Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Using LSA and text segmentation to improve automatic Chinese dialogue text summarization*

LIU Chuan-han†[1], WANG Yong-cheng[1], ZHENG Fei[2], LIU De-rong[1]

([1]*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200030, China*)

([2]*Center for Biomimetic Sensing and Control Research, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China*)

†E-mail: uuchliu@163.com

Received July 4, 2006; revision accepted Oct. 7, 2006

Abstract: Automatic Chinese text summarization for dialogue style is a relatively new research area. In this paper, Latent Semantic Analysis (LSA) is first used to extract semantic knowledge from a given document, all question paragraphs are identified, an automatic text segmentation approach analogous to TextTiling is exploited to improve the precision of correlating question paragraphs and answer paragraphs, and finally some "important" sentences are extracted from the generic content and the question-answer pairs to generate a complete summary. Experimental results showed that our approach is highly efficient and improves significantly the coherence of the summary while not compromising informativeness.

Key words: Automatic text summarization, Latent semantic analysis (LSA), Text segmentation, Dialogue style, Coherence, Question-answer pairs

doi:10.1631/jzus.2007.A0079 **Document code:** A **CLC number:** TP391.1

INTRODUCTION

Text summarization is an increasingly pressing practical problem due to the explosion of the amount of textual information available. Informally, the goal of text summarization is to take a textual document, extract content from it and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs (Mani, 2001). Instead of having to go through an entire text, one can understand a document quickly and easily by means of its concise summary. It is said that a professional abstractor can edit only 55 summaries at most in one day (15 summaries at least and 27 summaries on average) (Cunningham and Wicks, 1992). This makes it necessary to use computer for making summary automatically, so that the speed of summary publication can keep up with that of document publication.

Automatic text summarization is an extremely active research field making connections with many other research areas, such as Information Retrieval, Natural Language Processing, and Machine Learning. It can be classified as extracting-based summarization and understanding-based summarization (Wu *et al.*, 1998). In the research areas of Artificial Intelligence and Natural Language Understanding, many problems are hard to be solved by far so that the text summarization approaches based on understanding go forward slowly. The text summarization approaches based on extracting are mainly dependent on the normality of discourse structure of the document and do not semantically analyze the sentences or paragraphs of the document so that they have many obvious shortcomings, especially, some topics are missed or the content of the summary is not coherent when the document includes multiple topics.

Dialogue document belongs to a special multi-topics document, where there are two or more than

* Project (No. 2002AA119050) supported by the National Hi-Tech Research and Development Program (863) of China

two dialoguers. One of the participants asks some questions, others answer or refuse to answer these questions, or one of them affirms or reviews some opinions of other dialoguers, etc. If the conventional automatic text summarization approaches for the non-dialogue style are applied to the dialogue style's document, the dialogue between two participants may be not interrelated in the summary, so that the coherence, readability and logic validity of the summary are reduced significantly.

Along with the vast appearance of dialogues databases and dialogue documents in the Internet, such as portraits visit, news conference, text living of video programming in the Internet, one urgently needs a high-quality automatic summarization system for dialogue documents, which especially helps in indexing, classification, and retrieval of various dialogue documents, such as multi-meeting, exchange in business, negotiating with client.

In this paper, a new automatic summarization approach for Chinese dialogue documents is presented. It is based on extracting "important" sentences and text segmentation, where Latent Semantic Analysis (LSA) (Deerwester *et al.*, 1990) is used to extract semantic knowledge from a given document.

The rest of this paper is organized as follows. Section 2 simply describes the work related to this paper. Section 3 presents in detail some critical techniques for automatic summarization of dialogue documents. Section 3 has five subsections: Subsection 3.1 gives the system overview of automatic summarization for dialogue documents, Subsection 3.2 introduces the method of identifying the style of document, Subsection 3.3 introduces the use of LSA, Subsection 3.4 presents the method of identifying question paragraphs and investigates how to correlate each answer paragraph with its corresponding question paragraph, and Subsection 3.5 uses the correlation information of the question-answer pairs to generate the summary so that the problems using the conventional automatic text summarization method to summary a dialogue document can be avoided as much as possible, namely, the local coherence and readability of the summary are greatly improved. Section 4 gives the experimental results of automatic summarization for dialogue documents. Experimental results showed that, our approach has the high accuracy of identifying the style for dialogue documents and the high precision of correlating between question paragraphs and answer paragraphs so that it significantly improves the coherence of the summary while not compromising informativeness of the summary for dialogue documents. Finally, Section 5 summarizes the work and describes the future research directions.

