Word Topic Models for Spoken Document Retrieval and Transcription

BERLIN CHEN
National Taiwan Normal University

Statistical language modeling (LM), which aims to capture the regularities in human natural language and quantify the acceptability of a given word sequence, has long been an interesting yet challenging research topic in the speech and language processing community. It also has been introduced to information retrieval (IR) problems, and provided an effective and theoretically attractive probabilistic framework for building IR systems. In this article, we propose a word topic model (WTM) to explore the co-occurrence relationship between words, as well as the long-span latent topical information, for language modeling in spoken document retrieval and transcription. The document or the search history as a whole is modeled as a composite WTM model for generating a newly observed word. The underlying characteristics and different kinds of model structures are extensively investigated, while the performance of WTM is thoroughly analyzed and verified by comparison with the well-known probabilistic latent semantic analysis (PLSA) model as well as the other models. The IR experiments are performed on the TDT Chinese collections (TDT-2 and TDT-3), while the large vocabulary continuous speech recognition (LVCSR) experiments are conducted on the Mandarin broadcast news collected in Taiwan. Experimental results seem to indicate that WTM is a promising alternative to the existing models.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Language model, information retrieval, speech recognition, word topic model, adaptation

ACM Reference Format:
Chen, B. 2009. Word topic models for spoken document retrieval and transcription. ACM Trans. Asian Lang. Inform. Process. 8, 1, Article 2 (March 2009), 27 pages.
DOI = 10.1145/1482343.1482345. http://doi.acm.org/10.1145/1482343.1482345.

This work was supported in part by the National Science Council of Taiwan under grants NSC95-2221-E-003-014-MY3, NSC96-2628-E-003-015-MY3, and NSC97-2631-S-003-003.
Author's address: B. Chen, Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan; email: berlin@csie.ntnu.edu.tw.

1. INTRODUCTION

Statistical language modeling (LM), which aims to capture the regularities in human natural language and quantify the acceptability of a given word sequence, has continuously been a focus of active research in the speech and language processing community over the past three decades. For example, the n-gram modeling (especially the bigram and trigram modeling) approach, which determines the probability of a word given the preceding n-1 word history, is most prominently used [Jelinek and Mercer 1980; Rosenfeld 2000; Bellegarda 2004]. This statistical paradigm was first introduced for the information retrieval (IR) problems by Ponte and Croft [1998], Song and Croft [1999], and Miller et al. [1999], indicating very good potential, and was then extended in a number of publications [Berger and Lafferty 1999; Hoffmann 1999; Lafferty and Zhai 2001; Chen et al. 2004b]. In these approaches, the relevance measure between a query Q and a document D is expressed as $P(D|Q)$; that is, the probability that D is relevant given that the query Q is posed. Based on the Bayes theorem and some assumptions, this relevance measure can be approximated by $P(Q|D)$, or the probability of the query Q being posed, under the hypothesis that document D is relevant. Documents in the collection therefore can be ranked on the basis of this relevance measure.

In practice, the relevance measure $P(Q|D)$ is usually computed by two different matching strategies, namely, literal term matching and concept matching [Lee and Chen 2005]. The n-gram-based [Ponte and Croft 1998; Song and Croft 1999] and hidden Markov model (HMM)-based [Miller et al. 1999; Chen et al. 2004b] approaches are the most popular examples for literal term matching. In these approaches, each document is interpreted as a generative model composed of a mixture of n-gram probability distributions for observing a query, while the query is considered as observations, expressed as a sequence of indexing terms (or words). However, most of these approaches often suffer from the problem of word usage diversity (or so-called vocabulary mismatch), which will make the retrieval performance degrade severely as a given query and its relevant documents are using quite a different set of words. In contrast, concept matching tries to discover the latent topical information inherent in the query and documents, based on which the retrieval is performed; the probabilistic latent semantic analysis (PLSA) [Hoffmann 1999], referred to as the document topic model (DTM) in this article, is often considered as a representative of this category. PLSA introduces a set of latent topic variables to characterize the "word-document" co-occurrence relationships: typically, documents are mixtures of topics, where each topic has a probability distribution over words of the language. Therefore, the relevance measure of a query and a document is not computed directly based on the frequency of the query words occurring in a document, but instead based on the frequency of these words in the latent topics as well as the likelihood that the document generates the respective topics, which in fact exhibits some sort of concept matching. PLSA is usually trained in an unsupervised way [Hoffmann 2001] by maximizing the total log-likelihood of the document collection. Excellent survey articles of using statistical language modeling approaches for information retrieval can

be found in Croft and Lafferty [2003], Allan et al. [2003], Liu and Croft [2005], and Steyvers and Griffiths [2007].

With the rapid growth of accessible audio-visual content over the Internet, large volumes of real-world spoken documents, such as broadcast radio and television programs, digital libraries and so on, are now being accumulated and made available to the public. Substantial efforts and very encouraging results on speech recognition of spoken documents have been reported in the last few years [Woodland 2002; Byrne et al. 2004]. However, for complicated speech recognition tasks such as broadcast news transcription, it is still extremely difficult to build well-estimated language models, because the subject matters and lexical characteristics for the linguistic contents of news articles are very diverse and are often changing with time. Various attempts have been made to adapt the language model by making use of either the contemporaneous corpus or the recognition hypotheses observed so far [Jelinek et al. 1991; Federico and Bertoldi 2001]. Two of the most widespread approaches to language model adaptation are count merging and model interpolation, each of which can be respectively viewed as a maximum a posteriori (MAP) language model adaptation with a different parameterization of the prior distribution, and can be easily integrated into the *n*-gram modeling framework to capture the local regularities of word usage in the new task domain [Bacchiani and Roark 2003]. In the recent past, the probabilistic latent topic modeling approaches, which were originally formulated in IR, have been introduced to dynamic language model adaptation and investigated to complement the *n*-gram models as well. Among them, PLSA [Gildea and Hoffmann 1999] has been widely studied and demonstrated effective in a few speech recognition tasks. However, it merely targets on maximizing the collection likelihood but not directly on its language model prediction capability, and it also suffers from the problem that part of its model parameters have to be dynamically estimated on the fly during the speech recognition process, which would be time-consuming and makes it impractical for realistic speech recognition applications. Interested readers may refer to Bellegarda [2004] for a comprehensive overview of the major language model adaptation approaches that have been successfully developed and applied to speech recognition.

Based on these observations, in this article a word topic model (WTM) is proposed to explore the co-occurrence relationship between words, as well as the underlying global topic information, for language modeling in spoken document retrieval and transcription. Each word of the language is treated as a WTM model for predicting the occurrences of the other words, while the corresponding model parameters can be estimated by leveraging the vicinity information of all occurrences of the word in the document collection. The document or the search history as a whole thus is represented as a composite WTM model for generating the newly observed word. The underlying characteristics and different kinds of model structures are investigated, while the performance of WTM is analyzed and compared with PLSA and the other models.

The rest of this article is organized as follows. Section 2 sheds light on the structural characteristics of the word topic model and its applications to information retrieval and dynamic language model adaptation, as well as its

comparison to the probabilistic latent semantic analysis. Section 3 describes the speech and language corpora used in this article, as well as the experimental setup. The experiments on spoken document retrieval and transcription are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this article with future work.

2. DOCUMENT AND WORD TOPIC MODELS

In this section, the probabilistic latent semantic analysis (PLSA), namely the document topic model (DTM), is reviewed, followed by an introduction of our proposed word topic model (WTM). We will concentrate on elucidating and comparing the structural characteristics of PLSA and WTM, as well as their applications to information retrieval and language model adaptation.

2.1 PLSA for Information Retrieval

In information retrieval (IR), the relevance measure between a query Q and a document D can be expressed as $P(D|Q)$, that is, the probability that the document D is relevant given that the query Q was posed, which can be transformed to the following equation by applying the Bayes theorem:

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

where $P(Q|D)$ is the probability of the query Q being generated by the document D, $P(D)$ is the prior probability of document D being relevant, and $P(Q)$ is the prior probability of query Q being posed. $P(Q)$ in Equation (1) can be eliminated because it is identical for all documents and will not affect the ranking of the documents. Furthermore, because the way to estimate the probability $P(D)$ is still under active study [Miller et al. 1999; Liu and Croft 2005], we may simply assume that $P(D)$ is uniformly distributed, or identical for all documents. In this way, the documents can be ranked by means of the probability $P(Q|D)$ instead of using the probability $P(D|Q)$. If the query Q is treated as a sequence of input observations (words or terms), $Q = w_1 w_2 \ldots w_N$, where the query words are assumed to be conditionally independent given the document D and their order is also assumed to be of no importance (i.e., the so-called "bag-of-words" assumption), the relevance measure $P(Q|D)$ can be decomposed as a product of the probabilities of the query words generated by the document:

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

where $c(w_i, Q)$ is the number of times that each distinct word w_i occurs in Q.

