Mandarin Chinese Broadcast News Retrieval and Summarization Using Probabilistic Generative Models

Hsin-Min Wang* and Berlin Chen†
*Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: whm@iis.sinica.edu.tw
†Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan
E-mail: berlin@csie.ntnu.edu.tw

***Abstract*—This paper presents our recent research work on applying probabilistic generative models to Mandarin Chinese broadcast news retrieval and summarization. Most models can be trained in either a supervised or unsupervised manner. In addition, both literal term matching and concept matching strategies have been intensively investigated. This paper also presents a prototype web-based Mandarin Chinese broadcast news retrieval system, which is based on technologies such as automatic story segmentation, automatic speech recognition, spoken document retrieval and summarization.**

I. INTRODUCTION

Nowadays multimedia content continues to grow and fill our computers, networks and daily lives. Since speech is one of the most important sources of information about this content, multimedia access based on associated spoken documents has attracted much research in recent years [1]. Substantial efforts and very encouraging results for spoken document transcription, retrieval, and summarization have been reported.

A standard approach to spoken document retrieval (SDR) is to automatically transcribe spoken documents into word (or subword) sequences, which can be matched against queries. The indexing terms can be either word or subword *N*-grams, or both. In vector space model (VSM) based spoken document retrieval, a document d (or a query q) is represented as a set of feature vectors, each consisting of TF-IDF (term frequency and inverse document frequency) information for one type of indexing term, and the similarity between the document and the query is measured as the cosine measures of their feature vectors. A probabilistic model based approach [2] ranks the documents according to the probability that document d is relevant given that query q is observed. These approaches are based on matching the terms, thus they often suffer from the problem of word usage diversity (or vocabulary mismatch) because the query and its relevant documents might use different words with similar meanings to describe the same thing. In contrast, the concept matching strategy tries to discover the latent topical information inherent in the query and documents. The latent semantic indexing (LSI) model [3] and the probabilistic latent semantic analysis (PLSA) model [4] are two good examples.

Spoken document summarization (SDS), which aims at distilling the important information and removing redundant and incorrect information from spoken documents, can help users to efficiently review the spoken documents and understand the associated topics quickly. 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 concatenates them to form a summary. In contrast, abstractive summarization produces a concise abstract of a certain length that reflects the key concepts of the document. The latter is more difficult to achieve, thus much research has focused on the former. For example, the vector space model (VSM) represents the whole document and each of its sentences in vector form consisting of the weighted statistics associated with indexing terms in the sentence or document. The sentences with the highest proximity scores (usually calculated as the cosine measure of two vectors) to the whole document are included in the summary. The latent semantic analysis (LSA) model for information retrieval (IR) can also be used to represent 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 [5]. In another example, each sentence in a document, represented as a sequence of terms, is given a significance score, which is evaluated using a weighted combination of statistical and linguistic measures. Sentences are then selected according to their significance scores [6]. More recently, a probabilistic generative framework has been applied in SDS in [7].

This paper presents our recent research work on applying probabilistic generative models to Mandarin Chinese broadcast news retrieval and summarization. Both literal term matching and concept matching strategies have been intensively investigated. The remainder of this paper is organized as follows. Considerations of using word and subword indexing features for Mandarin Chinese broadcast

This work was supported in part by Taiwan e-Learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under Grant: NSC98-2631-001-013.

09-0103310337©2009 APSIPA. All rights reserved.

news retrieval and summarization are discussed in Section II. The retrieval task and the summarization task are presented in Sections III and IV, respectively. A prototype web-based Mandarin Chinese broadcast news retrieval system is presented in Section V. Finally, conclusions are drawn in Section VI.

