discussion

By observing the above results, we can conclude the followings.

(1) The performance of strategy 2 outperforms the performance of strategy 1.

(2) The threshold should be determined according to the content similarity of processing documents rather than using the same threshold..

The computation complexity can be reduced largely using strategy 2.

2.2 Clustering Meaningful Units

After the document clustering, those documents in a cluster denote the same event.(Chen and Huang, 1999) Thus, in order to be able to generate the extract summary of an event, we must cluster the similar MUs among documents and then choose a representative MU from each cluster. The two methods using position free strategy described in section 2.2 are proposed as the following.

(1) Complete link using all MUs

We compute MU similarity between any two MUs and then use complete link strategy to cluster the MUs.

(2) Complete-link within a cluster

In order to tackle the computational issue with complete link, we compute the MU similarity between the processing MU and all the MUs in a cluster. If the all the MU similarities are greater than the threshold, the processing MU is added to that cluster. Otherwise, we use another cluster to compute the MU similarity. On the other hand, if there is no cluster that the processing MU can be added, the processing MU will become a new cluster.

(3) Subsumption-based link

In order to further tackle the computational issue, we introduce the concept of information subsumption. We use the following formula (4) to compute the information score of each MU in a cluster and decide the most informative MU which has the highest score to be the ccentroid MU(Radev, et. al, 2000). Meanwhile, we use the inverse document frequency (IDF) of each word in a cluster to select the higher 25 words to be the topic words of that cluster.

$$\text{info\_score}(MU) = (|Mn|+|Mv|+|Tt|) \qquad (4)$$

where MU denotes a meaning unit,

$|Mn|$: the number of nouns in a MU,

$|Mv|$: the number of verb in a Mu,

$|Tt|$: the number of words which is a topic words

Thus, we only compute the MU similarity between the processing MU and the centroid MU rather than all the MUs in that cluster to perform the MU clustering. In order to show the subsumption of two MUs, the MU similarity formula is defined as formula (5). The larger the score is , the more the subsumption is.

$$mu\_sim(M_i,M_j)=\frac{|M_i \cap M_j|}{min(|M_i|,|M_j|)} \quad \text{....}\ (5)$$

where M_i and M_j are two sets denoting two MUs, $M_i \cap M_j$ denotes the common occurrences of two MUs[4], and| M_i |, | M_j | and | $M_i \cap M_j$ | denotes the number of elements in the the sets M_i, M_j, and $M_i \cap M_j$, respectively

2.2.1 Experiment

We collected the one day news articles from the above news sites (2001/5/8). After the manual clustering, we selected five events shown below and the related numbers of English and Chinese articles are shown in Table 8.

(1) Investment for bio-informatics

(2) The relation between president Chen and vice president Lu

(3) Mr. Hsiao Wuan-Chang visits mainland China.

(4) Can the management of Kaoshong harbor return to city government?

(5) The court rejected the application from The Journalist Magazine.

Table 8 Experiment data for MU clustering

	Article Number (C)	**Article Number (E)**	**MU Number (C)**	**MU Number (E)**
Event 1	4	3	69	25
Event 2	5	2	87	39
Event 3	5	3	92	40
Event 4	5	2	82	16
Event 5	2	3	23	46

Furthermore, we also cluster the related MUs manually within the documents of each events and there are 662 correct links. Table 9, Table 10 and Table 11 show the experimental results using the above methods respectively.

Table 9 Experimental results using complete link

Threshold	**Total link Number**	**Number of Correct Link**	**Precision**	**recall**
0.2	852	436	0.511	0.658
0.25	702	408	0.581	0.616
0.3	668	384	0.574	0.580

Table 10 Experimental results using complete link within a cluster

Threshold	**Total link Number**	**Number of Correct Link**	**Precision**	**recall**
0.5	892	892	0.536	0.722
0.55	718	420	0.585	0.634
0.6	622	376	0.604	0.567

Table 11 Experimental results using subsumption-based link

Threshold	**Total link Number**	**Number of Correct Link**	**Precision**	**recall**
0.5	874	462	0.529	0.698
0.55	708	418	0.590	0.631
0.6	602	358	0.595	0.540

2.2.2 Discussion

By observing the best precision of each threshold in the Table 9 and Table 10, we can conclude that the performance of strategy 2 is better than strategy 1. On the one hand, although the performance of strategy 3 is worse than strategy 2, the attraction point of strategy 3 is its time complexity. After analyzing the experimental results, we found that if the score function(formula (4)) can be further improved to obtain more representative MU, the algorithm of strategy 3 is promising.

3 Visualization

In English and Chinese multi-document summarization, how to display the summarization to the readers is an important issue. Two models, i.e., focusing model and browsing model, are proposed. Moreover, the readers' preference for reading the multilingual summarization is also taken into consideration. For example, the Chinese readers prefer to read more Chinese summarization than English. On

[4] The synonym matching specified in Section 1.1 may be adopted.

the contrary, the English readers would prefer to read more English summarization than Chinese.

3.1 Focus Model

A summarization is presented by voting from reporters. For each event, reporter records a news story from his own viewpoint. Recall that a news article is composed of several MUs. Those MUs that similar in a specific event are common focuses of different reporters. In other words, they are worth reading. In the current implementation, the MUs are reported than twice are our target. For readability, the original sentences that cover the MUs are extracted. For each set of similar MUs, only the longest sentence is displayed. The display order of the extracted sentences is determined by the related position in the original news articles shown as Figure 7 and the related position score function is defined as formula (6).

$$Posi\_score(sentence) = \frac{position(sentence, Doc)}{sizeof(Doc)} \quad .(6)$$

where position(sentence, Doc) : the sentence starting position in Doc
sizeof(Doc): the size of document Doc.