In the PLSA modeling approach for IR, each individual document D can be interpreted as a generative document topic model (DTM), denoted as M_D, in which a set of K latent topics characterized with unigram (or multinomial) distributions are used to predict the query terms, and each of the latent topics is associated with a document-specific weight. That is, each document D can

belong to many topics and the probability of a query word w_i generated by D is expressed by

$$P_{\mathrm{PLSA}}(w_i|\mathrm{M}_D) = \sum_{k=1}^{K} P(w_i|T_k)\,P(T_k|\mathrm{M}_D), \tag{3}$$

where $P(w_i|T_k)$ denotes the probability of a query word w_i occurring in a specific latent topic T_k, and $P(T_k|\mathrm{M}_D)$ is the posterior probability (or weight) of the topic T_k conditioned on the document model M_D, with the constraint $\sum_{k=1}^{K} P(T_k|\mathrm{M}_D) = 1$ imposed. More precisely, the topical unigram distributions, e.g., $P(w_i|T_k)$, are shared among the entire DTM models, while each model M_D has its own probability distribution over the latent topics, for example, $P(T_k|\mathrm{M}_D)$. The key idea we wish to illustrate here is that the relevance measure of a query word w_i and a document D is not computed directly based on the frequency of w_i occurring in D, but instead based on the frequency of w_i in the latent topic T_k as well as the likelihood that D generates the respective topic T_k, which in fact exhibits some sort of concept matching [Lee and Chen 2005]. The likelihood of a query Q generated by D is thus represented by

$$P_{\mathrm{PLSA}}(Q|\mathrm{M}_D) = \prod_{w_i \in Q} \left[\sum_{k=1}^{K} P(w_i|T_k)\,P(T_k|\mathrm{M}_D) \right]^{c(w_i,Q)}. \tag{4}$$

In the practical implementation of PLSA [Hoffmann 2001], the corresponding DTM models are usually trained in an unsupervised way by maximizing the total log-likelihood of the document collection $\mathbf{D}$ in terms of the unigram $P_{\mathrm{PLSA}}(w_i|\mathrm{M}_D)$ of all words w_i observed in the document collection, or more specifically, the total log-likelihood of all documents generated by their own DTM models:

$$\begin{aligned} \log L_{\mathbf{D}} &= \sum_{D \in \mathbf{D}} \log P_{\mathrm{PLSA}}(D|\mathrm{M}_D) \\ &= \sum_{D \in \mathbf{D}} \sum_{w_i \in D} c(w_i,D) \log P_{\mathrm{PLSA}}(w_i|\mathrm{M}_D). \end{aligned} \tag{5}$$

The optimization of Equation (5) can be conducted iteratively via the following three expectation-maximization (EM) update equations [Dempster et al. 1977; Rabiner 1989]:

$$\hat{P}(w_i|T_k) = \frac{\sum_{D \in \mathbf{D}} c(w_i,D)\,P(T_k|w_i,\mathrm{M}_D)}{\sum_{D \in \mathbf{D}} \sum_{w_s \in D} c(w_s,D)\,P(T_k|w_s,\mathrm{M}_D)}, \tag{6}$$

$$\hat{P}(T_k|M_D) = \frac{\sum_{w_i \in D} c(w_i,D)\,P(T_k|w_i,\mathrm{M}_D)}{\sum_{w_s \in D} c(w_s,D)}, \tag{7}$$

$$P(T_k|w_i,\mathrm{M}_D) = \frac{P(w_i|T_k)\,P(T_k|\mathrm{M}_D)}{\sum_{l=1}^{K} P(w_i|T_l)\,P(T_l|\mathrm{M}_D)}, \tag{8}$$

where $P(T_k|w_i, \mathrm{M}_D)$ is the probability that the latent topic T_k occurs given the query word w_i and the document model M_D, which is computed using the probability quantities $P(w_i|T_k)$ and $P(T_k|\mathrm{M}_D)$ obtained in the previous training iteration.

In addition to conventional unsupervised training of PLSA, we propose supervised training of PLSA (or the corresponding DTM models) in this article. Given a training set of query exemplars $\mathbf{Q}_{TrainSet}$ and their associated query-document relevance information, the DTM models can be optimized by instead finding the model parameters that can maximize the total log-likelihood of the training set of query exemplars $\mathbf{Q}_{TrainSet}$ generated by their relevant documents:

$$\begin{aligned}\log L_{\mathbf{Q}_{TrainSet}} &= \sum_{Q\in\mathbf{Q}_{TrainSet}} \sum_{D\in\mathbf{D}_{R\,\mathrm{to}\,Q}} \log P_{\mathrm{PLSA}}(Q|\mathrm{M}_D) \\ &= \sum_{Q\in\mathbf{Q}_{TrainSet}} \sum_{D\in\mathbf{D}_{R\,\mathrm{to}\,Q}} \sum_{w_i\in Q} c(w_i, Q)\log P_{\mathrm{PLSA}}(w_i|\mathrm{M}_D),\end{aligned} \tag{9}$$

where $\mathbf{D}_{R\ \mathrm{to}\ Q}$ denotes the set of documents that are relevant to a specific training query exemplar Q. This leads to the following two EM update equations:

$$\hat{P}(w_i|T_k) = \frac{\sum_{Q\in\mathbf{Q}_{TrainSet}} \sum_{D\in\mathbf{D}_{R\,\mathrm{to}\,Q}} c(w_i, Q)\,P(T_k|w_i, \mathrm{M}_D)}{\sum_{Q\in\mathbf{Q}_{TrainSet}} \sum_{D\in\mathbf{D}_{R\,\mathrm{to}\,Q}} \sum_{w_l\in Q} c(w_l, Q)\,P(T_k|w_l, \mathrm{M}_D)}, \tag{10}$$

$$\hat{P}(T_k|\mathrm{M}_D) = \frac{\sum_{\substack{Q\in\mathbf{Q}_{TrainSet}\\ \text{st. } D\in\mathbf{D}_{R\,\mathrm{to}\,Q}}} \sum_{w_s\in Q} c(w_s, Q)\,P(T_k|w_s, \mathrm{M}_D)}{\sum_{\substack{Q\in\mathbf{Q}_{TrainSet}\\ \text{st. } D\in\mathbf{D}_{R\,\mathrm{to}\,Q}}} \sum_{w_l\in Q} c(w_l, Q)}, \tag{11}$$

where $Q \in \mathbf{Q}_{TrainSet}$ st. $D \in \mathbf{D}_{R\,\mathrm{to}\,Q}$ means that the query exemplar Q in the training query set can satisfy the condition that the document D in the collection is relevant to it; and $P(T_k|w_s, \mathrm{M}_D)$ can be computed using Equation (8).

2.2 WTM for Information Retrieval

In this article, we present an alternative probabilistic latent topic approach for information retrieval. Instead of treating each document in the collection as a document topic model, we regard each word w_j of the language as a word topic model (WTM) M_{w_j} for predicting the occurrences of a particular word w_i:

$$P_{\mathrm{WTM}}(w_i|\mathrm{M}_{w_j}) = \sum_{k=1}^{K} P(w_i|T_k)\,P(T_k|\mathrm{M}_{w_j}). \tag{12}$$

where $P(w_i|T_k)$ and $P(T_k|\mathrm{M}_{w_j})$ are respectively the probability of a word w_i occurring in a specific latent topic T_k and the probability of the topic T_k

conditioned on M_{w_j}. During the retrieval process, we can linearly combine the associated WTM models of the words involved in a document D to form a composite WTM model for D, and the likelihood of a query Q being generated by D can be expressed by

$$P_{\mathrm{WTM}}(Q\,|\mathrm{M}_D) = \prod_{w_i \in Q} \left[\sum_{w_j \in D} \alpha_{j,D} \sum_{k=1}^{K} P(w_i\,|T_k)\, P\left(T_k\,\middle|\mathrm{M}_{w_j}\right) \right]^{c(w_i, Q)}, \tag{13}$$

where the weighting coefficient $\alpha_{j,D}$ is set to be in proportion to the frequency of w_j occurring in D and summed to 1 ($\sum_{w_j \in D} \alpha_{j,D} = 1$). In this way, the relevance measure between a query and document is determined by the product of a weighted sum of the probabilities that the respective WTM models of the words involved in the document generating each query word, and the documents having the highest probabilities expressed by Equation (13) are therefore believed to be more relevant to Q.

