Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 426–437
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Extractive spoken document summarization for information retrieval

Berlin Chen *, Yi-Ting Chen

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan

Received 18 January 2006; received in revised form 14 October 2007
Available online 17 November 2007

Communicated by O. Siohan

Abstract

The purpose of extractive summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document according to a target summarization ratio and then sequence them to form a concise summary. In this paper, we proposed the use of probabilistic latent topical information for extractive summarization of spoken documents. Various kinds of modeling structures and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with several conventional spoken document summarization models. The experiments were performed on the Chinese broadcast news collected in Taiwan. Noticeable performance gains were obtained. The proposed summarization technique has also been properly integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices.

Keywords: Extractive summarization; Information retrieval; Topical mixture model; Spoken documents; Speech recognition

1. Introduction

Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will play a more active role and will serve as the major human–machine interface for the interaction between people and different kinds of smart devices in the near future (Juang and Furui, 2000). On the other hand, huge quantities of multimedia content including speech information, such as that in broadcast radio and television programs, lectures, voice mails, digital libraries, and so on, are continuously growing and filling our computers, networks and lives. It is obvious that speech is one of the most important sources of information about multimedia content. However, unlike text documents, which are structured with titles and paragraphs and are thus easier to retrieve and browse, associated spoken documents of multimedia content are only presented with video or audio signals; hence, they are difficult to browse from beginning to end. Even though spoken documents are automatically transcribed into words, incorrect information (resulting from recognition errors and inaccurate sentence or paragraph boundaries) and redundant information (generated by disfluencies, fillers, and repetitions) prevent them from being accessed easily. Hence, in the recent past, several attempts have been made to investigate the possibility of understanding and organization of multimedia content using speech (Lee and Chen, 2005; Koumpis and Renals, 2005b), and substantial efforts and very encouraging results on spoken document transcription, retrieval and summarization have been reported (Woodland, 2002; Meng et al., 2004; Furui et al., 2004).

Spoken document summarization, which attempts to distill important information and remove redundant and incorrect content from spoken documents, can help users review spoken documents efficiently and understand

* Corresponding author. Tel.: +886 2 29322411x203; fax: +886 2 29322378.
E-mail address: berlin@csie.ntnu.edu.tw (B. Chen).
URL: http://berlin.csie.ntnu.edu.tw (B. Chen).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.10.022

associated topics quickly. Although research into automatic summarization of text documents dates back to the early 1950s, for nearly four decades, research work has suffered from a lack of funding. However, the development of the World Wide Web led to a renaissance of the field and summarization was subsequently extended to cover a wider range of tasks, including multidocument, multilingual, and multimedia summarization (Mani and Maybury, 1999). Generally, summarization can be either extractive or abstractive. Extractive summarization selects indicative sentences, passages, or paragraphs from an original document according to a target summarization ratio and sequences them to form a summary. Abstractive summarization, on the other hand, produces a concise abstract of a certain length that reflects the key concepts of the document. The latter is more difficult to achieve, thus recent research has focused on the former.

Quite several approaches have been developed for extractive spoken document summarization, and they may roughly fall into three main categories: (1) approaches based on the sentence structure or location information, (2) approaches based on statistical measures, and (3) approaches based on sentence classification. In (Baxendale, 1958; Hirohata et al., 2005), the authors suggested that important sentences can be selected from the significant parts of a document. For example, sentences can be selected form the introductory and/or concluding parts. However, such approaches can be only applied to some specific domains or document structures. On the other hand, statistical approaches for extractive spoken document summarization attempt to select salient sentences based on statistical features of the sentences or of the words in the sentences. Statistical features, for example, can be the term (word) frequency, language model probability, linguistic score and recognition confidence measure, as well as the prosodic information. The associated methods based on these features have gained much attention of research. Among them, the vector space model (VSM) (Gong and Liu, 2001), latent semantic analysis (LSA) method (Gong and Liu, 2001), maximum marginal relevance (MMR) method (Murray et al., 2005), sentence significant score method (Goldstein et al., 1999; Furui et al., 2004) are the most popular for spoken document summarization. Besides, a bulk of classification-based methods using statistical features and/or sentence structure (or position) information also have been developed, such as the Gaussian mixture models (GMM) (Murray et al., 2005), Hidden Markov Models (HMM) (Conroy and O'Leary, 2001; Barzilay and Lee, 2004; Maskey and Hirschberg, 2006), Bayesian network classifier (Kupiec et al., 1995; Maskey and Hirschberg, 2005), support vector machine (SVM), conditional random fields (CRF) (Galley, 2006), and logistic regression (Zhu and Penn, 2005). In these methods, sentence selection is usually formulated as a binary classification problem. A sentence can either be included in a summary or not. These methods, however, need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers. In recent years, there also has a great deal of research on exploring other extra information clues (e.g., word-clusters, WordNet or event relevance) and novel ranking algorithms for extractive text document summarization (Amini et al., 2005; Bellare et al., 2004; Li et al., 2006; Liu et al., 2007; Bollegala et al., 2006). Excellent survey articles on the development of a wide variety of summarization methods can be found in Mani and Maybury (1999).

The above approaches can be applied to both text and spoken documents. Nevertheless, the spoken documents bring extra difficulties such as the recognition errors, problems with spontaneous speech, and lack of correct sentence or paragraph boundaries. In order to avoid the redundant or incorrect parts while selecting the important and correct information, multiple recognition hypotheses, confidence scores, language model scores and other grammatical knowledge have been utilized (Furui et al., 2004; Hirohata et al., 2005). In addition, prosodic features (e.g., intonation, pitch, energy, pause duration) can be used as important clues for summarization as well; although reliable and efficient approaches to use these prosodic features are still under active research (Koumpis and Renals, 2005a; Maskey and Hirschberg, 2005). On the other hand, the summary of spoken documents can be in either text or speech form. The former has the advantage of easier browsing and further processing, but it is subject to speech recognition errors, as well as the loss of the speaker's emotional/prosodic information, which can only be conveyed by speech signals.

In contrast to conventional approaches, we address the issue of extractive summarization under a probabilistic generative framework (Chen et al., 2007). Each sentence of a spoken document to be summarized is treated as a probabilistic generative model for generating the document, and sentences are ranked and selected according to their likelihoods. We investigate the use of a topical mixture model (TMM) (Chen et al., 2006) for spoken document summarization, whereby each sentence of a spoken document to be summarized is modeled as a TMM for generating the document. Various kinds of modeling structures and training approaches are investigated. Moreover, the summarization capabilities are verified by comparison with the other summarization models. The proposed summarization model has also been successfully integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices (Chen et al., 2005).