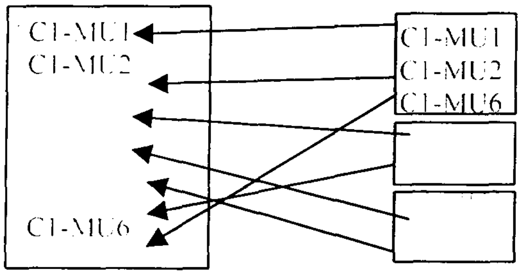


Figure 7 Display order of extracted sentences

Furthermore, the display order of English and Chinese multi-document summarization considering the readers preference is shown as Figure 8.

3.2 Browsing Model

The news articles are listed by information decay. The first article is shown to the user in its whole content. In the latter shown news articles, the MUs denoting the information mentioned before are shadowed (or eliminated), so that the reader can focus on the new information. The amount of information in a news article is measured in terms of the number of MUs, so that the article that contains more MUs is displayed before the others. For readability, a sentence is a display unit. On the other hand, we also consider the readers' preference in English and Chinese multi-document summarization. Thus, the news articles of the preference language are shown before the news articles of the other language is shown.

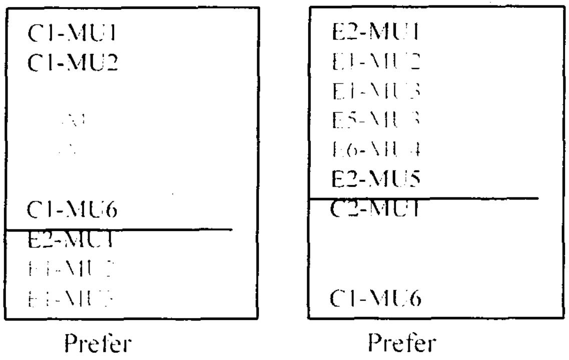


Figure 8 Display order considering the readers' preference

4 Concluding Remarks

This paper presents a Chinese-English Multi-document summarization system. It summarizes clusters of news articles automatically grouped by event clustering system. We used five strategies for the similarity computation between words and sentences of Chinese and English documents and found that the position free strategy is the most promising. On the one hand, we also used two strategies for multilingual document clustering and found that the performance of two phase multilingual document clustering is better than one phase multilingual document clustering. Furthermore, three strategies using complete link to tackle the MUs clustering of multilingual documents of an event and found that the complete link within a cluster has the best precision, but the subsumption-based clustering has the advantage of lower computation complexity and similar precision with the former strategy. Finally, we also presented two promising visualization models, e.g. focusing and browsing models, considering the readers' preference. In the future, we would like to expand our Chinese and English multi-document summarization system to multilingual multi-document summarization system and further test the readability of the proposed two visualization models.

Acknowledgements

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

References

Barzilay, R., et al., (1997) "*Using Lexical Chains for Text Summarization*," Proceedings of ACL/EACL 1997 Workshop on The Intelligent Scalable Text Summarization,pp10-16

Bian, G.W. and Chen, H.H (2000). "*Cross Language Information Access to Multilingual Collections on the Internet.*" Journal of American Society for Information Science, Special Issue on Digital Libraries, 51(3), 2000,pp 281-296

Chen, H.H. (1994) "*The Contextual analysis of Chinese Sentences with Punctuation Marks*," Literal and Linguistics computing, Oxford University Press, 9(4), 1994, pp.281-289

Chen, H.H. and Huang, S.J.(1999) "*A Summarization System for Chinese News from Multiple Sources,*" Proceeding of 4th International Workshop on Information Retrieval with Asia Language (1999) 1-7

Chen, H.H. and Lin, J.J. (2000) "*A Multilingual News Summarizer*", Proceedings of 18th International Conference on Computational Linguistics, pp159-165

Fellbaum, C. (1998) : *WordNet*. The MIT Press, Cambridge Masschusettes, 1998

Goldstein, J., et al. (2000) "*Multi-docuemt summarization by sentence extraction*," Proceedings of NAACL2001 Workshop on Automation Summarization, Seattle, WA

Goldstein, J., et al. (2000) "*Creating and Evaluating Multi-Document Sentence Extract Summaries*," Proceedings of the 2000 ACMCIKM International Conference on Information and Knowledge Management, pp165-172

Hatzivassiloglou, et al. (2001) "*SIMFINDER: A Flexible Clustering Tool for Summarization*," Proceedings of NAACL2001 Workshop on Automation Summarization, pp41-49

Mani, I. Et al. (1999) "*Summarizing Similarirties and Difference Among Related Documents*," Information Retrieval 1(1-2), pp35-67

Mckeown, K., et al. (1999), "*Towards Multidocument Summarization by Reformulation,* ," Proceedings of AAAI-99, pp453-460

Mei, et al. (1982) *tong2yi4ci2ci2lin2*, shanghai Dictionary Press, 1982

Radev, D.R., et al. (2000) "*Subsumption-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies.*" Workshop on Summarization, ANLP/NAACL, 2000

Salton, G. (1989) *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, 1989

Yang, Y. (1981) *The Research on Punctuation Marks*, Tian-jian Publishing Company, Hong Kong, 1981