The WTM models can also be optimized by the EM algorithm either with supervision or without supervision. For unsupervised training of WTM, each WTM model M_{w_j} can be trained by concatenating those words occurring within a vicinity of, or a context window of size S (S is experimentally set to 21 in this study) around, each occurrence of w_j, which are postulated to be relevant to w_j, in the spoken documents collection to form a relevant observation sequence Q_{w_j} for training M_{w_j}. The words in Q_{w_j} are also assumed to be conditionally independent given M_{w_j}. Therefore, the WTM models of the words in the vocabulary set $\mathbf{w}$ can be estimated by maximizing the total log-likelihood of their corresponding relevant observation sequences respectively generated by themselves:

$$\begin{aligned} \log L_{\mathbf{w}} &= \sum_{w_j \in \mathbf{w}} \log P_{\mathrm{WTM}}\left(Q_{w_j}\,\middle|\mathrm{M}_{w_j}\right) \\ &= \sum_{w_j \in \mathbf{w}} \sum_{w_i \in Q_{w_j}} c\left(w_i, Q_{w_j}\right) \log P_{\mathrm{WTM}}\left(w_i\,\middle|\mathrm{M}_{w_j}\right). \end{aligned} \tag{14}$$

The parameters of the WTM models hence can be optimized iteratively using the following three EM update equations:

$$\hat{P}(w_i\,|T_k) = \frac{\sum_{w_j \in \mathbf{w}} c\left(w_i, Q_{w_j}\right) P\left(T_k\,\middle|w_i, \mathrm{M}_{w_j}\right)}{\sum_{w_l \in \mathbf{w}} \sum_{w_s \in Q_{w_l}} c\left(w_s, Q_{w_l}\right) P\left(T_k\,\middle|w_s, \mathrm{M}_{w_l}\right)}, \tag{15}$$

$$\hat{P}\left(T_k\,\middle|\mathrm{M}_{w_j}\right) = \frac{\sum_{w_s \in Q_{w_j}} c\left(w_s, Q_{w_j}\right) P\left(T_k\,\middle|w_s, \mathrm{M}_{w_j}\right)}{\sum_{w_l \in Q_{w_j}} c\left(w_l, Q_{w_j}\right)}, \tag{16}$$

$$P\left(T_k\,\middle|w_i, \mathrm{M}_{w_j}\right) = \frac{P(w_i\,|T_k)\, P\left(T_k\,\middle|\mathrm{M}_{w_j}\right)}{\sum_{l=1}^{K} P(w_i\,|T_l)\, P\left(T_l\,\middle|\mathrm{M}_{w_j}\right)}, \tag{17}$$

where $P(T_k \mid w_i, \mathrm{M}_{w_j})$ is the probability that the latent topic T_k occurs given the word w_i and the word model M_{w_j}, which is computed using the probability quantities $P(w_i \mid T_k)$ and $P(T_k \mid \mathrm{M}_{w_j})$ obtained in the previous training iteration.

On the other hand, supervised training of WTM can be performed in a similar way as that of PLSA described previously. That is, given a training set of query exemplars with the corresponding query-document relevance information that is available, the model parameters of WTM can be estimated by maximizing the total log-likelihood of the training set of query exemplars $\mathbf{Q}_{TrainSet}$ generated by their relevant documents:

$$\log L_{\mathbf{Q}_{TrainSet}} = \sum_{Q \in \mathbf{Q}_{TrainSet}} \sum_{D \in \mathbf{D}_{R \text{ to } Q}} \log P_{\mathrm{WTM}}(Q \mid \mathrm{M}_D). \tag{18}$$

The EM update equations for supervised training of WTM are omitted here for brevity.

It is noteworthy that in recent years the use of training query exemplars and the respective query-document relevance information (or the clickthrough information that to some extent reflects users' relative preferences of document relevance) also has been extensively studied for training various machine-learning-based retrieval models like SVM (Support Vector Machines) [Joachims and Radlinski 2007]. Such a training approach in essence has the ability to associate documents with a query exemplar even though they do not share any of the query words.

2.3 PLSA for Language Model Adaptation

The task of language modeling in speech recognition can be interpreted as calculating the probability $P(w_i \mid H_{w_i})$, in which w_i is a decoded word and H_{w_i} is one of its possible search histories. When PLSA is applied to language model adaptation in speech recognition, for a decoded word w_i, we can conceptually regard it as a (single-word) query and each of its corresponding search histories H_{w_i} as a document, or more precisely a DTM model $M_{H_{w_i}}$, used for predicting the occurrence probability of w_i [Gildea and Hoffmann 1999; Mrva and Woodland 2004]:

$$P_{\mathrm{PLSA}}\left(w_i \mid \mathrm{M}_{H_{w_i}}\right) = \sum_{k=1}^{K} P(w_i \mid T_k)\, P\left(T_k \mid \mathrm{M}_{H_{w_i}}\right). \tag{19}$$

However, the search histories are not known in advance and their number could be enormous and varying during speech recognition. Thus, the corresponding DTM model of a search history has to be estimated on the fly. One possible solution is that we can collect a set of contemporaneous (or in-domain) articles and use them to construct the latent topical factors $P(w_i \mid T_k)$ in an unsupervised manner beforehand, by treating each individual document in the collection as a relevant observation sequence to train its own DTM model, as earlier illustrated in Equations (6) to (8). Then, during the speech recognition process, we can keep the topic-specific unigram $P(w_i \mid T_k)$ unchanged, but let the search history's probability distribution over the latent topics $P(T_k \mid M_{H_{w_i}})$

be gradually updated as path extension is performed, by treating the history itself as a relevant observation sequence of words and using the following two EM update formulas [Gildea and Hoffmann 1999]:

$$\hat{p}\left(T_k \left| \mathrm{M}_{H_{w_i}}\right.\right) = \frac{\sum\limits_{w_j \in H_{w_i}} c\left(w_j, H_{w_i}\right) h\left(T_k \left| w_j, \mathrm{M}_{H_{w_i}}\right.\right)}{\sum\limits_{w_s \in H_{w_i}} c\left(w_s, \mathrm{H}_{w_i}\right)}, \tag{20}$$

and

$$h\left(T_k \left| w_j, \mathrm{M}_{H_{w_i}}\right.\right) = \frac{P\left(w_j | T_k\right) P\left(T_k \left| \mathrm{M}_{H_{w_i}}\right.\right)}{\sum\limits_{l=1}^{K} P\left(w_j | T_l\right) P\left(T_l \left| \mathrm{M}_{H_{w_i}}\right.\right)}, \tag{21}$$

where $c\left(w_j, H_{w_i}\right)$ is the number of occurrences of a specific word w_j in H_{w_i}.

2.4 WTM for Language Model Adaptation

The WTM model previously introduced in Section 2.2 for information retrieval also can be applied for language model adaptation [Chiu and Chen 2007]. Each WTM model M_{w_j} can be trained by concatenating those words occurring within a vicinity of, or a context window of size S around each occurrence of w_j, which are postulated to be relevant to w_j, in the contemporaneous (or in-domain) collection to form a relevant observation sequence of words O_{w_j} for training M_{w_j}, similar to that previously described in Equations (14) to (17). During the speech recognition process, for a decoded word w_i, we can again interpret it as a single-word observation; while for each of its search histories H_{w_i}, expressed by $H_{w_i} = w_1, w_2, \ldots, w_{i-1}$, we can linearly combine the associated WTM models of the words occurring in H_{w_i} to form a composite WTM model for predicting w_i:

$$\begin{aligned} P_{\mathrm{WTM}}\left(w_i \left| \mathrm{M}_{H_{w_i}}\right.\right) &= \sum_{j=1}^{i-1} \beta_j P_{\mathrm{WTM}}\left(w_i \left| \mathrm{M}_{w_j}\right.\right) \\ &= \sum_{j=1}^{i-1} \beta_j \sum_{k=1}^{K} P\left(w_i | T_k\right) P\left(T_k \left| \mathrm{M}_{w_j}\right.\right) \\ &= \sum_{k=1}^{K} P\left(w_i | T_k\right) P'\left(T_k \left| \mathrm{M}_{H_{w_i}}\right.\right), \end{aligned} \tag{22}$$

where the values of the nonnegative weighting coefficients β_j are empirically set to be exponentially decayed as the word w_j is being apart from w_i and summed to 1 ($\sum_{j=1}^{i-1} \beta_j = 1$), which have the form:

$$\beta_j = \phi_j \prod_{s=j+1}^{i-1} (1 - \phi_s), \tag{23}$$

where ϕ_j is set to a fixed value (between 0 and 1) for $j = 2, \cdots, i-1$, and set to 1 for $j = 1$; and β_j will be equal to ϕ_j when $j = i-1$. The search history's probability distribution over the latent topics $P'\left(T_k \left| M_{H_{w_i}}\right.\right)$ is thus represented as:

$$P'\left(T_k \left| \mathrm{M}_{H_{w_i}}\right.\right) = \sum_{j=1}^{i-1} \beta_j P\left(T_k \left| \mathrm{M}_{w_j}\right.\right). \tag{24}$$

The above two topic-based language models [cf. Equations (19) and (22)] have the advantage of taking account of the whole search history of a word irrespective of the length of the search history, and to some extent can dynamically capture the underlying global topical information of the search history. On the other hand, the background n-gram language probability can provide the general constraint information of lexical regularities. Thus, there is good reason to combine the PLSA or WTM language model with the background n-gram (e.g., trigram) language model to form an adaptive language model for guiding the speech recognition process:

$$\tilde{P}_{\mathrm{Adapt-1}}(w_i \,|w_{i-2}w_{i-1}) = \lambda_1 \cdot P_{\mathrm{PLSA}}\left(w_i \left| \mathrm{M}_{H_{w_i}}\right.\right) + (1-\lambda_1)\cdot P_{\mathrm{BG}}(w_i \,|w_{i-2}w_{i-1}), \tag{25}$$

$$\tilde{P}_{\mathrm{Adapt-2}}(w_i \,|w_{i-2}w_{i-1}) = \lambda_2 \cdot P_{\mathrm{WTM}}\left(w_i \left| \mathrm{M}_{H_{w_i}}\right.\right) + (1-\lambda_2)\cdot P_{\mathrm{BG}}(w_i \,|w_{i-2}w_{i-1}), \tag{26}$$

where $P_{\mathrm{PLSA}}\left(w_i \left| \mathrm{M}_{H_{w_i}}\right.\right)$ and $P_{\mathrm{WTM}}\left(w_i \left| \mathrm{M}_{H_{w_i}}\right.\right)$ are respectively the language model probabilities of the PLSA and WTM models, $P_{\mathrm{BG}}(w_i \,|w_{i-2}w_{i-1})$ is the background trigram language model probability, and λ_1 and λ_2 are tunable weighting parameters.