RELATED WORK

Klaus Zechner is the first who studied automatic generation of concise summaries for dialogue documents (Zechner and Lavie, 2001; Zechner, 2001; 2002). The texts he processed are some human transcripts of spoken dialogues so that there was the issue of speech recognition errors and sentence boundaries were typically not available in the first place. He discussed and tried to address some challenges works on spoken dialogue summarization, such as coping with speech disfluencies, identifying the units for extraction, maintaining cross-speaker coherence, and coping with speech recognition errors. Now that the questions in the processed spoken dialogues are the individual questions, he trained a decision tree classifier (C4.5) using a corpus annotated manually and exploited this classifier to detect which sentences are questions. He had found from statistics that for more than 75% of the yes-no-questions and Wh-question, the answer was to be found in the first sentence of the speaker following the speaker uttering the question and in the remainder of cases the majority of answers were in the second (instead of the first) sentence of the other speaker. Hence, he devised a heuristic search to detect answers using some features, such as matching words between questions and answers.

Recently, some related studies were presented (Chen *et al.*, 2005; Hsueh *et al.*, 2006; Wu *et al.*, 2005; Zhang and Soergel, 2006). In (Chen *et al.*, 2005), the sentence similarity model was exploited to calculate the similarities between questions and answers. Wu *et al.*(2005) proposed an approach to domain-specific FAQ retrieval using independent aspects, where the idea of topic classification using paragraph-based LSA of question-answer pairs is presented. In (Hsueh *et al.*, 2006), the problem of automatically predicting segment boundaries in spoken multiparty dialogue was investigated. Zhang and Soergel (2006) dis-

cussed some knowledge-based approaches to the segmentation of oral history interviews. They applied the knowledge on discourse structure and questions as an indicator of topicality to the segmentation of speech transcripts, and suggested a segmentation approach combing multiple sources of evidence.

LSA is an approach to automatic indexing and information retrieval that attempts to overcome these problems by mapping documents as well as terms to a representation in the so-called latent semantic space. It usually uses high dimensional vector space representation of documents based on term frequencies as a starting point and applies a dimension reducing linear projection. The specific form of this mapping is determined by a given document collection and is based on Singular Value Decomposition (SVD) (Golub and van Loan, 1996) of the corresponding term-document matrix. The general claim is that similarities between documents or between documents and queries can be more reliably estimated in the reduced latent space representation than in the original representation. The foundational principle is that documents sharing frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. LSA thus performs some sort of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic. In many applications this has proven to result in more robust word processing. Choi *et al.*(2001) used LSA to estimate inter-sentence similarity matrix, where the "meaning" of a sentence was represented by the sum of the LSA feature vectors. In their experiments, semantic knowledge was acquired from a corpus containing the texts to be segmented in the test phase. Bestgen (2006) reanalyzed Choi *et al*.(2001)'s algorithm and reported two experiments. His experiments showed that the presence of the test materials in the LSA corpus has an important effect and that the generic semantic knowledge derived from large corpora clearly improves the segmentation accuracy.

In the last ten years, many methods for text segmentation have been proposed (Beeferman *et al.*, 1999; Bestgen, 2006; Hearst, 1997; Hsueh *et al.*, 2006; Kaufmann, 1999; Kehagias *et al.*, 2003; Kozima, 1993; Li and Yamanishi, 2003; Ponte and Croft, 1997; Reynar, 1999; Salton *et al.*, 1996; Zhang and Soergel, 2006). In (Hearst, 1997), TextTiling created for each segmentation candidate two pseudo-blocks, one preceding it and the other following it, and calculated the cosine value of the two pseudo-blocks' word frequency vectors as the similarity. It then conducted the segmentation at valley points whose similarity values are lower to a pre-determined value than each of the values of its left "peak" and right "peak". In (Li and Yamanishi, 2003), a Stochastic Topic Model (STM) was employed to represent a word distribution within a text. Two pseudo-blocks near a candidate point of segmentation were obtained in a way similar to that of the method mentioned in (Hearst, 1997). The significant differences were calculated as the similarity between STMs of two pseudo-blocks. In (Wang *et al*., 2005), a simple method of text segmentation was described as follows. Suppose the given document has n paragraphs. r denotes the relevance degree of the two adjacent paragraphs. r' denotes the mean of all r's. If there exists r^* which satisfies the following conditions: (1) r^* is lower to a pre-determined value ξ than its left and right, (2) $r^* \leq r'$, then r^* is the segmentation. Additionally, Kaufmann (1999) considered collocational word similarity as a source of text cohesion that is hard to measure and quantify and evaluated this method in the text segmentation task. His experimental results showed that adding collocational information from the training corpus improves the prediction of section breaks. Kehagias *et al.*(2003) proposed a segmentation algorithm based mainly on dynamic programming that equals or even outperforms the results of (Choi *et al*., 2001). This algorithm does not depend on additional semantic knowledge. According to (Bestgen, 2006), this algorithm could still be improved by taking into account such knowledge.