The remainder of this paper is organized as follows. In Section 2, we elucidate the structural characteristics of the topical mixture model and its use for extractive spoken document summarization. The experimental settings are introduced in Section 3, while a series of summarization experiments are presented in Section 4. Then, a prototype system for voice retrieval of Mandarin broadcast via mobile devices is described in Section 5. Finally, conclusions are drawn in Section 6.

2. Extractive spoken document summarization

In this section we will first describe the proposed probabilistic generative framework for extractive spoken document summarization, and then concentrate on elucidating the structural characteristics of the topical mixture model (TMM) exploited for this purpose.

2.1. Probabilistic generative framework

In the probabilistic generative framework for extractive spoken document summarization, an important sentence S of a document D to be summarized can be selected (or ranked) based the posterior probability of the sentence given the document, i.e., $P(S|D)$, which is transformed to the following equation by applying Bayes' rule:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)}, \tag{1}$$

where $P(D|S)$ is the sentence generative probability, i.e., the likelihood of D being generated by S; $P(S)$ the prior probability of S being important; and $P(D)$ is the prior probability of D. $P(D)$ in Eq. (1) can be eliminated because it is identical for all sentences and will not affect the ranking of them. The sentence generative probability $P(D|S)$ can be taken as a relevance measure between the document and sentences, while the sentence prior probability to some extent is a measure of importance of the sentences themselves. Normally, because the way to estimate the prior probability of the sentences is still an open issue, we might simply assume that the prior probability is uniformly distributed in this study. Therefore, the sentences of the spoken document D to be summarized can be ranked by means of the probability $P(D|S)$ instead of using the probability $P(S|D)$.

If the document D is treated as a sequence of input observations (terms or words), $D = w_1 w_2 \ldots w_n \ldots w_N$, where the document terms are assumed to be conditionally independent given the sentence S, the relevance measure $P(D|S)$ can be decomposed as a product of the probabilities of the document terms generated by the sentence S:

$$P(D|S) = \prod_{w_n \in D} P(w_n|S)^{c(w_n,D)}, \tag{2}$$

where $c(w_n, D)$ is the occurrence count of a term (or word) w_n in the document D.

2.2. Topical mixture model (TMM)

Each individual sentence S of the spoken document D to be summarized can be further interpreted as a probabilistic generative topical mixture model (TMM), as depicted in Fig. 1. In this model, a set of K latent topical distributions characterized by unigram language models are used to predict the document terms, and each of the latent topics is associated with a sentence-specific weight. That is, each document can belong to many topics. The sentence generative probability therefore can be further expressed as

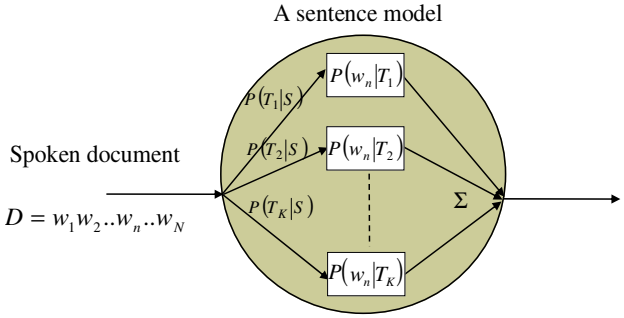

Fig. 1. The TMM model for a specific sentence S.

$$P_{\mathrm{TMM}}(D|S) = \prod_{w_n \in D} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S) \right]^{c(w_n,D)}, \tag{3}$$

where $P(w_n|T_k)$ and $P(T_k|S)$ denote, respectively, the probability of the term w_n occurring in a specific latent topic T_k and the posterior probability (or weight) of topic T_k conditioned on the sentence S. More precisely, the topical unigram distributions, e.g., $P(w_n|T_k)$, are tied among the sentences, while each sentence S has its own probability distribution over the latent topics, e.g., $P(T_k|S)$. Notice that such a relevance measure is not computed directly based on the frequency of the document terms occurring in the sentence, but instead through the frequency of the document terms in the latent topics as well as the likelihood that the sentence generates the respective topics, which in fact exhibits some sort of concept matching (Lee and Chen, 2005).

During training, a set of contemporary (or in-domain) text news documents D with corresponding human-generated titles (a title can be viewed as an extremely short summary of a document) can be collected to train the latent topical distributions $P(w_n|T_k)$ of the sentence TMM models. For each document D_j of the text news collection $\boldsymbol{D}$, the human-generated title H_j of D_j is also treated as a TMM model used to generate D_j itself:

$$P_{\mathrm{TMM}}(D_j|H_j) = \prod_{w_n \in D_j} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|H_j) \right]^{c(w_n,D_j)}. \tag{4}$$

The K-means algorithm (Ball and Hall, 1967) is first used to partition the entire titles of the document collection into K topical clusters in an unsupervised manner, and the initial topical unigram distribution $P(w_n|T_k)$ for a cluster topic T_k can be estimated according to the underlying statistical characteristics of document titles being assigned to it. The probabilities for each title generating the topics, i.e., $P(T_k|H_j)$, are measured according to its proximity to the centroid of each respective cluster as well. Then, the

probability distributions $P(w_n|T_k)$ and $P(T_k|H_j)$ can be further optimized by maximizing the total log-likelihood L_T of all the documents D_j in the collection $\boldsymbol{D}$ generated by their corresponding title TMM models, using the expectation-maximization (EM) algorithm (Dempster et al., 1977):

$$L_T = \sum_{D_j \in \boldsymbol{D}} \log P_{TMM}(D_j|H_j). \quad (5)$$

Interested readers can refer to (Manning and Schutze, 1999, p. 523) for more derivation details of the training of (topical) mixture models using the EM algorithm.

Our postulation is that the latent topical factors $P(w_n|T_k)$ properly constructed based on the "title-document" relationships might provide very helpful clues for the subsequent spoken document summarization task. In this way, when performing extractive summarization of a broadcast news document D, we can keep the latent topical factors $P(w_n|T_k)$ unchanged, as those previously obtained by the contemporary (or in-domain) text news documents, but estimate the probability distributions of the sentence TMM model over the latent topics $P(T_k|S)$ on the fly, by maximizing the log-likelihood of the document D generated by the sentence TMM model, using the EM algorithm. Once the TMM models for the sentences are estimated, they can thus be used to predict the occurrence probability of the terms in the spoken document, and the sentences with highest probabilities can be thus selected and sequenced to form the final summary according to different summarization ratios. Fig. 2 depicts a schematic representation of extractive spoken document summarization using the TMM models.