2.5 Theoretical Analysis of WTM and PLSA

WTM and PLSA can be analyzed from several perspectives. First, PLSA models the co-occurrence relationship between words and documents (or search histories), while WTM directly models the co-occurrence relationship between words in the collection. We may be able to view PLSA in a different way by regarding it as nonnegative (or probabilistic) matrix factorization. Given a vocabulary set of V distinct words and a collection of N documents, PLSA starts with a $V \times N$ "term-document" matrix $\mathbf{A}$ where each column n ($1 \le n \le N$) represents a document D_n in the document collection, and each entity of it, denoted by $a_{i,n}$ ($1 \le i \le V$), is conceptualized as the probability that a word w_i would occur in D_n, that is, $P_{\mathrm{PLSA}}\left(w_i \left| \mathrm{M}_{D_n}\right.\right)$. Then, matrix factorization of $\mathbf{A}$ will result in an approximation of $\mathbf{A}$ by a product of two nonnegative matrices:

$$\mathbf{A} \approx \mathbf{G} \times \mathbf{H}^{\mathrm{T}}, \tag{27}$$

where $\mathbf{G}$ is a $V \times K$ matrix ($K \le min(V,\ N)$), each entity $g_{i,k}$ ($1 \le k \le K$) of which accounts for the probability of a word w_i that would be generated by a latent topic T_k, that is, $P(w_i \,|T_k)$; $\mathbf{H}$ is an $N \times K$ matrix, each entity $h_{n,k}$ of which accounts for the probability that a document D_n will generate a latent topic T_k, that is, $P\left(T_k \left| \mathrm{M}_{D_n}\right.\right)$; and T denotes matrix transposition. Accordingly,

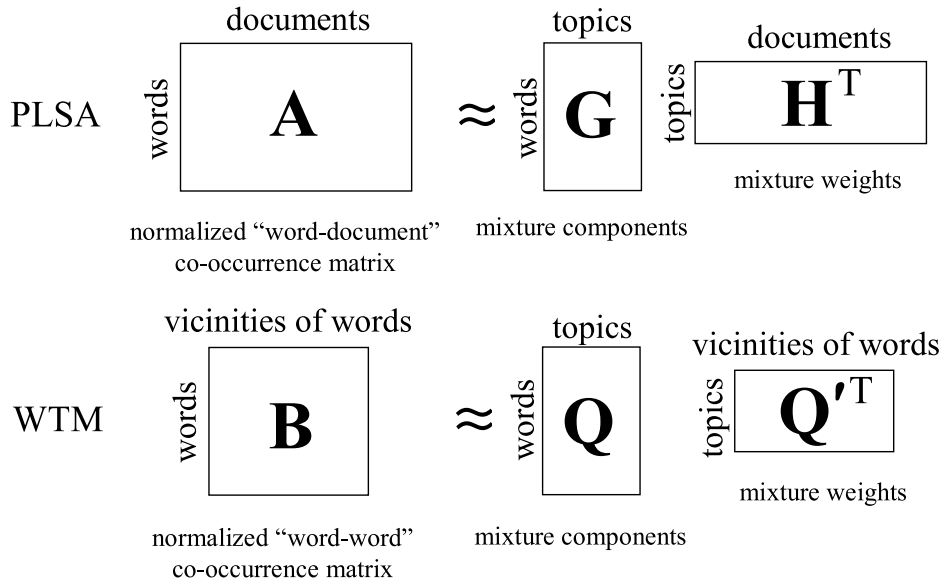

Fig. 1. A schematic comparison for the matrix factorizations of PLSA and WTM.

each entity $a_{i,n}$ of $\mathbf{A}$ can be efficiently approximated through the inner product of two vectors with probabilistic entities represented by the i-th row of $\mathbf{G}$ and the n-th row of $\mathbf{H}$, respectively. Such matrix factorization in fact is analogous to the singular value decomposition (SVD) procedure performed by the algebraic (or nonprobabilistic) counterpart of PLSA, namely, the latent semantic analysis (LSA) [Furnas et al. 1988]. Along a similar vein, we can also treat WTM as nonnegative (or probabilistic) matrix factorization, starting with a $V \times V$ "term-term" matrix $\mathbf{B}$ where each column j encodes the vicinity information of all occurrences of a word w_j ($1 \leq j \leq V$) in the document collection. More concretely, each entity $b_{i,j}$ ($1 \leq i \leq V$) of it is conceptualized as the probability that another word w_i would also appear in the vicinity of each occurrence of w_j in the collection, that is, $P_{\mathrm{WTM}}\left(w_i \left| \mathrm{M}_{w_j}\right.\right)$. Then, matrix factorization of $\mathbf{B}$ will lead to an approximation of $\mathbf{B}$ by a product of two nonnegative matrices:

$$\mathbf{B} \approx \mathbf{Q} \times \mathbf{Q}'^{T}, \tag{28}$$

where $\mathbf{Q}$ is a $V \times K$ ($K \leq V$) matrix and each entity $q_{i,k}$ of $\mathbf{Q}$ accounts for the probability of w_i being generated by a latent topic T_k, that is, $P(w_i|T_k)$; $\mathbf{Q}'$ is an $V \times K$ matrix, each entity $q'_{j,k}$ of which accounts for the probability that a latent topic T_k would be chosen given the vicinity information of w_j in the collection is known, that is, $P\left(T_k \left| \mathrm{M}_{w_j}\right.\right)$. Each entity $b_{i,j}$ of $\mathbf{B}$, similarly, can be efficiently approximated through the inner product of two vectors with probabilistic entities represented by the i-th row of $\mathbf{Q}$ and the j-th row of $\mathbf{Q}'$, respectively. Figure 1 shows a schematic comparison for the matrix factorizations of PLSA and WTM.

Second, for information retrieval, the topic mixture weights $P(T_k|\mathrm{M}_D)$ of PLSA for a new document D have to be estimated online using EM training, no matter whether the training is conducted in a supervised or unsupervised manner; on the contrary, the topic mixture weights $P(T_k|\mathrm{M}_D)$ of WTM can be efficiently estimated on the basis of the topic mixture weights $P\left(T_k\left|\mathrm{M}_{w_j}\right.\right)$ of words w_j involved in the document without using the time-consuming EM training procedure. For language model adaptation, PLSA again needs to perform EM training to estimate topic mixture weights $P\left(T_k\left|\mathrm{M}_{H_{w_i}}\right.\right)$ of a search history H_{w_i} on the fly; however, the topic mixture weights $P\left(T_k\left|\mathrm{M}_{H_{w_i}}\right.\right)$ of WTM [cf. Equation (22)] can be obtained on the basis of the topic mixture weights $P\left(T_k\left|\mathrm{M}_{w_j}\right.\right)$ of words w_j involved in the search history, which are instead trained in an offline manner. That is, unlike PLSA where the topic mixture weights trained with the contemporaneous (or in-domain) collection are entirely discarded during the speech recognition process, the topic mixture weights of WTM models are instead retained and exploited for language model adaptation. For our speech recognition test data, it was experimentally shown that the language model access time of WTM was roughly 1/30 of that of PLSA for language model adaptation, as the iteration number of the online EM estimation of $P\left(T_k\left|\mathrm{M}_{H_{w_i}}\right.\right)$ for PLSA was set to 5 [cf. Equations (20) and (21)]. This seems to indicate that WTM is more feasible than PLSA for practical speech recognition tasks.