SUMMARIZATION SYSTEM

Overview of system

Automatic summarization system for dialogue documents, as shown in Fig.1, consists of (1) document preprocessing, (2) style identification, (3) identification of the correlation between question paragraphs and answer paragraphs, and (4) summary generation.

In the document preprocessing stage, the system transforms those documents with different file formats, such as TXT, HTML, DOC, into the salient feature documents with unified normal format.

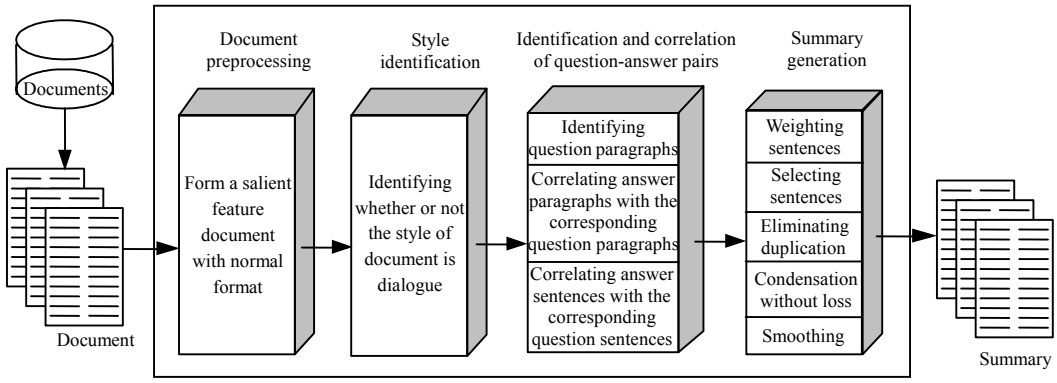

Fig.1 Architecture of automatic summarization system for dialogue documents

In the style identification stage, the system judges whether or not a given document belongs to dialogue documents. Generally speaking, the documents' styles are diversified and human summarizers apply the different text summarization methods to those documents with different styles. Hence it helps to improve the quality of the summary using the corresponding automatic summarization method for a given document with a certain style.

In the stage of identification and correlation of question-answer pairs, the ultimate goal is to identify the correlation between each answer sentence and its corresponding question sentence. Namely, after correctly identifying the style of a document, the system firstly identifies whether or not each sentence is a question sentence in the given document, secondly correlates all of the answer sentences with their corresponding question sentences.

In the summary generation stage, the heuristic rule will be used to calculate the scores of all sentences and all question-answer pairs in the given document, and then the "important" sentences are selected from the general content and question-answer pairs and composed logically to generate the summary.

Identification of document style

In the processing of automatic summarization of dialogue documents, it is firstly identified whether or not the style of a given document is dialogue. The identifying result will determine that the next processing exploits the method based on the conventional extracting for non-dialogue style or exploits the particular method based on extracting proposed in this paper.

After some samples with dialogue style were analyzed manually, it was found that dialogue documents have some specific features that can be used to judge whether or not the style of a document is dialogue. After 200 samples with dialogue style had been randomly selected from the corpus of dialogues, which are collected from on-line news sites, such as people.com.cn, xinhuanet.com, sina.com, and analyzed manually, it was found that the tagged features of all participants occur in the first sentence of paragraphs and end with colon or other symbols in dialogue documents. The statistical data of dialogue features in Chinese dialogue documents is shown in Table 1 indicating that 99.5% of dialogue documents have the salient tagged features for participants. Counting the number of sentences that occur in the first sentence of each paragraph and end with feature symbols, such as "：", ":", "】", "]", and whose length was restricted by a given length, it could be judged whether or not the given document was dialogue type.