II. Considerations of Using Word- and Subword-Level Indexing Features

Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of spoken Mandarin, if the differences in tones are disregarded. In contrast, an inventory of about 13,000 characters provides full textual coverage of written Chinese[1]. Each word is composed of one or several characters, and each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are easily generated by combining a few characters. For example, the combination of the characters "電(electricity)" and "腦(brain)" yields the word "電腦(computer)" while the combination of "火(fire)" and "山(mountain)" yields the word "火山(volcano)". Examples of such new words also include many proper nouns, such as personal names, organization names, and domain specific terms. The construction of words from characters is very often quite flexible. One example phenomenon is that different words describing the same or similar concepts can be constructed by slightly different characters, e.g., both "中華文化(Chinese culture)" and "中國文化(Chinese culture)" mean the same, but the second characters in these two words are different. Another example phenomenon is that a longer word can be arbitrarily abbreviated to a shorter word, e.g., "國家科學委員會(National Science Council)" is often abbreviated to "國科會". In addition, there is a many-to-many mapping between characters and syllables. For example, the character "乾" may be pronounced as /gan1/ or /qian2/, while all the characters "甘", "干", "柑", "肝", "竿", "尷", and "疳" are also pronounced as /gan1/, and all the characters "前", "錢", "潛", "黔", "虔", and "捐" are pronounced as /qian2/. Consequently, a foreign word can be transliterated into different Chinese words based on its pronunciation. For example, Kosovo may be transliterated into "科索沃/ke1-suo3-wo4/", "科索佛/ke1-suo3-fo2/", "科索夫/ke1-suo3-fu1/", "科索伏/ke1-suo3-fu2/", "柯索佛/ke1-suo3-fo2/", etc., while Sarkozy may be transliterated into "薩科齊/sa4-ke1-chi2/", "薩科奇/sa4-ke1-chi2/", "薩爾科齊/sa4-er3-ke1-chi2/", etc. Different transliterations usually have some syllables in common, or may have exactly the same syllables.

The characteristics of the Chinese language lead to some special considerations when performing Mandarin Chinese speech recognition, e.g., syllable recognition is believed to be a key problem [8]. Recognition performance evaluation is usually based on syllable accuracy and character accuracy, rather than word accuracy. The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task [9][10]. Word-level indexing features possess more semantic information than subword-level features; thus, word-based retrieval enhances the precision. On the other hand, subword-level indexing features are more robust against the Chinese word tokenization ambiguity, Chinese homophone ambiguity, open vocabulary problem, and speech recognition errors; thus, subword-based retrieval enhances the recall. Accordingly, there is good reason to fuse the information obtained from indexing features of different levels. It has been shown [9] that syllable-level indexing features are very effective for Mandarin Chinese spoken document retrieval, and the retrieval performance can be improved by integrating information from character-level and word-level indexing features.

[1] According to the BIG5 character set. There are about 6,800 simplified Chinese characters in GB-2312 code.

III. Mandarin Chinese Broadcast News Retrieval

In probabilistic model based spoken document retrieval, the documents are ranked according to $p(d \text{ is relevant} \mid q)$, i.e., the probability that document d is relevant given that query q is observed. By Bayes' rule, $p(d \text{ is relevant} \mid q)$ can be expressed as

$$p(d \text{ is relevant} \mid q) = \frac{p(q \mid d \text{ is relevant})\, p(d \text{ is relevant})}{p(q)}, \quad (1)$$

where $p(q \mid d \text{ is relevant})$ is the probability of query q given that document d is relevant; $p(d \text{ is relevant})$ is the prior probability of document d being relevant; and $p(q)$ is the prior probability of query q. Note that $p(q)$, in (1), can be omitted because it is identical for all documents and will not affect their ranking. Moreover, $p(d \text{ is relevant})$ is usually assumed uniformly distributed because the way to estimate it is still an open issue. Therefore, in most practical implementations, the documents are ranked according to $p(q \mid d \text{ is relevant})$, which can be implemented by a probabilistic generative model. For notational convenience, $p(q \mid d \text{ is relevant})$ will be denoted by $p(q \mid d)$ in the following discussion.

A. Literal Term Matching

Each document d can be treated as a probabilistic generative model consisting of N-gram distributions for predicting a query q, while the query q is considered as an input observation sequence of indexing terms, i.e., $q=t_1 t_2 \ldots t_j \ldots t_J$, where t_j is the j-th indexing term in q and J is the length of q. When only the unigrams are considered, the relevance score for document d and query q can be expressed as

$$p_{LM}(q \mid d) = \prod_{j=1}^{J} \left[m_1 p(t_j \mid d) + m_2 p(t_j \mid C) \right], \quad (2)$$

where m_i is a mixture weight; $p(t_j \mid d)$ is the unigram probability of a specific indexing term t_j within document d; and $p(t_j \mid C)$ is the unigram probability estimated from a large text corpus related to the spoken document collection. For a broadcast news retrieval task, a newswire text corpus can be used for this purpose.