Third, PLSA has $V \times K + K \times N$ parameters and WTM has $V \times K \times 2$ parameters; as shown previously, V denotes the size of the vocabulary set, N denotes the number of the documents, and K denotes the number of the latent topics used for training the IR or language models. It is obvious that the parameter number of WTM will be larger than that of PLSA, when the number of training documents in the collection is less than the number of distinct words in the vocabulary ($N < V$). The parameter number of PLSA grows linearly with the number of documents used for training PLSA; the parameter number of WTM instead remains the same regardless of the number of training documents, as the IR or speech recognition systems adopt a closed set of vocabulary. Recently, a latent Dirichlet allocation (LDA) method [Blei et al. 2003; Steyvers and Griffiths 2007] has been developed to address the above issue for PLSA. However, such a method still requires an iterative variational inference [Jordan 1999] procedure for online estimating the associated parameters of a newly observed document or search history.

Finally, it should be noted that for language model adaptation, if the context window for capturing the vicinity information of WTM is reduced to one word ($S = 1$), WTM can be either degenerated to a unigram model as the latent topic number K is set to 1, or viewed as analogous to a bigram model as $K = V$, or an aggregate Markov model as $1 < K < V$ [Saul and Pereira 1997]. Thus, with the appropriate values of S and K being chosen, WTM seems to be a good way to approximate the bigram or skip-bigram models for sparse data. On the other hand, WTM can be regarded as close in spirit to the class-based model (CBM) as well, by relating the latent topics of the former to the word

classes of the latter [Brown et al. 1992]. WTM differs from CBM in that WTM disregards word order information and contains word co-occurrence data from longer spans of the context window, whereas most of the approaches to using CBM are based purely on modeling word bigram sequences. This might mean that latent topics of WTM are semantic clusters whereas the classes of CBM are based on distributional clusters.

3. EXPERIMENTAL SETUP

3.1 Spoken Document Retrieval

We used two Topic Detection and Tracking (TDT) collections [LDC 2000] for the spoken document retrieval (SDR) task. TDT is a DARPA-sponsored program, where participating sites tackle tasks such as identifying the first time a news story is reported on a given topic, or grouping news stories with similar topics from the audio and textual streams of newswire data. Both the English and Mandarin Chinese corpora have been studied in the recent past. The TDT corpora have also been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project [Meng et al. 2004]. In this article we use the Mandarin Chinese collection of the TDT corpora for the retrospective retrieval task, such that the statistics for the entire document collection is obtainable. The Chinese news stories (text) from Xinhua News Agency are used as our test queries (or training query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts are used as the spoken documents. All news stories are exhaustively tagged with event-based topic labels, which serve as the relevance judgments for performance evaluation. Table I describes the details for the corpora used in this article. The TDT-2 collection is taken as the development set, which forms the basis for tuning the parameters in various retrieval models, including the dimensionality of the latent vector space in LSA, the weighting parameter between unigram and bigram probabilities in the HMM-based retrieval model, and the number of topical mixtures in the PLSA and WTM retrieval models. The TDT-3 collection is taken as the evaluation set; that is, all the experiments performed on it were conducted following the parameter setting that was optimized based on the TDT-2 development set. Therefore, the experimental results can validate the effectiveness of the proposed approaches on comparable real-world data.

The Dragon large-vocabulary continuous speech recognizer [Zhan et al. 1999] provided Chinese word transcripts for our Mandarin audio collections (TDT-2 and TDT-3). To assess the performance level of the recognizer, we spot-checked a fraction of the TDT-2 development set (about 39.90 hours) by comparing the Dragon recognition hypotheses with manual transcripts, and obtained a word error rate (WER) of 35.38%. Spot-checking approximately 76 hours of the TDT-3 test set gave a WER of 36.97%. Notice that Dragon's recognition output contains word boundaries (tokenizations) resulting from its own language models and vocabulary definition, while the manual transcripts are running texts without word boundaries. Since Dragon's lexicon is not

Table I. Statistics for TDT-2 and TDT-3 Collections Used for Spoken Document Retrieval

	TDT-2 (Development set) 1998, 02∼06				TDT-3 (Evaluation set) 1998, 10∼12			
# Spoken documents	2,265 stories, 46.03 hours of audio				3,371 stories, 98.43 hours of audio			
# Distinct test queries	16 Xinhua text stories (Topics 20001∼20096)				47 Xinhua text stories (Topics 30001∼30060)			
# Distinct training queries	819 Xinhua text stories (Topics 20001∼20096)				731 Xinhua text stories (Topics 30001∼30060)			
	Min.	Max.	Medium	Mean	Min.	Max.	Medium	Mean
Doc. length (characters)	23	4841	153	287.1	19	3667	159	415.1
Length of test query (in characters)	183	2623	329	532.9	98	1477	368	443.6
# Relevant documents per test query	2	95	13	29.3	3	89	12	20.1
Length of training query (in characters)	84	2980	412	510.0	49	4112	447	533.7
# Relevant documents per training query	2	95	87	74.4	3	60	13	20.6

available, we augmented the LDC Mandarin Chinese Lexicon with 24k words extracted from Dragon's word recognition output, and for computing error rates used the augmented LDC lexicon (about 51,000 words) to tokenize the manual transcripts. We also used this augmented LDC lexicon to tokenize the text training and test queries in the retrieval experiments.

The retrieval results, assuming manual transcripts for the spoken documents to be retrieved (denoted TD, text documents, in the tables below) are known, are also shown for reference, compared to the results when only the erroneous transcripts by speech recognition are available (denoted SD, spoken documents, below). The retrieval results are expressed in terms of non-interpolated mean average precision (*m*AP) following the TREC evaluation [Harman 1995; Baeza-Yates and Ribeiro-Neto 1999], which is computed by the following equation:

$$m\,\mathrm{AP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}}, \tag{29}$$

where L is the number of test queries, N_i is the total number of documents that are relevant to query Q_i, and $r_{i,j}$ is the position (rank) of the j–th document that is relevant to query Q_i, counting down from the top of the ranked list.

In this article, when PLSA and WTM are employed in evaluating the relevance between a query Q and a document D for IR, we additionally incorporate the unigram probabilities of a query term occurring in the document $P(w_i|D)$ and a general text corpus $P(w_i|Corpus)$ into PLSA and WTM, respectively, for probability smoothing and better performance. For example, the probability

of a word w_i generated by the WTM model of a word w_j involved in D [i.e., $P_{\mathrm{WTMM}}\left(w_i|\mathrm{M}_{w_j}\right)$ in Equation (12)] is therefore modified as follows:

$$\hat{P}_{\mathrm{WTM}}\left(w_i|\mathrm{M}_{w_j}\right) = (1-\rho_1-\rho_2)\cdot P_{\mathrm{WTM}}\left(w_i|\mathrm{M}_{w_j}\right) + \rho_1 \cdot P(w_i|D) + \rho_2 \cdot P(w_i|\mathit{Corpus}), \tag{30}$$

where ρ_1 and ρ_2 are weighting parameters ($0 < \rho_1, \rho_2 < 1$ and $\rho_1 + \rho_2 < 1$). Similar treatments also have been studied for the PLSA- [Hoffmann 1999] and the LDA-based [Wei and Croft 2006] retrieval models. The above two kinds of unigram probabilities, that is, $P(w_i|D)$ and $P(w_i|\mathit{Corpus})$, can be simply estimated using the maximum likelihood (ML) criterion, while their corresponding weights, that is, ρ_1 and ρ_2, can be further optimized using the EM algorithm [Chen et al. 2004b].

3.2 Spoken Document Transcription

The speech data set consists of about 112 hours of FM radio broadcast news [Chen et al. 2002], which were collected from several radio stations located at Taipei during November 1998 to April 2004. All the speech materials were manually segmented into separate stories, and each of them is a news abstract pronounced by one anchor speaker. Some of these stories contain background noise and music. Only 7.7 hours of speech data have corresponding orthographic transcripts, of which approximately 4.0 hours collected during 1998 to 1999 are used to bootstrap the acoustic training and the other 3.7 hours (506 stories) collected in September 2002 are used for testing. The remaining 104.3 hours of untranscribed speech data (about 18,600 stories) are reserved for unsupervised acoustic model training and unsupervised language model adaptation [Chen et al. 2004a].

The front-end processing is conducted with the data-driven linear discriminant analysis- based [Duda and Hart 1973] feature extraction approach. The states of each HMM were taken as the unit for class assignment. The outputs of 18 Mel-frequency filter banks are chosen as the basic vector. The basic vectors from every nine successive frames were spliced together to form the spliced vectors for the construction of a transformation matrix, which was then used to project the spliced vectors to a lower feature space for better discrimination. The dimension of the resultant vectors was set to 39 [Lee and Chen 2008]. The recognition lexicon consists of about 72,000 words. The language models used in this article consist of unigram, bigram, and trigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) (the Chinese Gigaword Corpus released by LDC). The *n*-gram language models were trained with the SRI language modeling toolkit (SRILM) [Stolcke 2000]. The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree search as well as a lexical prefix tree organization of the lexicon. The recognition hypotheses were organized into a word graph for further language model rescoring.