Table 1 Statistical data of dialogue features in Chinese dialogue documents

End symbol	Examples	Account
： or :	问：最难过的是什么时候？ 【问：】你怎么看待新闻自由？你自由吗？	195
】or]	[网友六神无主] 徐老师，您觉得新东方的成功是否意味着传统正规大学的教育的失败呢？	4
Other	今天上午布什总统谈及台湾问题时是否使用了"一个中国政策"这个字眼？	1

Latent semantic analysis

For each document, LSA (Deerwester *et al.*, 1990) is trained on the set of sentences $S=\{s_1, \ldots, s_m\}$ with the set of terms $\{t_1, \ldots, t_n\}$ in the document. An $n\times m$ matrix $\boldsymbol{A}=(a_{ij})$ is calculated, where a_{ij} denotes the number of times t_i occurs in s_j, without loss of generality $m\geq n$, and rank($\boldsymbol{A}$)=r.

Singular value decomposition (Golub and van Loan, 1996) is then applied to generate

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}_{n\times n}$, $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{I}_{m\times m}$, and $\boldsymbol{\Sigma}$=diag($\sigma_1, \ldots, \sigma_n$), $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_r>\sigma_{r+1}=\ldots=\sigma_n=0$. The first r columns of the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ define the orthonormal eigenvectors associated with the r nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, respectively. The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are referred to as the left and right singular vectors, respectively, and the singular values of $\boldsymbol{A}$ are defined as the diagonal elements of $\boldsymbol{\Sigma}$ which are the nonnegative square roots of the n eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.

If the largest K singular values in $\boldsymbol{\Sigma}$ are kept and the remaining smaller ones are set to zero, the product of the resulting matrices is a matrix $\tilde{\boldsymbol{A}}$ which is only approximately equal to $\boldsymbol{A}$, and is of rank K. Since zeros were introduced into $\boldsymbol{\Sigma}$, the representation can be simplified by deleting the zero rows and columns of $\boldsymbol{\Sigma}$ to obtain a new diagonal matrix $\tilde{\boldsymbol{\Sigma}}$, and then deleting the corresponding columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ to obtain $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ respectively. So

$$\boldsymbol{A}\approx\tilde{\boldsymbol{A}}=\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^{\mathrm{T}}. \tag{2}$$

K can be selected so that it satisfies

$$\sum_{i=1}^{K}\sigma_i\Big/\sum_{i=1}^{r}\sigma_i\geq\eta, \tag{3}$$

where $\eta\in[0, 1]$ is a pre-determined value.

In latent semantic subspace a latent semantic vector of sentence s_j is represented as

$$\boldsymbol{w}_j=\boldsymbol{A}_j^{\mathrm{T}}\tilde{\boldsymbol{U}}, \tag{4}$$

where $\boldsymbol{A}_j$, the jth column of $\boldsymbol{A}$, denotes the term frequency vector of sentence s_j. In the stage of identifying question-answer pairs, the latent semantic vectors of sentences replacing the frequency vectors of sentences will be used to calculate the similarity of two pseudo-blocks.

Identification of question-answer pairs

In the question-answer pairs' identification stage, all of the dialogue content in the given document will be segmented into many different information units so that each unit embodies only one dialogue content between one questioner and its corresponding answerer. The identifying correctness and the question-answer pairs' correlation will directly influence the quality of the summary, that is to say, the coherence, readability and logical validity of the summary.

Firstly, the question paragraphs spoken by questioner are identified from all paragraphs in the given document. Secondly, the corresponding answer paragraphs are identified for each question paragraph. Finally, each answer sentence in answer paragraphs is correlated with its corresponding question sentence in the question paragraph. The purpose of identification and correlation of question paragraph and answer paragraph is to improve the correlation precision between the question sentences and their answer sentences, and reduce the computational complexity of the correlation between one sentence and the other sentences.