Structures similar to the presented topical mixture model also have been extensively investigated in the information retrieval (IR) task recently (Hofmann, 2001; Blei et al., 2003; Chen et al., 2004c). More detailed elucidation and comparison of the training of these models for IR can be found in Chen (2006).

3. Experimental setup

3.1. Speech and text corpora

The speech data set was comprised of approximately 176 h of radio and TV broadcast news documents collected from several radio and TV stations in Taipei between 1998 and 2004 (Chen et al., 2005). From them, a collection of 200 broadcast news documents (1.6 h) collected in August 2001, was reserved for the summarization experiments (Lee and Chen, 2005). These spoken documents were divided into two parts, each of which contained 100 spoken documents. The first part of spoken documents was taken as the development set, which formed the basis for tuning parameters or settings. The rest part of spoken documents was taken as the evaluation set; i.e., all the summarization experiments on it were conducted following the same training (or parameter) settings and model complexities that were optimized based on the development set. Therefore, the experimental results can validate the effectiveness of the proposed approach on comparable real-world data.

The remainder of the speech data was used to train a set of acoustic models for speech recognition, of which about 4.0 h of data with corresponding orthographic transcripts was used to bootstrap the acoustic model training, while 104.3 h of the remaining un-transcribed speech data was reserved for unsupervised acoustic model training (Chen et al., 2004a). The acoustic models were further optimized by the minimum phone error (MPE) training algorithm.

On the other hand, a large number of text news documents collected from the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) was also used. The text news documents collected in 2000 and 2001 were used to train the N-gram language models for speech recognition with the SRI Language Modeling Toolkit (Stolcke, 2005). A subset of about 14,000 text news documents collected in the same time period as that of the broadcast news documents to be summa-

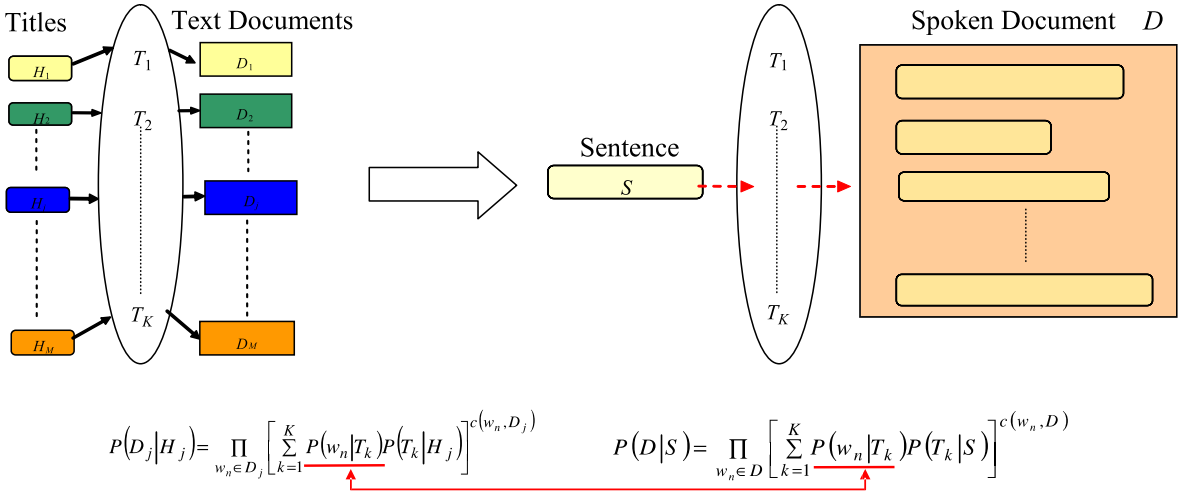

Fig. 2. A schematic representation of extractive spoken document summarization using the TMM models.

rized (August, 2001) were also used to train the latent topical distributions for TMM, and to estimate the model parameters for VSM, LSA, MMR, as well as the sentence significant score method.

3.2. Broadcast news transcription system

The front-end processing was conducted with the HLDA (Heteroscedastic Linear Discriminant Analysis)-based data-driven Mel-frequency feature extraction approach (Kumar, 1997) and then processed by MLLT (Maximum Likelihood Linear Transformation) for feature de-correlation (Saon et al., 2000). Utterance-based feature mean subtraction and variance normalization were further applied. On the other hand, the speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree-copy search as well as a lexical prefix tree organization of the lexicon (Aubert, 2002). At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding unigram language model look-ahead scores and syllable-level acoustic model look-ahead scores (Chen et al., 2004a), was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores higher than a predefined threshold, their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, will be kept in order to build a word graph for further language model rescoring (Ortmanns et al., 1997). In this study, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word graph rescoring procedure. The Chinese character error rate (CER) achieved for the 200 broadcast news documents to be summarized was 14.17%.

3.3. Evaluation metrics

Three human subjects were instructed to do the human summarization, to be taken as the references for evaluation, in two forms. The first was simply to select 50% most important sentences in the reference transcript of the spoken (broadcast news) document and then to rank these sentences by importance without assigning a score to each sentence. The second was to write an abstract for the document by the subject himself with a length being roughly 25% of the original broadcast news story. The automatic summarization results were tested based on several summarization ratios (10%, 20%, 30% and 50%), which are the ratios of the length of the automatic summary to that of the document (Lee and Chen, 2005). On the other hand, the agreements in the cosine measure (see in the next paragraph) between the human subjects for important sentence ranking are about 0.55, 0.59, 0.65 and 0.75 for the summarization ratios of 10%, 20%, 30% and 50%, respectively; while the agreement for written abstracts is 0.55.

Two evaluation metrics were employed in this paper. The first is the cosine measure (Saggion et al., 2002), for which let E_R denote the extractive summary that was obtained from the concatenation of the top several important sentences selected by the human subject, and A_R the abstractive summary that was written by the subject. The summarization accuracy, Acc_D, of a broadcast news document D is then the averaged similarity measure for the automatic summary, E, with respect to E_R and A_R:

$$\mathrm{Acc}_D = \frac{1}{2}[\mathrm{sim}(E, E_R) + \mathrm{sim}(E, A_R)], \tag{6}$$

where the similarity measures $\mathrm{sim}(E, E_R)$ and $\mathrm{sim}(E, A_R)$ are calculated in the cosine score based on the vector representations of the automatic and reference summaries. In this way, higher accuracy would be obtained if more sentences that are important in the broadcast news document are included in the automatic summary. The final summarization accuracy is defined as the average of Acc_D in Eq. (6) over all the broadcast news documents and all the three human subjects.