The N-gram probabilities for generating the query observations in a specific document and in the large text corpus can be obtained by maximum likelihood estimation (MLE). The weights m_i, which are tied among all the documents, can be optimized using the expectation-maximization (EM) algorithm, given a training set of query exemplars and their corresponding query-document relevance information. For example, m_1 can be estimated using the following equation:

$$m_1 = \frac{\sum_{q \in Q_{train}} \sum_{d \in D_{R \text{ to } q}} \sum_{t_j \in q} \left[\frac{\hat{m}_1 p(t_j \mid d)}{\hat{m}_1 p(t_j \mid d) + \hat{m}_2 p(t_j \mid C)} \right]}{\sum_{q \in Q_t} |q| \cdot |D_{R \text{ to } q}|}, \quad (3)$$

where $\hat{m}_1$ and $\hat{m}_2$ are the weights estimated in the previous iteration, Q_{train} is the set of training query exemplars, $D_{R \text{ to } q}$ is the set of documents that are relevant to a specific training query exemplar q, $|q|$ is the length of the query q, and $|D_{R \text{ to } q}|$ is the total number of documents relevant to the query q. If the training query set is not available, we can adopt online weight estimation. First, for a query, an initial retrieval is performed with the equal weights. After the initial retrieval, the top L retrieved documents are assumed relevant to the query; hence can be used to train the query specific weights. Finally, a second retrieval can be performed based on the newly estimated weights. The details of this method can be found in [11].

B. Concept Matching

The probabilistic latent semantic analysis (PLSA) [4] model introduces a set of latent topic variables, $\{T_k,\ k = 1,2,..,K\}$, to characterize the "term-document" co-occurrence relationships. A query q is again treated as an observation sequence of indexing terms, i.e., $q=t_1t_2\ldots t_j\ldots t_J$. The probability of a document d generating a term t_j is parameterized by

$$p(t_j \mid d) = \sum_{k=1}^{K} p(t_j \mid T_k) \cdot p(T_k \mid d). \quad (4)$$

The relevance score between query q and document d can then be expressed as:

$$p_{PLSA}(q \mid d) = \prod_{j=1}^{J} \left[\sum_{k=1}^{K} p(t_j \mid T_k) \cdot p(T_k \mid d) \right]. \quad (5)$$

Note that this relevance score is not obtained directly from the frequency of the respective query term t_j occurring in d, but instead through the frequency of t_j in the latent topic T_k as well as the likelihood that d generates the latent topic T_k. A query and a document thus may have a high relevance score even if they do not share any terms in common.

TABLE I
STATISTICS OF THE TDT-2 COLLECTION.

No. of spoken documents	2,265 stories, 46 hours of audio		
No. of distinct text queries	16 Xinhua text stories (Topics 20001~20096)		
	Min.	Max.	Mean
Document length (characters)	23	4841	287.1
Query length (characters)	183	2623	532.9
No. of relevant documents per query	2	95	29.3

TABLE II
RETRIEVAL RESULTS ACHIEVED BY THE LM METHOD.

Indexing features	LM-E	LM-S	LM-U
Word unigram	0.5300	0.5658	0.5748
Syllable unigram+bigram	0.5061	0.5307	0.5410

TABLE III
RETRIEVAL RESULTS ACHIEVED BY DIFFERENT METHODS.

LM-S	PLSA-S	PLSA-U	VSM	LSA
0.5658	0.6068	0.5707	0.5122	0.5362

The PLSA model is usually trained in an unsupervised way by maximizing the total log-likelihood L_T of the document collection D in terms of the unigram $p(t \mid d)$ of all terms t in D, using the EM algorithm:

$$L_T = \sum_{d \in D} \sum_{t \in d} n(t,d) \cdot \log p(t \mid d), \quad (6)$$

where $n(t,d)$ is the frequency count for the term t in the document d, and $p(t \mid d)$ is the probability obtained in (4). However, the PLSA model can also be trained in a supervised manner, given a training set of query exemplars and their corresponding query-document relevance information. The details of this method can be found in [12].