Each question paragraph possibly includes several question sentences, and there are several corresponding answer sentences in their corresponding answer paragraphs, how to correlate each answer sentence in the answer paragraphs with its corresponding question sentence must be considered. Manual analysis of 200 dialogue samples mentioned in the above subsection showed that not one of the question sentences have one or more than one corresponding answer sentences. In some dialogue samples, certain question paragraphs embody more than one question sentence, but some question sentences among them did not have the corresponding answer sentences in the corresponding answer paragraph or several question sentences among them have some common answer sentences. Moreover, two question sentences in certain question paragraphs have the same or close meaning, with the latter being the complement or detailed explanation of the former. Although the correlation between questions and answers was tried in (Chen *et al.*, 2005), it is difficult to

correlate each answer sentence with its corresponding question sentence in some cases. Consequently, the correlation between question paragraphs and their corresponding answer paragraphs using text segmentation technique will be emphasized in this study.

1. Identification of question paragraphs

The participant information can be denoted as a triplet $\{T, O, S\}$, where T denotes the tagged feature symbol of participant, O denotes the occurrence number of the identity string of participant in the given document, S denotes the status of participant, whose values indicate that the participant is a questioner or an answerer.

After 200 dialogue samples mentioned in subsection 3.2 had been analyzed, more than 100 feature words and symbols had been collected, such as "?", "谁 (who)", "为什么 (why)", "何时 (when)", "多少 (how many/much)", each of which indicates that the sentences including it are likely question sentences, so a dictionary was constructed to store these question feature words and symbols.

In the stage of identifying question paragraph the main task is to identify all question paragraphs in the given document. The operating steps are as follows:

(1) Counting the information of all participants. Now that the information of all participants occurs in the first sentence of some paragraphs, the identity string of all participants can be recorded orderly to T and their occurrence number in the document can be counted to O.

(2) Judging the status of each participant. After the information of all participants is counted, the statistical method will be used to judge that the status of each participant is a questioner or an answerer. Using the question features dictionary, it can be judged whether or not the current analyzed paragraph is a question paragraph according to the occurrence number of these question feature words in the current paragraph and the length of the given document. Finally the number of question paragraphs spoken by each participant is counted. If this number is more than half of all the occurrence number of this participant, it can be judged that this participant is a questioner, or an answerer.

(3) Identifying all question paragraphs. According to the above status information of all participants, all the paragraphs whose status is questioner are marked as question paragraphs.

2. Correlating question paragraphs and answer paragraphs

According to the statistical analysis of (Zechner and Lavie, 2001; Zechner, 2001; 2002), in dialogue document each question paragraph has one or more than one corresponding answer paragraphs, and the answer paragraphs usually locate one by one after the corresponding question paragraph, rarely are two sequential question paragraphs followed by common answer paragraphs, namely, a question paragraph is seldom not followed by any answer paragraphs. After 200 dialogue samples were analyzed, it can be known that this judgment presented above is still true for Chinese dialogue documents. For this characteristic, all paragraphs between two question paragraphs are regarded as the answer paragraphs of the previous question paragraph, all paragraphs following the last question paragraph as its answer paragraphs. Using this simple rule, the corresponding answer paragraphs of each question paragraph were identified in (Chen *et al*., 2005).

In fact, if there is only one paragraph between two question paragraphs, it must be the answer paragraph of the previous question paragraph; if there are more than one paragraph between two question paragraphs, some of their paragraphs are possibly not the answer paragraphs of the previous question paragraph but some generic content written by the author for the purpose of playing a connecting link between the preceding and the following paragraphs. Apparently, the boundary between answer paragraphs and generic content is clear. Some paragraphs preceding it are the answer paragraphs of the previous question paragraph while some following it are the generic content. So text segmentation will be used to separate the candidate answer paragraphs of a question paragraph into two parts: answer paragraphs and generic paragraphs, which are not related with its previous question paragraph.

Here, suppose that a question paragraph QP is followed by l candidate answer paragraphs AP_1, AP_2, …, AP_l ($l\geq1$). QP, AP_1, AP_2, …, AP_l are merged into a pseudo-text PT. If $l=1$, such pseudo-text need not be segmented. If $l\geq2$, an analogous method of TextTitling (Hearst, 1997) is exploited to segment the pseudo-text PT into two topic groups: answer paragraphs and generic paragraphs.