In addition, the ROUGE measure (Lin, 2003; Liu et al., 2005; Hirohata et al., 2005; Maskey and Hirschberg, 2005) was also used to evaluate the performance levels of the proposed models and the conventional models. This measure evaluates the quality of the summarization by counting the number of overlapping units, such as N-grams and word sequences, between the automatic summary and a set of reference (or manual) summaries. ROUGE-N is an N-gram recall measure defined as follows:

$$\mathrm{ROUGE}-N = \frac{\sum_{E_R \in \boldsymbol{E_R}} \sum_{\mathrm{gram}_N \in E_R} \mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_N)}{\sum_{E_R \in \boldsymbol{E_R}} \sum_{\mathrm{gram}_N \in E_R} \mathrm{Count}(\mathrm{gram}_N)}, \tag{7}$$

where E_R is an individual reference (or manual) summary; $\boldsymbol{E_R}$ is a set of reference summaries; $\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_N)$ is the maximum number of N-grams co-occurring in the automatic summary and the corresponding reference summary; and $\mathrm{Count}(\mathrm{gram}_N)$ is the number of N-grams in the reference summary. Notice here that in this paper, the ROUGE-2 measure using word bigrams as matching units was adopted, which was computed by matching the bigrams of the automatic extractive summary with that of the reference extractive summary.

4. Experimental results

4.1. Summarization results of TMM on the development set

In this study, when TMM was employed in the extractive spoken document summarization task, we additionally incorporated the unigram probability of a term (or a word) w_n occurring in the sentence S of a document D to be summarized into Eq. (3) for probability smoothing:

$$\widehat{P}_{\mathrm{TMM}}(D|S) = \prod_{w_n \in D} \left[\alpha \left(\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S) \right) + (1-\alpha) P(w_n|S) \right]^{c(w_n,D)}, \quad (8)$$

where $P(w_n|S)$ is the unigram probability of a term (or a word) w_n occurring in the sentence S, and α is a weighting parameter whose value also can be optimized using the EM algorithm.

We first evaluate the summarization performance of the TMM summarization approach on the development set. The latent topical distributions $P(w_n|T_k)$ were estimated beforehand using a set of contemporary text news documents and their corresponding human-generated titles, as previously described in Sections 2.2 and 3.1; while the probability distributions of the sentence TMM models over the latent topics $P(T_k|S)$ were estimated on the fly in an unsupervised manner during the summarization process. That is, the document-reference summary information of the 100 spoken documents in the development set was not utilized for training the sentence TMM models. The summarization results of TMM on the development set in the cosine measure and the ROUGE-2 measure were, respectively, shown in Tables 1 and 2, where each column illustrates the results for different summarization ratios and different latent topics used. Moreover, the results assuming manual transcripts with correct sentence boundaries for the spoken documents to be summarized (denoted TD, text documents) are also shown for reference, compared to the results when only erroneous recognition transcripts with sentence boundaries determined by speech pauses are available (denoted SD, spoken documents). As can be seen, the summarization results are not always improved as the topic number increases. The improvements seem to saturate for most cases when the topic number increases up to 32, thought the performance difference between different topic numbers is not so significant. The cosine accuracies for the TD case are respectively about 0.43, 0.47, 0.53 and 0.63 for summarization ratios of 10%, 20%, 30% and 50%, while the accuracies for the SD one are 0.34, 0.38, 0.43 and 0.52, respectively. On the other hand, the ROUGE-2 recall rates for the TD case are respectively about 0.42, 0.47, 0.51 and 0.65 for summarization ratios of 10%, 20%, 30% and 50%, while the recall rates for the SD one are 0.31, 0.34, 0.36 and 0.47, respectively. Speech recognition errors might be the main reason for the considerable performance gap between the TD case and the SD one, while another reason might result from the poor sentence boundary detection algorithm employed in this paper, which simply utilizes pause information for such purpose. Therefore, a more sophisticated sentence boundary detection algorithm using either prosodic or lexical

Table 1
The results (in the cosine measure) on the development set achieved by the TMM models using different mixture numbers and under different summarization ratios

		2	4	8	16	32	64
10%	TD	0.4217	0.4217	0.4209	0.4252	0.4263	0.4248
	SD	0.3400	0.3400	0.3400	0.3405	0.3377	0.3421
20%	TD	0.4704	0.4683	0.4657	0.4648	0.4707	0.4697
	SD	0.3775	0.3763	0.3763	0.3769	0.3758	0.3796
30%	TD	0.5191	0.5187	0.5187	0.5204	0.5293	0.5287
	SD	0.4212	0.4206	0.4208	0.4210	0.4239	0.4258
50%	TD	0.6259	0.6253	0.6254	0.6250	0.6282	0.6248
	SD	0.5117	0.5140	0.5151	0.5149	0.5161	0.5164

The latent topical distributions of the TMM models were trained with the contemporary text news documents.

Table 2
The results (in the ROUGE-2 measure) on the development set achieved by the TMM models using different mixture numbers and under different summarization ratios

		2	4	8	16	32	64
10%	TD	0.4195	0.4195	0.4242	0.4195	0.4242	0.4242
	SD	0.3135	0.3135	0.3135	0.3154	0.3055	0.3133
20%	TD	0.4615	0.4615	0.4599	0.4545	0.4683	0.4673
	SD	0.3495	0.3484	0.3484	0.3496	0.3395	0.3471
30%	TD	0.5002	0.5002	0.5052	0.4995	0.5090	0.5123
	SD	0.3622	0.3605	0.3599	0.3600	0.3632	0.3633
50%	TD	0.6516	0.6516	0.6516	0.6495	0.6529	0.6482
	SD	0.4607	0.4630	0.4651	0.4657	0.4661	0.4584

The latent topical distributions of the TMM models were trained with the contemporary text news documents.

information would be helpful for the summarization task studied here (Maskey and Hirschberg, 2005).

On the other hand, in most real-world applications, it is not always the case that the spoken document summarization systems can have contemporary or in-domain text news documents for model training. Thus, we study here the use of unsupervised training for TMM by merely using all possible "sentence-document" pairs of the spoken (broadcast news) documents in the development set to construct the latent topical space. That is, each sentence of the spoken document in the development set, regardless of whether it belongs to the reference summary or not, was treated as a sentence TMM model and involved in the construction of the latent topical distributions $P(w_n|T_k)$, while the probability distributions of the sentence TMM models over the latent topics $P(T_k|S)$ were again estimated on the fly during the summarization process, as described previously in Section 2.2. The results are shown in Tables 3 and 4 for different evaluation metrics. Compared to the results in Tables 1 and 2, it can be found that the results obtained by using smaller mixture numbers are quite competitive to those of TMM trained with the contemporary text news documents, but the results for larger mixtures are apparently degraded instead. One possible reason might be that the "sentence-document" pairs used here may not be adequate for the correct construction of the latent topical space.