C. Experimental Results

Our experiments were performed on the Topic Detection and Tracking collection (TDT-2). The Chinese news stories (text) from Xinhua News Agency were used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. Table I describes the details for the corpus used in this paper. The speech recognition error rates for the spoken documents are about 35% (word), 18% (character), and 13% (syllable). 819 training queries with their corresponding query-document relevance information to the TDT-2 spoken document collection were used in supervised training of LM and PLSA models. The retrieval performance was evaluated in terms of mean average precision (mAP), defined as follows:

$$\text{mAP} = \frac{1}{|Q_{test}|} \sum_{i=1}^{|Q_{test}|} \frac{1}{|D_{R \text{ to } q_i}|} \sum_{k=1}^{|D_{R \text{ to } q_i}|} \frac{k}{rank_{ik}}, \quad (7)$$

where $|Q_{test}|$ is the total number of test queries; $|D_{R \text{ to } q_i}|$ is the total number of documents relevant to query q_i; and $rank_{ik}$ is the rank of the k-th relevant document for query q_i.

Table II shows the retrieval results of the LM method. LM-E adopted the equal weights. For LM-S, the weights were trained in a supervised manner with the 819 training queries while, for LM-U, the weights were trained online in an unsupervised manner with the top 20 ranked documents given by the initial retrieval. It is obvious that both LM-S and LM-U outperform LM-E. Table III shows the retrieval results of different retrieval approaches based on the word-level indexing features. For PLSA-U, the PLSA model was trained in an unsupervised manner while, for PLSA-S, the PLSA model was trained in a supervised manner with the 819 training queries. For both PLSA-S and PLSA-U, the number of latent topics was set to 8. From Table III, we observe that LM-S outperforms VSM and LSA and PLSA-S outperforms LM-S. The detailed experimental results can be found in [11-12].

IV. Mandarin Chinese Broadcast News Summarization

Extractive spoken document summarization can also be performed with probabilistic generative models. The importance of a sentence *s* in a document *d* to be summarized can be modeled by $p(s|d)$; i.e., the posterior probability of the sentence *s* given the document *d*. According to Bayes' rule, $p(s|d)$ can be expressed as:

$$p(s \mid d) = \frac{p(d \mid s)P(s)}{p(d)}, \quad (8)$$

where $p(d|s)$ is the sentence generative probability, i.e., the probability of *d* being generated by *s*; $p(s)$ is the prior probability of *s* being important; and $p(d)$ is the prior probability of *d*. Note that $p(d)$, in (8), can be omitted because it is identical for all sentences and will not affect their ranking. The sentence generative probability $p(d|s)$ can be taken as a relevance measure between the document *d* and the sentence *s*, while the sentence prior probability $p(s)$ is a measure of the importance of the sentence itself. Therefore, all the sentences of the spoken document *d* can be ranked according to the product of the sentence generative probability $p(d|s)$ and the sentence prior probability $p(s)$. Then, the sentences with the highest probabilities are selected and sequenced to form a summary.

A. Sentence Generative Probability for Literal Term Matching

A language model (LM) can be applied in extractive spoken document summarization, where each sentence *s* of a document *d* to be summarized is treated as a probabilistic generative model comprised of *N*-gram distributions for predicting the document *d*; and the indexing terms in *d* are taken as an input observation sequence. When only the unigrams are considered, the probability of the document *d* given the sentence *s* is expressed as:

$$p_{LM}(d \mid s) = \prod_{t \in d} \left[\lambda \cdot p(t \mid s) + (1-\lambda) \cdot p(t \mid C)\right]^{n(t,d)}, \quad (9)$$

where λ is a weighting parameter and $n(t,d)$ is the occurrence count of the term *t* in *d*. The sentence model $p(t|s)$ and the collection model $p(t|C)$ are estimated, respectively, from the sentence *s* itself and a large external text collection *C* using the MLE method. The weighting parameter λ can be empirically tuned by using a development data set, or optimized by applying the EM training algorithm to a training data set. Since the relevance measure is computed according to the frequency that document terms occur in the sentence, it is obvious that this LM method adopts literal term matching.