All sentence-ending periods are first set as the

candidate points of segmentation within *PT*. For each candidate *i*, two pseudo-blocks B_1 and B_2 are created, one consisting of the *k* sentences preceding it, and the other of the *k* sentences following it (when fewer than *k* exist in any direction, those which do exist are simply used). Next, the similarity between the preceding pseudo-block and the following pseudo-block is calculated by a cosine measure

$$sim(i)=sim(B_1,B_2)=\frac{\boldsymbol{w}_{B_1}\boldsymbol{w}_{B_2}^{\mathrm{T}}}{\|\boldsymbol{w}_{B_1}\|\cdot\|\boldsymbol{w}_{B_2}\|}=\sum_{i=1}^{K}w_{iB_1}w_{iB_2}\Bigg/\left(\sqrt{\sum_{i=1}^{K}w_{iB_1}^2}\sqrt{\sum_{i=1}^{K}w_{iB_2}^2}\right), \tag{5}$$

where *K* is the dimension of latent semantic subspace, and

$$\boldsymbol{w}_{B_1}=\sum_{j=1}^{k}\boldsymbol{w}_{i-j+1},\quad \boldsymbol{w}_{B_2}=\sum_{j=1}^{k}\boldsymbol{w}_{i+j}, \tag{6}$$

$\boldsymbol{w}_j$ denotes a latent semantic vector of sentence s_j in latent semantic subspace (see Eq.(4)), and w_{iB_1}, w_{iB_2} are the *i*th element of $\boldsymbol{w}_{B_1}$ and $\boldsymbol{w}_{B_2}$, respectively.

The set of candidate points is denoted as *CP*, which should only include points whose left and right are candidate answer paragraphs. A point cp^* whose similarity score is the minimum score is searched in *CP*. If its similarity score is lower to a pre-determined value θ than each of the scores of its left peak and right peak, it is a reasonable segmentation, otherwise, the segmentation does not exist, namely, all of the candidate answer paragraphs are the answer paragraphs. The segmentation algorithm of a pseudo-text is shown in Fig.2.

Fig.3 shows a graph of calculated similarity scores for each of the candidates in certain text where k=3, θ=0.05. Points 5, 8, 12, and 17 in Fig.3 form the set of candidate points *CP*. Point cp^*=17 is the candidate point of segmentation whose similarity score is the minimum score in *CP*. The similarity score of its left peak is 0.39 and that of its right peak is 0.29. Because 0.39−0.12 and 0.29−0.12 are greater than θ, the segmentation is performed at candidate point 17.

Summary generation

In this subsection, it is supposed that the length of the given document is *L*, the compression ratio of

```
n: the number of sentences in pseudo-text
l: the number of candidate answer paragraphs in pseudo-text
k: the number of sentences in pseudo-block
θ: a pre-determined value
S(i): the status of the ith sentence, 1 indicates it is the last
      sentence of a paragraph, otherwise not
if (l==1) return 0;  // there is no segmentation point
// calculate the similarity scores between two pseudo-blocks
// near the ith sentence
for (i=1; i<n; i++) sim(i);  // using Eq.(5)
// for the convenience of calculating
sim(0)=sim(1); sim(n)=sim(n−1);
// finding the sentence whose state is 1 and similarity score
// is minimum
cp=1;
for (i=2; i<n; i++)
   if (S(i)==1 && sim(i)<sim(cp)) cp=i;
// judging whether or not this point is segmentation point
if (sim(cp−1)>=sim(cp) && sim(cp+1)>=sim(cp)) {
   j=cp−1;
   while (j>0 && sim(j−1)>=sim(j)) j−−;
   P1=sim(j);
   j=cp+1;
   while (j<n && sim(j+1)>=sim(j)) j++;
   P2=sim(j);
   if (P1−sim(cp)>θ && P2−sim(cp)>θ)
      return cp;
   else
      return 0;
}
else
   return 0;
```

Fig.2 Segmentation algorithm of a pseudo-text

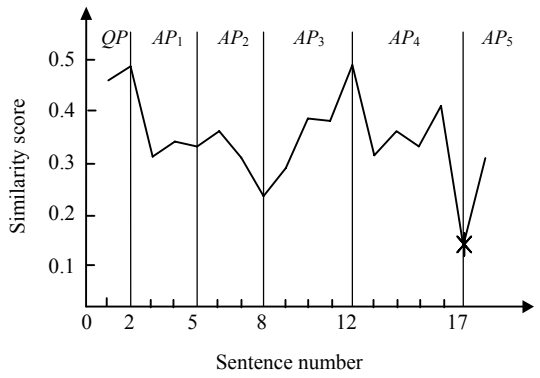

Fig.3 Similarity scores for segmentation candidates when the number of sentences in pseudo-block is 3 and the pre-determined value is 0.05

text summarization is γ, the generic content and all of the question-answer pairs form a set of information units denoted as $\{U_1, U_2, \ldots, U_N\}$.