4.2. Comparisons with other summarization models

Then, we try to compare the TMM model with the conventional VSM, LSA and MMR methods, as well as the sentence significant score method. VSM is a typical example for literal term matching, while LSA for concept matching (Lee and Chen, 2005). VSM represents each sentence of a document, and the whole document, in vector form. In this approach, each dimension specifies the weighted statistics, for example the product of the term frequency (TF) and inverse document frequency (IDF) (Baeza-Yates and Ribeiro-Neto, 1999), associated with an indexing term (or word) in the sentence or document. Sentences with the highest relevance scores (usually calculated by the cosine score of two vectors) to the whole document are included in the summary (Gong and Liu, 2001). LSA (denoted as LSA-1 below) instead represents each sentence of a document as a vector in the latent semantic space of the document, which is constructed by performing singular value decomposition (SVD) on the "term-sentence" matrix of the document. The right singular vectors with larger singular values represent the dimensions of the more important latent semantic concepts in the document. Therefore, the sentences with the largest index values in each of the top L right singular vectors are included in the summary (Gong and Liu, 2001). An alternative LSA-based approach (denoted as LSA-2 below) proposed in (Hirohata et al.,

Table 3
The results (in the cosine measure) on the development set achieved by the TMM models using different mixture numbers and under different summarization ratios

		2	4	8	16	32	64
10%	TD	0.4206	0.4333	0.4365	0.4083	0.3916	0.4003
	SD	0.3485	0.3428	0.3460	0.3466	0.3419	0.3304
20%	TD	0.4693	0.4752	0.4709	0.4394	0.4419	0.4351
	SD	0.3888	0.3733	0.3757	0.3746	0.3685	0.3636
30%	TD	0.5249	0.5156	0.5262	0.4998	0.4981	0.4947
	SD	0.4206	0.4148	0.4109	0.4159	0.4204	0.4035
50%	TD	0.6177	0.6209	0.6098	0.6059	0.6068	0.6085
	SD	0.5174	0.5203	0.5097	0.5089	0.5123	0.5065

The latent topical distributions of the TMM models were trained with the development set.

Table 4
The results (in the ROUGE-2 measure) on the development set achieved by the TMM models using different mixture numbers and under different summarization ratios

		2	4	8	16	32	64
10%	TD	0.4105	0.4340	0.4142	0.3834	0.3670	0.3843
	SD	0.3328	0.3058	0.3173	0.3021	0.3233	0.2946
20%	TD	0.4635	0.4707	0.4484	0.4241	0.4182	0.4218
	SD	0.3646	0.3466	0.3425	0.3340	0.3479	0.3122
30%	TD	0.4897	0.4940	0.4853	0.4510	0.4752	0.4642
	SD	0.3641	0.3566	0.3467	0.3425	0.3507	0.3276
50%	TD	0.6324	0.6270	0.6290	0.6192	0.6146	0.6111
	SD	0.4777	0.4801	0.4614	0.4644	0.4577	0.4516

The latent topical distributions of the TMM models were trained with the development set.

2005) was evaluated as well. LSA-2 simply computed the score of each sentence based on the norm of its vector representation in the lower m-dimensional latent semantic space, and a fixed number of sentences having relatively large scores were therefore selected to form the summary. The value of m was set at 5 in our experiments, which was just the same as that in (Hirohata et al., 2005). In this paper, these two LSA models were implemented with the MIT SVD Toolkit (Rohde, 2005). MMR actually is close in spirit to VSM, because MMR also represents each sentence of a document and the document itself in vector form and uses the cosine score for sentence selection. However, MMR performs sentence selection iteratively with the criteria of topic relevance and coverage. The next sentence S_u to be selected is ranked according to two criteria: (1) whether it is more similar to the whole document D and (2) whether it is less similar to the set $\boldsymbol{S}_l$ of sentences that have been selected so far, using the following formula:

$$\text{NextSen} = \max_{S_u}[\beta \cdot \text{sim}(S_u, D) - (1-\beta) \times \max_{S_j \in \boldsymbol{S}_l} \text{sim}(S_u, S_j)], \quad (9)$$

where β is a parameter used to trade off between relevance and redundancy (Murray et al., 2005). Therefore, MMR can not only select relevant sentences into the summary but also make it cover more different concepts.

On the other hand, the sentence significant score method (denoted as SIG below) selects indicative sentences from the spoken document based on the sentence significance score (Goldstein et al., 1999). For example, given a sentence $S = \{w_1, w_2, \ldots w_n, \ldots w_{N_S}\}$ with length N_S, the significance score of S can be expressed using the following formula:

$$\text{Sig}(S) = \frac{1}{N_S}\sum_{n=1}^{N_S}[\gamma \cdot I(w_n) + (1-\gamma)L(w_n)], \quad (10)$$

where $I(w_n)$ is the product of TF and IDF of term w_n, $L(w_n)$ is the logarithm of the bigram probability of w_n given its predecessor term w_{n-1} in S, which was estimated from a large contemporary text corpus, and γ is a weighting parameter. In this paper, a language modeling approach (denoted as LM below) to spoken document summarization was also proposed. In such an approach, each sentence of a document to be summarized was treated as a probabilistic generative model consisting of N-gram distributions for predicting the document, which were directly estimated from each sentence itself and smoothed by N-gram distributions estimated from a large text corpus. In this paper, only unigram modeling was initially investigated for LM:

$$P_{\text{LM}}(D|S) = \prod_{w_n \in D}[\lambda \cdot P(w_n|S) + (1-\lambda) \times P(w_n|\text{Corpus})]^{c(w_n,D)}, \quad (11)$$

where λ is a weighting parameter. The LM model can be viewed as a probabilistic counterpart of VSM for literal term matching, and the estimation of its model parameters is in fact conducted in an unsupervised manner without the use of the document-reference summary information.