The sentence model $p(t|s)$ might be poorly estimated because the sentence *s* only consists of a few terms. We can employ the relevance model (RM) [13] to obtain a more accurate estimation of the sentence model. In the extractive spoken document summarization task, each sentence *s* of a document *d* has its own associated relevant class R_s, which is defined as the subset of documents in the collection that are relevant to *s*. The relevance model of *s* is defined as the probability distribution $p(t|R_s)$, which gives the probability that we would observe a term *t* if we were to randomly select a document from the relevant class R_s and select a word from that document. After the relevance model of *s* has been constructed, it can be used to replace the original sentence model or it can be combined linearly with the original sentence model. Because we do not have prior knowledge about the subset of relevant documents for each spoken sentence *s*, we employ a relevance feedback procedure that submits *s* as a query to an IR system to obtain a ranked list of documents for sentence generative model expansion. It is assumed that the top *L* documents returned by the IR system are relevant to *s*, and the relevance model $p(t|R_s)$ of *s* can be constructed by the following equation:

$$p(t \mid R_s) = \sum_{d_l \in D_{topL}} p(d_l \mid s) \cdot p(t \mid d_l), \quad (10)$$

where D_{topL} is the set of top *L* retrieved documents, and the probability $p(d_l|s)$ can be approximated by the following equation using Bayes' rule:

$$p(d_l \mid s) \approx \frac{p(d_l) \cdot p(s \mid d_l)}{\sum_{d_u \in D_{topL}} p(d_u) \cdot p(s \mid d_u)}. \quad (11)$$

A uniform prior probability $p(d_l)$ can be assumed for the top *L* retrieved documents, and $p(s|d_l)$ is given by the probabilistic model based IR system. The relevance model $p(t|R_s)$ can then be combined linearly with the original sentence model $p(t|s)$ to form a more accurate sentence model

$$\hat{p}(t \mid s) = \alpha \cdot p(t \mid s) + (1-\alpha) \cdot p(t \mid R_s), \quad (12)$$

where α is a weighting parameter. The sentence generative model can be thus expressed as

$$p_{LM-RM}(d \mid s) = \prod_{t \in d} \left[\lambda \cdot \hat{p}(t \mid s) + (1-\lambda) \cdot p(t \mid C)\right]^{n(t,d)}. \quad (13)$$

TABLE IV
THE FEATURES EXPLOITED FOR MODELING THE SENTENCE PRIORITY PROBABILITY.

Lexical Features	average TF-ICF score of words in a spoken sentence (F1) average bi-gram scores of word pairs in a spoken sentence (F2)
Prosodic Features	average pitch value of words in a spoken sentence (F3) average energy value of words in a spoken sentence (F4) maximum energy of words in a spoken sentence (F5)
Confidence Feature	average posterior probability of words in a spoken sentence (F6)
Relevance Feature	average similarity among the retrieved text documents for a spoken sentence (F7)

The weighting parameter λ in (9) can also be estimated with the retrieved relevant text document set D_{topL}, using the following EM updating equation:

$$\hat{\lambda} = \frac{\sum_{d_l \in D_{topL}} \sum_{t \in d_l} n(t,d_l) \cdot \frac{\lambda \cdot p(t \mid s)}{\lambda \cdot p(t \mid s) + (1-\lambda) \cdot p(t \mid C)}}{\sum_{d_l \in D_{topL}} \sum_{t \in d_l} n(t,d_l)}. \quad (14)$$

We denote this model as LM-RT.

B. Sentence Generative Probability for Concept Matching

In the concept matching mode, each sentence s of a document d can be interpreted as a probabilistic sentence topic model (STM). Then, the sentence generative probability can be expressed as

$$p_{STM}(d \mid s) = \prod_{t \in d} \left[\sum_{k=1}^{K} p(t \mid T_k) p(T_k \mid s) \right]^{n(t,d)}, \quad (15)$$

where $p(t \mid T_k)$ and $p(T_k \mid s)$ denote, respectively, the probability of the term t 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, $p(t \mid T_k)$, k=1,…, K, are the same for all sentences, but each sentence s has its own probability distributions over the latent topics, i.e., $p(T_k \mid s)$, k=1,…, K. Note that this relevance measure is not computed directly according to the frequency that the document terms occur in the sentence. Instead, it is derived from the frequency of the document terms in the latent topics as well as the likelihood that the sentence will generate the respective topics.