Firstly, the score *score*(*s*) of each sentence *s* is calculated in the given document according to cue phrases, title and sentence position, and the score *score*(U_k) of U_k by

$$score(U_k)=\sum_{i=1}^{N_k} score(s_i), \quad (7)$$

where N_k denotes the number of sentences in U_k, s_i is a sentence with U_k.

Secondly, the information units $\{U_1, U_2, \ldots, U_N\}$ are ranked to $\{U'_1, U'_2, \ldots, U'_N\}$ in descending order of their scores, and the candidate sentences of the summary are selected from U'_1 to U'_N in turn. In each information unit U'_k, the sentences are selected according to their scores in descending order and the total length $Length(U'_k)$ of candidate sentences must satisfy

$$\left|Length(U'_k)-L_k\right|\le\delta, \quad k=1, 2, \ldots, N, \quad (8)$$

where

$$L_k=score(U'_k)L\gamma\Big/\sum_{i=1}^{N} score(U'_i).$$

and δ denotes the allowed error scope of length. When the processing information unit is a question-answer pair, if its answer sentence is selected into the summary, then its question paragraph must be selected into the summary unless Eq.(8) is not satisfied, that is to say, if its question paragraph cannot be selected into the summary due to Eq.(8) not being satisfied, then its corresponding answer sentence cannot be selected.

Finally, the following operation will be done: deleting duplicate sentences, refining sentences, compressing sentences with no information loss, smoothing the sentences in the summary, and sorting the sentences to form the summary in the order in which they appeared in the original document.

EXPERIMENTAL RESULTS

1175 dialogue samples and 1341 no-dialogue samples, all of which are web files, were collected specially as the corpus from some online news sites. The precision of identifying style and identifying question-answer pairs will be tested as follows.

(1) Style identification

In our experiment, the error ratio of style judgment for 1341 samples with no-dialogue style is zero and that for 1175 samples with dialogue style is 0.43%, that is to say, 5 dialogue samples are erroneously identified as no-dialogue documents. After these misidentified dialogue samples are analyzed, two main reasons of the wrong identification are found: one the size of document is extremely short, the other the format of the document is not normal, as can be seen in Table 1. The experimental result showed that this method of style identification is feasible.

(2) Question-answer pairs identification

500 samples were selected randomly from the dialogue documents identified correctly by our system, with various genre, such as news conference, portraits visit, Internet living. These 500 dialogue samples have 7425 question-answer pairs, with 6223 (83.8%) of them only followed by one paragraph (OA), 1091 (14.7%) followed by several answer paragraphs (MA), and 111 (1.5%) followed by answer paragraphs and generic content (AG). Our system is used to identify the question-answer pairs for the 500 dialogue samples mentioned above and the result is shown in Table 2, where NS (No Segmentation) denotes the segmentation technique is not used [this method was used in (Chen *et al*., 2005)], RS (Random Segmentation) denotes the segmenting operation was carried out with the probability of 20% and the segmentation point was set randomly if the number of the candidate answer paragraphs is greater than one, OM-LSA (Our Method without LSA) denotes the segmenting operation is analogous to our system but does not exploit LSA, and OM (Our Method) denotes the segmentation of our system.

Table 2 Precision of identifying question-answer pairs

	OA	MA	AG	MA+AG	Total
Num.	6223	1091	111	1202	7425
NS	1.00	1.000	0.000	0.908	0.985
RS	1.00	0.791	0.081	0.725	0.956
OM-LSA	1.00	0.833	0.541	0.806	0.969
OM	1.00	0.904	0.721	0.887	0.982

Note: the parameters k=3, θ=0.05. OA: one answer paragraph; MA: several answer paragraphs; AG: answer paragraphs and generic content; NS: no segmentation; RS: random segmentation; OM: our method; OM-LSA: our method without LSA

The experimental results showed that the identifying precision of our system satisfies the requirement of ensuring the summary coherence.

CONCLUSION AND FUTURE WORK

This paper presents some key techniques and their implementation of automatic summarization for dialogue documents. Latent Semantic Analysis (LSA) was first used to extract semantic knowledge from a given document so that the precision of text segmentation is improved 8% (i.e. 0.887−0.806). The method of automatically identifying and correlating of dialogue information units (i.e. question-answer pairs) has significantly improved the quality of the summary. With no loss of summary information and consideration of local coherence of the summary, the system farther improves the global coherence of the summary. It will be studied how to farther improve the precision of segmentation and how to determine automatically the value of the parameters in text segmentation algorithm by learning from the corpus of dialogue documents in the future.