The summarization results of the above models on the development set in the cosine measure and the ROUGE-2 measure are, respectively, shown in Tables 5 and 6, and the results obtained by random selection (denoted as RAND) are also listed for comparison. Notice here that all above models were tuned with optimum settings directly based on the development set and the contemporary text news documents collected in August 2001, as described in Section 3.1. However, the document-reference summary information of the development set was again not utilized in the construction of these models. As we compare the best results of TMM shown in Tables 1–4, several observations can be drawn. First, TMM is substantially better than LSA and SIG, and competitive with VSM, MMR and LM, when the summarization results are evaluated using the cosine measure. One possible reason for the good performance of VSM and MMR is mainly because that the sentence selection criterion for both VSM and MMR is based on the vector representation and the cosine score, which actually is quite analogous to those adopted in the cosine measure of summarization performance. Second, as the summarization performance is evaluated using the ROUGE-2 measure, TMM significantly outperforms VSM, MMR, LSA and SIG, and its performance is again on par with LM. Third, if we go a step further by looking into the detailed summarization results obtained by TMM and LM, we can find that the best results of TMM are

Table 5
The results (in the cosine measure) on the development set achieved by the conventional approaches under different summarization ratios

		VSM	LSA-1	LSA-2	MMR	SIG	LM	RAND
10%	TD	0.4199	0.3717	0.3273	0.4212	0.3914	0.4203	0.1923
	SD	0.3553	0.3189	0.3047	0.3553	0.3299	0.3415	0.1900
20%	TD	0.4657	0.4069	0.3978	0.4685	0.4448	0.4657	0.2407
	SD	0.3889	0.3384	0.3445	0.3904	0.3645	0.3789	0.2317
30%	TD	0.5148	0.4676	0.4766	0.5171	0.5171	0.5178	0.3303
	SD	0.4221	0.3761	0.3950	0.4226	0.4178	0.4281	0.2793
50%	TD	0.6206	0.5821	0.6171	0.6186	0.6132	0.6271	0.5013
	SD	0.5118	0.4870	0.5058	0.5108	0.5295	0.5159	0.4193

Table 6
The results (in the ROUGE-2 measure) on the development set achieved by the conventional approaches under different summarization ratios

		VSM	LSA-1	LSA-2	MMR	SIG	LM	RAND
10%	TD	0.3836	0.3487	0.3270	0.3855	0.3379	0.4157	0.0931
	SD	0.2991	0.2901	0.2861	0.2991	0.2841	0.3084	0.1215
20%	TD	0.4341	0.3797	0.3937	0.4335	0.3968	0.4591	0.1329
	SD	0.3337	0.3123	0.3237	0.3392	0.3133	0.3467	0.1470
30%	TD	0.4738	0.4269	0.4709	0.4792	0.4896	0.4947	0.2237
	SD	0.3328	0.3094	0.3451	0.3363	0.3563	0.3734	0.1702
50%	TD	0.6124	0.5710	0.6364	0.6165	0.6242	0.6527	0.4183
	SD	0.4441	0.4113	0.4568	0.4452	0.4823	0.4768	0.3177

slightly better than those of LM for most cases, especially for lower summarization ratios ($\leqslant$20%).

4.3. Further summarization results on the evaluation set

In order to validate the effectiveness of the proposed TMM summarization approach on comparable real-world data, we further conducted a series of corresponding spoken document summarization experiments on the evaluation set. For TMM and LM, the settings and/or model complexities for different experimental conditions (TD and SD cases) are set with the same configurations as those optimized using the development set; while for VSM, LSI, MMR and SIG the model parameters are also set at the same optimum values tuned based on the development set as well. Here we also compare the TMM model with the Bayesian network classifier (denoted as BNC below) on the evaluation set (Kupiec et al., 1995; Maskey and Hirschberg, 2005). The unigram features as well as the document-reference summary information of the 100 spoken documents in the development set were utilized for training the BNC model. The summarization results in the cosine measure and the ROUGE-2 measure achieved by using TMM and the other models are respectively shown in Tables 7 and 8, where TMM-1 denotes the results of TMM whose latent topical distributions were trained with the contemporary text news documents and TMM-2 denotes the results of TMM whose latent topical distributions were instead trained with the 100 spoken documents of the development set, as previously described in Section

Table 7
The results (in the cosine measure) on the evaluation set achieved by TMM and the other conventional approaches under different summarization ratios

		TMM-1	TMM-2	VSM	LSA-1	LSA-2	MMR	SIG	LM	BNC	RAND
10%	TD	0.3981	0.3959	0.3701	0.3681	0.3604	0.3693	0.3571	0.3841	0.3173	0.1896
	SD	0.3302	0.3140	0.3123	0.3065	0.2960	0.3123	0.3114	0.3155	0.2401	0.1871
20%	TD	0.4173	0.4152	0.3974	0.3861	0.3849	0.4016	0.3810	0.4064	0.3449	0.2173
	SD	0.3503	0.3445	0.3314	0.3214	0.3257	0.3349	0.3349	0.3381	0.2557	0.2089
30%	TD	0.4928	0.4930	0.4737	0.4292	0.4474	0.4780	0.4476	0.4855	0.4202	0.3203
	SD	0.4056	0.4037	0.3956	0.3523	0.3710	0.3977	0.3785	0.3993	0.3080	0.2546
50%	TD	0.6003	0.6041	0.6119	0.5647	0.6077	0.6169	0.5915	0.6037	0.5457	0.4917
	SD	0.4992	0.5014	0.4912	0.4676	0.4862	0.4940	0.4996	0.5027	0.4214	0.3996

Table 8
The results (in the ROUGE-2 measure) on the evaluation set achieved by TMM and the other conventional approaches under different summarization ratios

		TMM-1	TMM-2	VSM	LSA-1	LSA-2	MMR	SIG	LM	BNC	RAND
10%	TD	0.3871	0.3782	0.3212	0.3582	0.3667	0.3212	0.3298	0.3714	0.3270	0.1314
	SD	0.3210	0.3016	0.2847	0.2861	0.2751	0.2847	0.2888	0.2932	0.2187	0.1204
20%	TD	0.3953	0.3871	0.3426	0.3698	0.3859	0.3483	0.3525	0.3892	0.3412	0.1548
	SD	0.3333	0.3217	0.2980	0.2862	0.2997	0.3035	0.3060	0.3191	0.2218	0.1392
30%	TD	0.4659	0.4656	0.4396	0.3789	0.4351	0.4423	0.4274	0.4669	0.3893	0.2543
	SD	0.3741	0.3618	0.3476	0.2910	0.3336	0.3484	0.3261	0.3705	0.2507	0.1679
50%	TD	0.6280	0.6191	0.6138	0.5603	0.6311	0.6241	0.6206	0.6308	0.5088	0.4353
	SD	0.4605	0.4683	0.4468	0.3998	0.4431	0.4505	0.4656	0.4732	0.3377	0.3163

4.1. Comparatively speaking, the summarization results on the evaluation set obtained by TMM are significantly better than those obtained by the other models in most conditions.