In this study, the results for the speech recognition experiments are evaluated in terms of character error rate (CER), defined as the sum of the insertion

Table II. Retrieval Results on the TDT-2 Development Set, Achieved Respectively, with WTM and PLSA Trained in a Supervised Manner

No. Latent Topics		2	4	8	16	32	64	128	256
WTM-S	TD	0.6505	0.6630	0.6887	0.7177	0.7351	0.7532	0.7672	0.7852
	SD	0.5731	0.5962	0.6186	0.6730	0.6864	0.7387	0.7558	0.7858
PLSA-S	TD	0.6362	0.6721	0.6750	0.6769	0.6823	0.6930	0.7243	0.7794
	SD	0.5759	0.5894	0.5918	0.5988	0.6255	0.6528	0.6652	0.6591

(*Ins*), deletion (*Del*), and substitution (*Sub*) errors between the recognized and reference Chinese character strings, divided by the total number of Chinese characters in the reference string (*Ref*):

$$\text{CER} = \frac{Ins + Sub + Del}{Ref}. \tag{31}$$

4. EXPERIMENTS ON SPOKEN DOCUMENT RETRIEVAL

4.1 Retrieval Results of WTM

We first evaluate the retrieval performance of the word topic models trained with supervision (denoted as WTM-S) and varying model complexities on the TDT-2 development set. The model parameters were trained using the 819 training query exemplars with their corresponding query-document relevance information to the TDT-2 development set [Chen et al. 2004b]. It should be borne in mind that from now on, unless otherwise stated, the retrieval results reported in *m*AP were obtained by evaluating the ranked list of documents returned by the retrieval models in response to each of the test queries (but not the training query exemplars). The retrieval results of WTM-S are shown in the upper part of Table II, where each column illustrates the retrieval results in both the TD and SD cases by using different numbers of latent topics for modeling WTM-S. As can be seen, the retrieval performance is steadily improved as the topic number increases. The best retrieval result of 0.7852 is obtained for the TD case when the topic number is set to 256, while the best result is 0.7858 for the SD case with the same topic mixture number. Notice that although the word error rate (WER) for the spoken document collection is higher than 35%, the average degradation in retrieval performance is much smaller, especially when the topic mixture number becomes larger. Such an observation indicates that the WER does not cause much adverse effect on retrieval performance, which is quite in parallel with those reported by other groups [Renals et al. 2000; Srinivasan and Petkovic 2000]. For example, most of the retrieval systems participating in the TREC-SDR evaluations had declared that the retrieval performance obtained by using the automatic transcripts of the spoken documents was flat with respect to the WER variations in the range of 15% to 30%, and they had also reported that no severe degradation was observed when evaluating the retrieval performance using the automatic transcripts of spoken documents compared to that using the manual transcripts [Garofolo et al. 2000; Chelba et al. 2008]. One

Table III. Retrieval Results on the TDT-2 Development Set, Achieved Respectively, with WTM and PLSA Trained in an Unsupervised Manner

No. Latent Topics		2	4	8	16	32	64	128	256
WTM-U	TD	0.6336	0.6350	0.6359	0.6368	0.6382	0.6386	0.6395	0.6287
	SD	0.5693	0.5723	0.5734	0.5733	0.5739	0.5767	0.5737	0.5652
PLSA-U	TD	0.6277	0.6332	0.6266	0.5973	0.5949	0.6267	0.6041	0.5878
	SD	0.5545	0.5659	0.5681	0.5503	0.5534	0.5664	0.5484	0.5831

possible reason is that a query word (or phrase) might occur repeatedly (more than once) within a broadcast news story and it is not always the case that all the occurrences of the word would be misrecognized totally as other words. For example, a word spoken by the studio anchor might have higher recognition accuracy than the same word spoken by the field reporter or the interviewee, which is mainly because for the anchor speech, the corresponding bandwidth variability, recording environment and speaking style, as well as the amount of acoustic training data, can be well controlled. Therefore, the true meaning of the word occurring within the spoken document could be still preserved for the following retrieval process.

In most real-world applications, it is not always the case that the retrieval systems can have query exemplars correctly labeled with the query-document relevance information to be used for model training. Thus, in this article, we study unsupervised model training for WTM (denoted as WTM-U). The retrieval results for the experiments carried out on the TDT-2 collection are shown in the upper part of Table III. As it can be seen, the results are not always improved as the topic number increases. The best result of 0.6395 for the TD case is obtained when the document topic number is set to 128, while the best result of 0.5767 for the SD case when document topic number is 64. When comparing with the best results achieved in supervised training, there are at most about 0.15 (0.7852 vs. 0.6395) and 0.21 (0.7858 vs. 0.5767) decreases in precision, respectively, for the TD and SD cases.

To recap, for the WTM retrieval model, given a training set of query exemplars with the corresponding query-document relevance information, the retrieval results obtained based on the supervised training approach (WTM-S) are much better than those based on the unsupervised approach (WTM-U). Our hope is that, given a set of real user queries and the associated click-through information about the retrieved relevant documents, the performance of retrieval systems might be incrementally improved through use.

4.2 Comparison of WTM and PLSA

In an attempt to assess the performance level of WTM as well as the associated approaches for model training and relevance measure, we study here the use of PLSA [Hoffmann 2001] for Chinese spoken document retrieval. The conventional PLSA retrieval model is trained in a purely unsupervised manner. The retrieval results of such a modeling approach (denoted as PLSA-U) on the TDT-2 collection are shown in the lower part of Table III, where each column illustrates the retrieval results in both the TD and SD cases by

Table IV. Retrieval Results on the TDT-2 Development Set, Achieved with HMM, VSM, and LSA, Respectively

Retrieval Model	HMM/Unigram	HMM/Bigram	VSM	LSA
TD	0.6327	0.5427	0.5548	0.5510
SD	0.5658	0.4803	0.5122	0.5310

using a different number of latent topics for PLSA modeling. The best retrieval result of 0.6332 is obtained for the TD case when the latent topic number is set to 4, while the best result is 0.5831 for the SD case with 256 topic mixtures. They are comparable to those achieved by WTM trained without supervision (WTM-U, cf. Table III), but considerably worse than those achieved by WTM trained with supervision (WTM-S, cf. Table II).

We also explore a supervised training approach for PLSA (denoted as PLSA-S), as described in Section 2.1. The same set of 819 query exemplars and their corresponding query-document relevance information to the TDT-2 development set are employed here again to iteratively estimate parameters of PLSA by using Equations (10) and (11). The retrieval results of such an approach on the TDT-2 collection are shown in the lower part of Table II, in which the best result of 0.7794 is obtained for the TD case when the document topic number is set to 256 and the best result of 0.6652 is obtained for the SD case with 128 topics. Such results are better than those obtained by using either WTM or PLSA trained in an unsupervised manner (WTM-U or PLSA-U), but are considerably worse than those obtained by using the WTM trained in a supervised manner (WTM-S). We can thus conclude that for the Chinese spoken document retrieval task studied here, WTM is truly a good alternative to PLSA when the retrieval models are trained either with or without supervision.

4.3 Comparison of WTM and Other Retrieval Models

Moreover, we also compare WTM with three other popular retrieval models: the vector space model (VSM) [Salton and McGill 1983], the latent semantic analysis (LSA) [Furnas et al. 1988], and HMM [Miller et al. 1999; Chen et al. 2004b]. The retrieval results of these three models on the TDT-2 collection are listed in Table IV for comparison. VSM and LSA are implemented with the best parameter settings; while for HMM, both the unigram and bigram modeling strategies are used, and the corresponding models are trained with the same set of 819 query exemplars in a supervised manner [Chen et al. 2004b]. As compared with the results in Tables II and III, it can be observed that WTM significantly outperforms all these three retrieval models when supervised learning is adopted (WTM-S). Even though WTM is trained in an unsupervised manner (WTM-U), its retrieval performance is still apparently better than that of VSM and LSA, and achieves quite competitive results to that of the HMM trained in a supervised manner. It is interesting that the retrieval performance of HMM degrades as the model structure becomes more sophisticated (e.g., from unigram to bigram modeling), whereas the retrieval performance of WTM and PLSA tends to become better as the topic number increased, when both models were trained in a supervised manner. Since the number of distinct words (51,000) is large, the estimation of bigram

Table V. Retrieval Results on the TDT-3 Evaluation Set, Achieved with WTM and PLSA, Which Are Trained in Both Supervised and Unsupervised Modes and with the Best Model Complexities Set by the TDT-2 Development Set, Respectively

	Supervised Training		Unsupervised Training	
Retrieval Model	WTM-S	PLSA-S	WTM-U	PLSA-U
TD	0.8009 (256 Topics)	0.7870 (256 Topics)	0.6587 (128 Topics)	0.6585 (4 Topics)
SD	0.7858 (256 Topics)	0.7937 (256 Topics)	0.6322 (64 Topics)	0.6582 (256 Topics)

Table VI. Retrieval Results on the TDT-3 Evaluation Set, Achieved with HMM, VSM, and LSA, Respectively

Retrieval Model	HMM/Unigram	HMM/Bigram	VSM	LSA
TD	0.6569	0.6143	0.6505	0.6440
SD	0.6308	0.5808	0.6216	0.6390

probabilities for the HMM inherently suffers from the sparse data problem [Chen et al. 2004b].