During training, a set of contemporaneous (or in-domain) text news documents D_{text} 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 $p(t \mid T_k)$. First, the K-means algorithm is used to partition all the titles of the documents in D_{text} into K topical clusters in an unsupervised manner. Then, the initial topical unigram distribution $p(t \mid T_k)$ is estimated from the document titles assigned to topic T_k. In addition, the probability that each title h will generate topic T_k, i.e., $p(T_k \mid h)$, is measured according to the proximity of h to the centroid of the k-th topical cluster. Then, using the EM algorithm, the probability distributions $p(t \mid T_k)$ and $p(T_k \mid h)$ can be optimized by maximizing the total log-likelihood L_T of all the documents in D_{text} generated by their individual titles:

$$\begin{aligned} L_T &= \sum_{d \in D_{text}} \log p(d \mid h_d) \\ &= \sum_{d \in D_{text}} \log \prod_{t \in d} \left[\sum_{k=1}^{K} p(t \mid T_k) p(T_k \mid h_d) \right]^{n(t,d)}. \end{aligned} \quad (16)$$

We postulate that latent topical factors $p(t \mid T_k)$ properly constructed based on "document-title" relationships might provide very helpful clues for the subsequent spoken document summarization task. When performing extractive summarization of a broadcast news document d, we can apply the latent topical factors $p(t \mid T_k)$ trained in this way in (15), but use the EM algorithm to estimate the posterior probabilities, $p(T_k \mid s)$, k=1,…, K, on the fly by maximizing the log-likelihood of the document d generated by the STM model. A detailed account of the process can be found in [14].

If the contemporaneous or in-domain text documents are not accompanied by "document-title" pairs for model training, we can also use unsupervised training by exploiting all the sentences of the spoken (broadcast news) documents in the development set to construct the latent topical space [14]. That is, each sentence of a spoken document in the development set, regardless of whether it belongs to the reference summary or not, is treated as an STM model and included in the construction of the latent topical distributions $p(t \mid T_k)$. Meanwhile, the probability distribution $p(T_k \mid s)$ is estimated online during the summarization process. We denote this model as STM-U.

C. Sentence Prior Probability

In the probabilistic generative framework for extractive spoken document summarization, the sentence prior probability in (8), which can be regarded as the probability of a sentence being important in the document, is usually assumed uniformly distributed. However, the sentences in a spoken document should not be considered equally important. In fact, a sentence's importance may depend on a wide variety of factors, such as the structural (positional and lexical) information, recognition accuracy, and inherent prosodic properties. Therefore, we model the sentence prior probability (or importance) based on lexical, prosodic, and confidence features extracted from a spoken sentence. These features are presented in Table IV. The TF-ICF score is similar to the conventional TF-IDF measure widely used in IR systems, but the value of ICF (Inverse Collection Frequency) is calculated by [15]:

$$ICF = \log \frac{F_A}{F_t}, \quad (17)$$

where F_t is the occurrence count of a term t in a large contemporaneous text corpus, and F_A is the number of terms in the corpus. In addition, the prosodic features are extracted from the broadcast news speech by using the Snack toolkit [16]. The measure or score of each feature in Table IV is normalized such that it can be taken as the sentence prior probability that satisfies $\sum_{s \in d} p(s) = 1$.

We also model the sentence prior probability by calculating the average similarity of documents in the retrieved text document set D_{topL} for sentence s (cf. Section IV-A). Our assumption is that the relevant text documents retrieved for a summary sentence might have the same or similar topics because a summary sentence is usually indicative for some specific topic related to the document. In contrast, the relevant text documents retrieved for a non-summary sentence might cover diverse topics. Therefore, the relevance information estimated based on the similarity of documents in the retrieved text document set D_{topL} might be a good indicator for determining the importance of a spoken sentence. Consequently, the sentence prior probability can be approximated by using the sentence's relevance information as follows:

$$p(s) = \frac{avgSim(s)}{\sum_{s' \in d} avgSim(s')}, \quad (18)$$

where $avgSim(s)$ is the average similarity of documents in the retrieved text document set D_{topL} for a spoken sentence s computed by

$$avgSim(s) = \frac{\sum_{d_l \in D_{topL}} \sum_{\substack{d_u \in D_{topL} \\ d_l \neq d_u}} \frac{\vec{d}_l \cdot \vec{d}_u}{\|\vec{d}_l\| \cdot \|\vec{d}_u\|}}{L \cdot (L-1)}, \quad (19)$$

where $\vec{d}_l$ is the TF-IDF vector representation of the document d_l, and L is the number of documents in the retrieved relevant text document set D_{topL}.