Based on the experimental results achieved from this and previous sections, it has been clearly demonstrated that TMM does achieve superior performance over the other summarization models, which also evidences that TMM is indeed a good candidate of concept matching for the Chinese spoken document summarization task studied here. All the summarization experiments reported here have been carefully designed to avoid "testing on training"; i.e., all the training (or parameter) settings and model complexities are tuned or optimized by using the development set and tested on both the development set and the evaluation set. Generally speaking, the training settings and model complexities tuned from the development set perform rather well in the evaluation set.

5. Spoken document retrieval system

We have implemented a prototype system that allows the user to search for Mandarin broadcast news via the PDA using a spoken natural language query. The framework of the system is shown in Fig. 3. There is a small client program on the PDA, as illustrated in Fig. 4, which transmits the speech waveform or acoustic feature data of the spoken query to the information retrieval server. The information retrieval server then passes the speech waveform or acoustic feature data to the large vocabulary continuous speech recognition (LVCSR) server, which works in the similar way as the broadcast news transcription system shown earlier in Section 3.2. The recognition result is then passed back to the information retrieval server to act as the query to generate a ranked list of relevant documents. When the retrieval results are sent back to the PDA, the user can first browse the summaries of the retrieved documents and then click to read the speech transcripts of the relevant broadcast news documents or play the corresponding audio files from the audio streaming server. On the other hand, the huge collection of broadcast news documents, as described previously in Section 3, is offline recognized by the broadcast news transcription system and the resultant transcripts are then utilized by the multiscale indexer to generate the word-level and syllable-level indexing terms. Only the VSM model for literal term matching of the spoken query and the spoken documents was implemented here for simplicity, although our previous experiments on Mandarin spoken document retrieval have demonstrated the TMM-based or HMM-based retrieval models have superior retrieval performance over VSM (Chen, 2006; Chen et al., 2004b). The final retrieval indices, including the vocabularies and document occurrences of indexing terms of different types (word- and syllable-level indexing terms), are stored as inverted files for efficient searching and comparison (Baeza-Yates and Ribeiro-Neto, 1999).

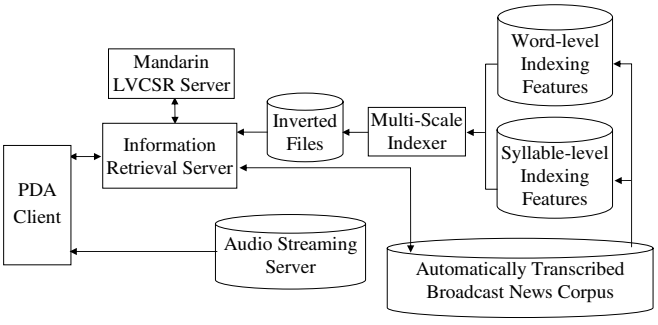

Fig. 3. The framework for speech retrieval of Mandarin broadcast news.

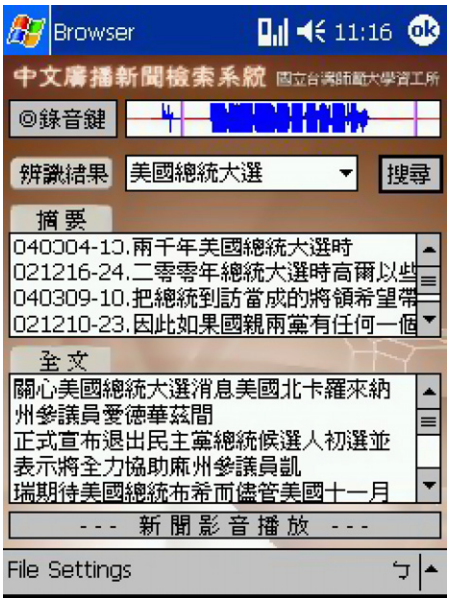

Fig. 4. A PDA-based broadcast news retrieval system that displays the retrieved spoken documents and their associated summaries for efficient browsing, in which the middle part lists the summaries of retrieved spoken documents, while the low part the transcript of the selected document.

In order to evaluate the performance level of the retrieval system, a set of 20 simple queries with length of one to several words, in both text and speech forms, was manually created. Four speakers (two males and two females) produced the 20 queries using an Acer n20 PDA with its original microphone in an environment of slight background noise. To recognize these spoken queries, another read speech corpus consisting of 8.5 h of speech produced by an additional 39 male and 38 female speakers over the same type of PDAs was used for training the speaker-independent acoustic models for recognition of the spoken queries. The character and syllable error rates for the spoken queries are 27.61% and 19.47%, respectively. The results are not as good as that of broadcast news transcription reported earlier, it is probably because that most of the test queries contain one to several out-of-vocabulary (OOV) words, such as personal names and new organization or event names, which apparently occur much more fre-

quently in the queries than in the broadcast news documents and may degrade the speech recognition performance severely. The retrieval experiments were performed with respect to a collection of about 21,000 broadcast news stories. The results in terms of *mean* average precision (*m*AP) (Harman, 1995) at a document cutoff number of ten were 0.8038 and 0.6237 for text and spoken queries, respectively. While at a document cutoff number of 30, the mean average precision were 0.6692 and 0.5232 for text and spoken queries, respectively (Chen et al., 2005).

6. Conclusions

In this paper, we have studied the use of topical mixture model for extractive spoken document summarization. Various kinds of modeling complexities and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with the other summarization models. Noticeable and consistent performance gains were obtained. The proposed summarization technique has also been properly integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices. More in-deep investigation of the combination of TMM with other linguistic and prosodic features (Chen et al., 2007), as well as the comparison of TMM with other language modeling approaches (Pingali et al., 2007; Chen and Chen, 2007; Croft and Lafferty, 2003), is currently undertaken.

References

Amini, M., Usunier, N., Gallinari, P., 2005. Automatic text summarization based on word-clusters and ranking algorithms. In: Proc. European Conf. on Information Retrieval Research, pp. 142–156.

Aubert, X.L., 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. Comput. Speech Lang. 16 (1), 89–114.

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

Ball, G.H., Hall, D.J., 1967. A clustering technique for summarizing multivariate data. Behav. Sci. 12, 153–155.

Barzilay, R., Lee, L., 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In: Proc. NAACL/HLT.

Baxendale, P.B., 1958. Machine-made index for technical literature – An experiment. IBM J., 354–361.

Bellare, K., Sarma, A.D., Sarma, A.D., Loiwal, N., Mehta, V., Ramakrishnan, G., Bhattacharya, P., 2004. Generic text summarization using WordNet. In: Proc. Internat. Conf. on Language Resources and Evaluation.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bollegala, D., Okazaki, N., Ishizuka M., 2006. A bottom-up approach to sentence ordering for multi-document summarization. In: Proc. Annual Meeting of the Assoc. of Comput. Linguistics, pp. 385–392.