4.4 Further Evaluation on the TDT-3 Collection

Finally, in order to validate the effectiveness of the proposed WTM retrieval model on comparable real-world data, we further conducted a series of corresponding information retrieval experiments on the TDT-3 evaluation set. The retrieval results achieved by using WTM and PLSA are shown in Table V, while the results achieved by using the other retrieval models, such as HMM, VSM, and LSA, are shown in Table VI. For WTM, PLSA, and HMM, the training settings and model complexities for different experimental conditions (TD and SD cases) are set with the same configurations as those optimized using the TDT-2 collection; while for VSM and LSA, the model parameters are also set at the same optimum values tuned based on the TDT-2 collection as well.

We first examine the WTM trained in a supervised manner (WTM-S). The retrieval results are shown in the second column of Table V, in which the document models were respectively trained by another training query set consisting of 731 query exemplars together with the corresponding query-document relevance information to the TDT-3 evaluation set. A retrieval result of 0.8009 for the TD case is obtained with the document topic number set to 256, while a result of 0.7858 for the SD case with the same topic number. Comparatively speaking, these results are comparable to the results achieved by the PLSA (as respectively shown in the third (PLSA-S) and rightmost (PLSA-U) columns of Table V), and substantially better than those obtained by the other retrieval models (as shown in Table VI).

We then examine the retrieval performance of the WTM trained without supervision (WTM-U). It is worth mentioning that both the probabilities $P(w_i|T_k)$ and $P\left(T_k|\mathrm{M}_{w_j}\right)$ used to construct WTM-U for the TDT-3 collection were directly adopted from that of the TDT-2 collection, as opposed to PLSA where $P(w_i|T_k)$ and $P(T_k|\mathrm{M}_D)$ had to be re-estimated using the EM algorithm for either supervised or unsupervised training. According to the results shown in the fourth column (WTM-U) of Table V, WTM trained

in an unsupervised manner is considerably worse than PLSA trained with supervision (PLSA-S); however, it still is comparable to PLSA trained without supervision (PLSA-U) and achieves better retrieval performance than the other models in most retrieval conditions, though the differences are not significant.

Based on the experimental results achieved from this and previous sections, it has been clearly demonstrated that the WTM trained in a supervised manner does achieve better performance than the other conventional retrieval models for the Chinese spoken document retrieval task studied here. All the IR experiments throughout this article have been carefully designed to avoid "testing on training"; that is, all the training (or parameter) settings and model complexities are tuned or optimized by using the TDT-2 development set and tested on both the TDT-2 development set and the TDT-3 evaluation set. Generally speaking, the training settings and model complexities tuned from the TDT-2 development set perform rather well in the TDT-3 evaluation set.

5. EXPERIMENTS ON SPOKEN DOCUMENT TRANSCRIPTION

As mentioned earlier in Section 3.2, a set of 506 broadcast news stories collected in September 2002 is used for testing. On the other hand, a set of about 39,000 text news stories collected from CNA during August to October 2002 is taken as the contemporaneous adaptation data, while another set of about 18,600 automatic transcripts (3.2 million Chinese characters) as the in-domain adaptation corpus. These two sets of corpora are postulated to be either temporally or stylistically consistent with the broadcast news speech to be tested, and therefore can be used to explore the global topical and local contextual information which might be helpful for speech recognition.

5.1 Comparison of WTM and PLSA

The baseline system results in a character error rate (CER) of 15.22% and a perplexity (PP) of 752.49 on the test set. We first compare the performance levels of WTM and PLSA for language model adaptation by varying the model complexities (the number of latent topics ranged from 16 to 256) and using either the contemporaneous newswire texts (denoted as Texts) or the in-domain automatic transcripts (denoted as Automatic Transcripts). The constant ϕ_j,$j = 2, \ldots, i-1$, in Equation (23) and the weighting parameters λ_1 and λ_2 in Equations (25) and (26) were respectively set at optimum values (ϕ_j=0.6, $j = 2, \ldots, i-1$; λ_1=0.1; and λ_2=0.3 in this research), using a held-out set different form the test set. As can be seen from Table VII, for these two variants of latent topic approaches, both CER and PP are steadily reduced as the topic mixture number increases, when the contemporaneous texts were used for language model adaptation. For PLSA, the best result of CER of 14.47% (4.93% relative reduction) and PP of 510.20 (32.20% relative reduction) is obtained when the topic number is set to 256, while for WTM, the best result of CER of 14.38% (5.52% relative reduction) and PP of 508.29 (32.45% relative reduction) is obtained with the same topic number. Though the performance seems not to be saturated yet, these results clearly demonstrate the effectiveness of these

Table VII. CER (%) and Perplexity (PP) Results, Achieved Using WTM and PLSA, Respectively, for Language Model Adaptation

		CER (%)		PP	
Baseline (Background Trigram Model)		15.22%		752.49	
		WTM		PLSA	
Adaptation Corpus	No. Latent Topics	CER (%)	PP	CER (%)	PP
Texts	16	14.77	566.10	14.83	588.51
	32	14.69	553.88	14.73	571.46
	64	14.60	540.62	14.58	552.80
	128	14.44	524.15	14.53	527.41
	256	14.38	508.29	14.47	510.20
Automatic Transcripts	16	14.87	574.60	14.99	591.21
	32	14.90	568.76	14.92	580.80
	64	14.85	564.56	14.82	569.93
	128	14.81	563.25	14.87	562.45
	256	14.96	567.53	14.92	565.85

two latent topic approaches for dynamic language model adaptation. WTM performs comparably to PLSA in most cases when the contemporaneous texts were used for language model adaptation; however, it has the advantage of not needing to use the EM update formulas [as illustrated in Equations (19), (20), and (21)] to iteratively update the weights of search histories over the latent topics during the speech recognition process. Therefore, WTM is believed to be more efficient than PLSA for the speech recognition task studied here (cf. Section 2.5).

On the other hand, as the in-domain automatic transcripts were instead used for language model adaptation, for PLSA, the best result of CER of 14.82% (2.62% relative reduction) is achieved when the topic number is 64, and the best result of PP of 562.45 (25.28% relative reduction) is achieved when the topic number is 256; while for WTM, the best result of CER of 14.81% (2.69% relative reduction) and PP of 563.25 (25.15% relative reduction) is achieved when 128 topics are used. However, when comparing the results obtained with different model complexities, it can be found that WTM is quite comparable with PLSA such that the differences between WTM and PLSA are almost negligible as the in-domain automatic transcripts were exploited. By and large, language model adaptation using the in-domain automatic transcripts is quite competitive with that using the contemporaneous newswire texts in PP reduction, but only reaches about half of the CER reduction as that provided by using the contemporaneous newswire texts.

Significance tests based on the standard NIST MAPSSWE [Pallett et al. 1990] also have been conducted on the speech recognition results of both the WTM and PLSA adaptation approaches, and they all indicated the statistical significance of CER improvements (with P-value <0.001) over the baseline trigram system when either the contemporaneous newswire texts or the in-domain automatic transcripts were adopted as the adaptation corpus. However, the influence of the recognition errors of the automatic transcripts on language model adaptation using the probabilistic latent topical information is still under extensive investigation.

5.2 Fusion of Topical and Contextual Information

Though WTM and PLSA, aiming at exploring the long-span latent topical information of the search histories, have been shown effective for language model adaptation, the local word regularity information inherent in the adaptation corpus is still vital and worthy of being taken into account when performing language model adaptation. Therefore, we investigate here the integration of WTM or PLSA with two variants of the widely-used *n*-gram language model adaptation approach, that is, count merging and model interpolation, each of which can be respectively viewed as maximum a posteriori (MAP) language model adaptation [Bacchiani and Roark 2003] with a different parameterization of the prior distribution and can be used to capture the local regularities of word usage in the new task domain. The integration is performed through simple linear model interpolation, and the contemporaneous newswire texts and in-domain automatic transcripts, as well as their combination, were exploited as well. More specifically, PLSA and WTM are combined with the count merging approach through the following two equations, respectively:

$$\tilde{P}_{\text{Adapt}-3}(w_i|w_{i-2}w_{i-1}) = \gamma_1 \cdot P_{\text{PLSA}}\left(w_i \middle| \mathbf{M}_{H_{w_i}}\right) + (1-\gamma_1)\cdot P_{\text{BG}-\text{CONT}}(w_i|w_{i-2}w_{i-1}), \tag{32}$$

$$\tilde{P}_{\text{Adapt}-4}(w_i|w_{i-2}w_{i-1}) = \gamma_2 \cdot P_{\text{WTM}}\left(w_i \middle| \mathbf{M}_{H_{w_i}}\right) + (1-\gamma_2)\cdot P_{\text{BG}-\text{CONT}}(w_i|w_{i-2}w_{i-1}), \tag{33}$$