D. Experimental Results

Our experiments were performed on 200 broadcast news stories, the first 100 stories were used as the development set for tuning the parameters, and the remaining 100 stories were used as the evaluation set. Automatic summarization was based on the best scoring sequence of words generated by the speech recognizer. A pause with duration more than 0.5 seconds was regarded as a sentence boundary. The average character error rate for the 200 news stories was 14.7%. A set of approximately 14,000 text news stories was used to estimate the collection model $p(t|C)$ for LM, LM-RM, and LM-RT and the latent topical distributions $p(t|T_k)$ for STM. It was also used to construct the retrieved text document set for each spoken sentence (cf. Section IV-A).

The summarization results were tested by using several summarization ratios (10%, 20%, and 30%), defined as the ratio of the number of sentences in the automatic summary to that in the reference transcript of a spoken document. We used the ROUGE package (Version 1.5.5) [17] to evaluate the performance. ROUGE-N is an N-gram recall measure, defined as follows:

$$\text{ROUGE} - N = \frac{\sum_{s_{ref} \in S_{ref}} \sum_{gram_N \in s_{ref}} Count_{match}(gram_N)}{\sum_{s_{ref} \in S_{ref}} \sum_{gram_N \in s_{ref}} Count(gram_N)}, \quad (20)$$

where s_{ref} is an individual manual (or reference) summary; S_{ref} is a set of manual summaries; $Count_{match}(gram_N)$ is the maximum number of N-grams co-occurring in the automatic summary and the manual summary; and $Count(gram_N)$ is the number of N-grams in the manual summary. In the experiments, each document had 3 manual summaries created by 3 subjects, and the ROUGE-2 measure was used.

In the first experiment, the sentence prior probability $p(s)$ was assumed to be uniform. The results are shown in Table V. It is clear that LM-RT and STM are the best among all the methods compared in this paper. In the second experiment, the LM-RT model was further integrated with the sentence prior probabilities derived by different features in Table IV. The results are shown in Table VI. Comparing these results with those in Table V, we observe that the performance at lower summarization ratios (10% and 20%) is in general improved by incorporating the sentence prior probability. The relevance feature F7, which is the average similarity among the top L retrieved text documents for a spoken sentence, is most helpful. L was set to 5 in the experiment. The details of our probabilistic generative framework and more experimental results and discussions can be found in [7].

TABLE V
SUMMARIZATION RESULTS ACHIEVED BY DIFFERENT MODELS, USING A UNIFORM SENTENCE PRIOR PROBABILITY.

	LM	LM-RM	LM-RT	STM	STM-U	VSM	LSA
10%	0.2932	0.3182	0.3316	0.3210	0.3016	0.3073	0.3034
20%	0.3191	0.3264	0.3412	0.3333	0.3217	0.3188	0.2926
30%	0.3705	0.3671	0.3739	0.3741	0.3618	0.3593	0.3286

TABLE VI
SUMMARIZATION RESULTS ACHIEVED BY LM-RT, WITH THE SENTENCE PRIOR PROBABILITY MODELED BY USING DIFFERENT FEATURES.

	F1	F2	F3	F4	F5	F6	F7
10%	0.3394	0.3408	0.3414	0.3347	0.3347	0.3203	0.3589
20%	0.3549	0.3448	0.3422	0.3443	0.3443	0.3430	0.3690
30%	0.3765	0.3696	0.3742	0.3622	0.3739	0.3765	0.3858

V. A PROTOTYPE MANDARIN CHINESE BROADCAST NEWS RETRIEVAL SYSTEM

We have implemented a web-based Mandarin Chinese broadcast news retrieval system, called SoVideo (http://sovideo.iis.sinica.edu.tw). The database consists of more than 400 hours of broadcast news, which yields 10,343 stories by automatic story segmentation. As depicted in Fig. 1, SoVideo allows users to input search terms to search for their desired news stories from the broadcast news database.