References

Beeferman, D., Berger, A., Lafferty, J., 1999. Statistical models for text segmentation. *Machine Learning*, **34**:177-210. [doi:10.1023/A:1007506220214]

Bestgen, Y., 2006. Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, **32**(1):5-12. [doi:10.1162/coli.2006.32.1.5]

Chen, W.P., Wang, Y.C., Liu, C.H., 2005. Research on automatic summarization of spoken dialogues. *Computer Simulation*, **22**(5):226-230 (in Chinese).

Choi, F.Y.Y., Wiemer-Hastings, P., Moore, J., 2001. Latent Semantic Analysis for Text Segmentation. Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, p.109-117.

Cunningham, A.M., Wicks, W., 1992. Guide to Careers in Abstracting and Indexing. National Federation of Abstracting and Information Services, Philadelphia.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6):391-407. [doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9]

Golub, G.H., van Loan, C.F., 1996. Matrix Computations (3rd Ed.). John Hopkins University Press, Baltimore and London, p.69-74.

Hearst, M.A., 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1):33-64.

Hsueh, P.Y., Moore, J., Renals, S., 2006. Automatic Segmentation of Multiparty Dialogue. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), p.273-280.

Kaufmann, S., 1999. Cohesion and Collocation: Using Context Vectors in Text Segmentation. Proceedings of the 37th Annual Meeting of the Association of for Computational Linguistics (Student Session), p.591-595.

Kehagias, A., Pavlina, F., Petridis, P., 2003. Linear Text Segmentation Using a Dynamic Programming Algorithm. Proceedings of the European Association of Computational Linguistics. Budapest, Hungary, p.171-178.

Kozima, H., 1993. Text Segmentation Based on Similarity Between Words. Proceedings of the 31st Annual Meeting of Association for Computational Linguistics (ACL'93), p.286-288.

Li, H., Yamanishi, K., 2003. Topic analysis using a finite mixture model. *Information Processing and Management*, **39**(4):521-541. [doi:10.1016/S0306-4573(02)00035-3]

Mani, I., 2001. Automatic Summarization. John Benjamins Publishing Company, Amsterdam/Philadelphia, p.1-25.

Ponte, J.M., Croft, W.B., 1997. Text Segmentation by Topic. Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries, p.120-129.

Reynar, J.C., 1999. Statistical Models for Topic Segmentation. Proceedings of the 37th Annual Meeting of Association for Computational Linguistics (ACL'99), p.357-364.

Salton, G., Singhal, A., Buckley, C., Mitra, M., 1996. Automatic Text Decomposition Using Text Segments and Text Themes. Proceedings of the 7th ACM Conference on Hypertext (Hypertext'96), p.53-65.

Wang, Z.Q., Wang, Y.C., Gao, K., 2005. A New Model of Document Structure Analysis. FSKD 2005, LNAI 3614, p.658-666.

Wu, Y., Liu, T., Wang, K.Z., Chen, B., 1998. Research on the method of Chinese automatic abstracting. *Journal of Chinese Information Processing*, **12**(2):8-16 (in Chinese).

Wu, C.H., Yeh, J.F., Chen, M.J., 2005. Domain-specific FAQ retrieval using independent aspects. *ACM Transactions on Asian Language Information Processing*, **4**(1):1-17. [doi:10.1145/1066078.1066079]

Zechner, K., 2001. Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains. Proceedings of the 24th ACM SIGIR International Conference on Research and Development in Information Retrieval. New Orleans, LA, USA, p.199-207. [doi:10.1145/383952.383989]

Zechner, K., Lavie, A., 2001. Increasing the Coherence of Spoken Dialogue Summaries by Cross-speaker Information Linking. Proceedings of the NAACL-01 Workshop on Automatic Summarization. Pittsburgh, PA, p.22-31.

Zechner, K., 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, **28**(4):447-485. [doi:10.1162/089120102762671945]

Zhang, P., Soergel, D., 2006. Knowledge-based Approaches to the Segmentation of Oral History Interviews. MALACH Technical Report. College of Information Studies, University of Maryland, College Park.