Chen, B., Kuo, J.W., Tsai, W.H., 2004a. Lightly supervised and data-driven approaches to Mandarin broadcast news transcription. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Process., vol. 1, pp. 777–780.

Chen, B., Wang, H.M., Lee, L.S., 2004b. A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents. ACM Trans. Asian Lang. Inform. Process. 3 (2), 128–145.

Chen, B., Kuo, J.W., Huang, Y.M., Wang, H.M., 2004c. Statistical Chinese spoken document retrieval using latent topical information. In: Proc. Internat. Conf. on Spoken Language Process., vol. 2, pp. 1621–1625.

Chen, B., Chen, Y.T., Chang, C.H., Chen, H.B., 2005. Speech retrieval of Mandarin broadcast news via mobile devices. In: Proc. European Conf. Speech Comm. Technol., pp. 109–112.

Chen, B., 2006. Exploring the use of latent topical information for statistical Chinese spoken document retrieval. Pattern Recogn. Lett. 27 (1), 9–18.

Chen, B., Yeh, Y.M., Huang, Y.M., Chen, Y.T., 2006. Chinese spoken document summarization using probabilistic latent topical information. In: Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., vol. I, pp. 969–972.

Chen, Y.T., Chiu, H.S., Wang, H.M., Chen, B., 2007. A unified probabilistic generative framework for extractive spoken document summarization. In: Proc. European Conf. on Speech Comm. Technol.

Chen, B., Chen, Y.T., 2007. Word topical mixture models for extractive spoken document summarization. In: Proc. IEEE Internat. Conf. on Multimedia & Expo.

Croft, W.B., Lafferty, J. (Eds.), 2003. Language Modeling for Information Retrieval. Kluwer – Academic Publishers.

Conroy, J., O'Leary, D., 2001. Text summarization via hidden Markov models and pivoted QR. matrix decomposition. Technical report, University of Maryland, College Park, Maryland.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39 (1), 1–38.

Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Trans. Speech Audio Process. 12 (4), 401–408.

Galley, M., 2006. Skip-chain conditional random field for ranking meeting utterances by importance. In: Proc. Empirical Methods in Natural Language Process., pp. 364–372.

Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summarizing text documents: Sentence selection and evaluation metrics. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 121–128.

Gong, Y., Liu, X., 2001. Generic text summarization using relevance measure and latent semantic analysis. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 19–25.

Harman, D., 1995. Overview of the fourth text retrieval conference (TREC-4). In: Proc. 4th Text Retrieval Conf., pp. 1–23.

Hirohata, M., Shinnaka Y., Iwano, K., Furui, S., 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Process., vol. 1, pp. 1065–1068.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196.

Juang, B.H., Furui, S., 2000. Automatic recognition and understanding of spoken language – A first step toward natural human–machine communication. Proc. IEEE 88 (8), 1142–1165.

Koumpis, K., Renals, S., 2005a. Automatic summarization of voicemail messages using lexical and prosodic features. ACM Trans. Speech Lang. Process. 2 (1), 1–24.

Koumpis, K., Renals, S., 2005b. Content-based access to spoken audio. IEEE Signal Process. Mag. 22 (5), 61–69.

Kumar, N., 1997. Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. Thesis, John Hopkins University, Baltimore.

Kupiec, J., Pedersen, J., Chen, F., 1995. A trainable documentsummarizer. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 68–73.

Lee, L.S., Chen, B., 2005. Spoken document understanding and organization. IEEE Signal Process. Mag. 22 (5), 42–60.

Li, W., Wu, M., Lu, Q., Xu, W., Yuan, C., 2006. Extractive summarization using inter- and intra-event relevance. In: Proc. Annual Meeting of the Assoc. of Comput. Linguistics, pp. 369–376.

Lin, C.Y., 2003. ROUGE: Recall-oriented understudy for gisting evaluation. Available from: <http://haydn.isi.edu/ROUGE/>.

Liu, M., Li, W., Wu, M., Lu, Q., 2007. Extractive summarization based on event term clustering. In: Proc. Annual Meeting of the Assoc. of Comput. Linguistics, pp. 185–188.

Liu, X., Croft, W.B., 2005. Statistical Language Modeling for Information Retrieval. In: Proc. 30th Annual Review of Information Sci. and Technol., vol. 39, Chapter 1.

Mani, I., Maybury, M.T. (Eds.), 1999. Advances in Automatic Text Summarization. MIT Press, Cambridge, MA.

Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Maskey, S., Hirschberg, J., 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In: Proc. European Conf. on Speech Comm. and Technol., pp. 621–624.

Maskey, S., Hirschberg, J., 2006. Summarizing speech without text using hidden Markov models. In: Proc. HLT-NAACL 2006.

Meng, H., Chen, B., Khudanpur, S., Levow, G.A., Lo, W.K., Oard, D., Schone, P., Tang, K., Wang, H.M., Wang, J., 2004. Mandarin English information (MEI): Investigating translingual speech retrieval. Comput. Speech Lang. 18 (2), 163–179.

Murray, G., Renals, S., Carletta, J., 2005. Extractive summarization of meeting recordings. In: Proc. European Conf. on Speech Comm. and Technol., pp. 593–596.

Ortmanns, S., Ney, H., Aubert, X., 1997. A word graph algorithm for large vocabulary continuous speech recognition. Comput. Speech Lang. 11, 43–72.

Pingali, P., Jagarlamudi, J., Varma, V., 2007. Experiments in cross language query focused multi-document summarization. In: Proc. Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies.

Rohde, D., 2005. Doug Rohde's SVD C Library, Version 1.34. Available from: <http://tedlab.mit.edu:16080/~dr/SVDLIBC/>.

Saggion, H., Teufel, S., Radev, D., Lam, W., 2002. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In: Proc. Internat. Conf. Comput. Linguistics.

Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., 2000. Maximum likelihood discriminant feature spaces. In: Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., vol. 2, pp. 1129–1132.

Stolcke, A., 2005. SRI Language Modeling Toolkit, Version 1.4.4. Available from: <http://www.speech.sri.com/projects/srilm/>.

Woodland, P.C., 2002. The development of the HTK broadcast news transcription system: An overview. Speech Commun. 37, 47–67.

Zhu, X., Penn G., 2005. Evaluation of sentence selection for speech summarization. In: Proc. 2nd Internat. Conf. on Recent Advances in Natural Language Processing (RANLP-05), Workshop on Crossing Barriers in Text Summarization Research, pp. 39–45.