We adopted a simple but effective multi-pass approach for automatic story segmentation. The first pass performs speaker

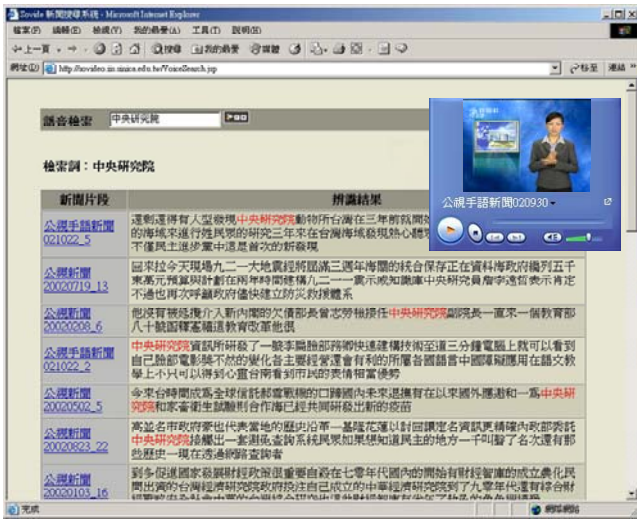

Fig. 1 The SoVideo web-based Mandarin Chinese broadcast news retrieval system.

and environment change detection. The second pass conducts hierarchical agglomerative clustering of audio segments. We assume that the largest cluster is the anchor reporter cluster, and every anchor speech segment is the first segment of a story. In this way, the number of anchor reporter segments corresponds to the number of stories in the audio stream. The details of our story segmentation approach can be found in [18]. The results of experiments on 5 one-hour news shows demonstrate that it works very well. There were 112 stories by hand segmentation. Automatic story segmentation resulted in 114 stories. Two detected stories were false alarms, i.e., two stories were respectively divided into two. The starting time errors of 108 detected stories were within 3 seconds, while the errors of the remaining 4 detected stories were 5, 6, 9, and 17 seconds, respectively.

VI. Conclusions

We have been working on Mandarin Chinese broadcast news transcription, retrieval, and summarization for about a decade. This paper has summarized our works on retrieval and summarization using the probabilistic generative models. Both the literal term matching and concept matching strategies have been applied in our probabilistic generative framework. This paper has also introduced a prototype web-based Mandarin Chinese broadcast news retrieval system. Currently, we are still improving the probabilistic generative models for spoken document retrieval and summarization.

References

[1] L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 22(5), pp. 42-60, 2005.

[2] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proc. *ACM SIGIR 1998*.

[3] G. W. Furnas, et al., "Information retrieval using a singular value decomposition model of latent semantic structure," in Proc. *ACM SIGIR 1988*.

[4] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. *ACM SIGIR 1999*.

[5] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in Proc. *ACM SIGIR 2001*.

[6] S. Furui, et al., "Speech-to-text and speech-to-speech summarization of spontaneous speech", *IEEE Trans. Speech and Audio Processing*, 12(4), pp. 401-408, 2004.

[7] Y. T. Chen, B. Chen, and H. M. Wang, "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Trans. on Audio, Speech, and Language Processing*, 17(1), pp. 95-106, 2009.

[8] L. S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 14(4), pp. 63-101, 1997.

[9] B. Chen, H. M. Wang, and L. S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," *IEEE Trans. on Speech and Audio Processing*, 10(5), pp.303-314, 2002.

[10] H. Meng, et al., "Mandarin-English Information (MEI): investigating translingual speech retrieval," *Computer Speech and Language*, 18(2), pp. 163-179, 2004.

[11] B. Chen, H. M. Wang, and L. S. Lee, "A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents," *ACM Trans. on Asian Language Information Processing*, 3(2), pp. 128-145, 2004.

[12] B. Chen, et al., "Statistical Chinese spoken document retrieval using latent topical information," in Proc. *ICSLP 2004*.

[13] W. B. Croft and J. Lafferty (Eds.), Language Modeling for Information Retrieval, Kluwer-Academic Publishers, 2003.

[14] B. Chen, et al., "Chinese spoken document summarization using probabilistic latent topical information," in Proc. *ICASSP 2006*.

[15] M. Hirohata, et al., "Sentence-extractive automatic speech summarization and evaluation techniques," *Speech Communication*, 48(9), pp. 1151-1161, 2006.

[16] Snack Sound Toolkit. Available from: <http://www.speech.kth.se/snack/>.

[17] C. Y. Lin, "ROUGE: recall-oriented understudy for gisting evaluation," 2003. Available from: < http://haydn.isi.edu/ROUGE/>.

[18] H. M. Wang, S. S. Cheng, and Y. C. Chen, "The SoVideo Mandarin Chinese broadcast news retrieval system," *International Journal of Speech Technology*, 7(2-3), pp. 189-202, 2004.