where $P_{\text{BG}-\text{CONT}}(w_i|w_{i-2}w_{i-1})$ was trained by merging the trigram counts of the background language model training corpus either with that of the contemporaneous news text, or the in-domain automatic transcripts, or the combination of the both. The values for γ_1 and γ_2 in Equations (32) and (33) were respectively set at optimum values (γ_1=0.1 and γ_2=0.3), using the held-out set. On the other hand, PLSA and WTM were combined with the model interpolation approach through the following two equations, respectively:

$$\begin{aligned}\tilde{P}_{\text{Adapt}-5}(w_i|w_{i-2}w_{i-1}) &= \nu_1 \cdot P_{\text{PLSA}}\left(w_i \middle| \mathbf{M}_{H_{w_i}}\right) + \kappa_1 \cdot P_{\text{CONT}}(w_i|w_{i-2}w_{i-1}) \\ &\quad + (1-\nu_1-\kappa_1)\, P_{\text{BG}}(w_i|w_{i-2}w_{i-1}),\end{aligned} \tag{34}$$

$$\begin{aligned}\tilde{P}_{\text{Adapt}-6}(w_i|w_{i-2}w_{i-1}) &= \nu_2 \cdot P_{\text{WTM}}\left(w_i \middle| \mathbf{M}_{H_{w_i}}\right) + \kappa_2 \cdot P_{\text{CONT}}(w_i|w_{i-2}w_{i-1}) \\ &\quad + (1-\nu_2-\kappa_2)\, P_{\text{BG}}(w_i|w_{i-2}w_{i-1}),\end{aligned} \tag{35}$$

where $P_{\text{CONT}}(w_i|w_{i-2}w_{i-1})$ was trained using either the contemporaneous news text, or the in-domain automatic transcripts, or the combination of both. The values for ν_1, κ_1, ν_2 and κ_2 in Equations (34) and (35) were respectively set at optimum values (ν_1=0.1, κ_1=0.4, ν_2=0.3 and κ_2=0.35) using the held-out set as well.

The results obtained by respectively combining the WTM and PLSA with two variants of the MAP-based approaches (count merging and model interpolation) are shown in Table VIII. As can be observed, fusion of these two kinds of information sources does provide additional gains in most experimental conditions. Moreover, the combination of WTM with the MAP-based approaches

Table VIII. CER (%) and Perplexity (PP) Results, Achieved by Combining Either WTM or PLSA with the Two Variants of MAP-based Adaptation Approaches, Respectively

	CER (%)	PP
WTM (256 Topics, Texts) + Count Merging (Texts)	13.49	357.31
WTM (256 Topics, Texts) + Model Interpolation (Texts)	13.50	365.60
WTM (128 Topics, Automatic Transcripts) + Count Merging (Automatic Transcripts)	14.85	508.04
WTM (128 Topics, Automatic Transcripts) + Model Interpolation (Automatic Transcripts)	15.06	499.04
WTM (256 Topics, Texts, and Automatic Transcripts) + Count Merging (Texts and Automatic Transcripts)	13.45	334.09
PLSA (256 Topics and Texts) + Count Merging (Texts)	13.27	336.84
PLSA (256 Topics and Texts) + Model Interpolation (Texts)	13.37	339.16
PLSA (64 Topics and Automatic Transcripts) + Count Merging (Automatic Transcripts)	14.65	506.22
PLSA (64 Topics and Automatic Transcripts) + Model Interpolation (Automatic Transcripts)	15.00	471.67
PLSA (256 Topics, Texts, and Automatic Transcripts) + Count Merging (Texts and Automatic Transcripts)	13.23	313.20

Table IX. CER (%) and Perplexity (PP) Results, Achieved by Combining WTM with PLSA

		WTM + PLSA	
Adaptation Corpus	No. Latent Topics	CER (%)	PP
Texts	16	14.78	551.62
	32	14.61	530.00
	64	14.47	506.19
	128	14.34	474.87
	256	14.21	449.09
Automatic Transcripts	16	14.95	546.54
	32	14.97	531.40
	64	14.82	516.82
	128	14.81	504.77
	256	14.94	502.06
Texts + Automatic Transcripts	256	14.10	441.07

yields slightly worse results than the combination of PLSA with the MAP-based ones, though WTM has been previously shown to yield better results than PLSA when the contemporaneous newswire texts are exploited and to be competitive with PLSA when in-domain automatic transcripts are exploited. The combination of the latent topic approach (PLSA with 256 topics) with the count merging approach, as shown in the last row of Table VIII, achieves the best result of CER of 13.23% (13.07% relative reduction) and PP of 313.20 (58.38% relative reduction). On the other hand, attempts also have been made to combine PLSA with WTM, and the results are shown in Table IX. However, no significant performance gain is observed in word error rate as well as

perplexity reductions compared to the results obtained by combining either WTM or PLSA with the two variants of MAP-based approaches, respectively.

6. CONCLUSIONS

In this article, we have proposed a word topic model (WTM) to explore the co-occurrence relationship between words, as well as the long-span latent topical information, for language modeling in spoken document retrieval and transcription. The document or the search history as a whole is modeled as a composite WTM model for predicting the newly observed word. The underlying characteristics and different kinds of model structures were extensively investigated, while the performance of WTM was analyzed and verified by comparison with the well-known probabilistic latent semantic analysis (PLSA) model as well as the other models. Experimental results on the Mandarin spoken document retrieval and transcription tasks both demonstrate WTM is indeed a feasible alternative to the existing models. Future work on the WTM-based approaches includes better model inference or adaptation of WTM [Blei et al. 2003; Steyvers and Griffiths 2007], discriminative training of WTM [Chen et al. 2004b; Kuo and Chen 2005; Gao et al. 2006], and applying WTM to spoken document summarization [Chen et al. 2009].

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for valuable comments that greatly improved the quality of this article.

REFERENCES

ALLAN, J. ED., ASLAM, J., BELKIN, N., BUCKLEY, C., CALLAN, J., CROFT, W. B. ED., DUMAIS, S., FUHR, N., HARMAN, D., HARPER, D., HIEMSTRA, D., HOFFMANN, T., HOVY, E., KRAAIJ, W., LAFFERTY, J., LAVRENKO, V., LEWIS, D., LIDDY, L., MANMATHA, R., MCCALLUM, A., PONTE, J., PRAGER, J., RADEV, D., RESNIK, P., ROBERTSON, S., ROSENFELD, R., ROUKOS, S., SANDERSON, M., SCHWARTZ, R., SINGHAL, A., SMEATON, A., TURTLE, H., VOORHEES, E., WEISCHEDEL, R., XU, J., AND ZHAI, C. 2003. Challenges in information retrieval and language modeling. *SIGIR Forum, 37*, 1, 1–17.

BACCHIANI, M. AND ROARK, B. 2003. Unsupervised language model adaptation. In *Proceedings of the IEEE International Conference Acoustics, Speech, Signal Processing (ICASSP'03)*, 224–227.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval.* Addison Wesley, Reading, MA.

BELLEGARDA, J. R. 2004. Statistical language model adaptation: Review and perspectives. *Speech Commun. 42*, 11, 93–108.

BERGER, A. AND LAFFERTY, J. 1999. Information retrieval as statistical translation. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, 222–229.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BROWN, P. F., DELLAPIETRA, V. J., DESOUZA, P. V., LAI, J. C., AND MERCER, R. L. 1992. Class-based *n*-gram models of natural language. *Comput. Linguist. 18*, 4, 467–479.

BYRNE, W., DOERMANN, D., FRANZ, M., GUSTMAN, S., HAJIČ, J., OARD D., PICHENY M., PSUTKA, J., RAMABHADRAN B., SOERGEL, D., WARD, T., AND ZHU, W. J. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech Audio Process. 12*, 4, 420–435.

CHELBA C., HAZEN, T. J., AND SARACLAR, M. 2008. Retrieval and browsing of spoken content. *IEEE Signal Process. Mag. 25*, 3, 39–49.

CHEN, B., WANG, H.-M., AND LEE, L.-S. 2002. Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese. *IEEE Trans. Speech Audio Process. 10*, 5, 303–314.

CHEN, B., KUO, J.-W., AND TSAI W.-H. 2004a. Lightly supervised and data-driven approaches to Mandarin broadcast news transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, 777–780.

CHEN, B., WANG, H.-M., AND 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, Y.-T., CHEN, B., AND WANG, H.-M., 2009. A probabilistic generative framework for extractive broadcast news speech summarization. *Trans. Audio Speech Lang. Process. 17*, 1, 95–106.

CHIU, H.-S. AND CHEN, B. 2007. Word topical mixture models for dynamic language model adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, 169–172.

CROFT, W. B. AND LAFFERTY, J. Eds. 2003. *Language Modeling for Information Retrieval*. Kluwer-Academic Publishers.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B 39*, 1, 1–38.

DUDA, R. O. AND HART, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York.

FEDERICO, M. AND BERTOLDI, N. 2001. Broadcast news LM adaptation using contemporary texts. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'01)*, 239–242.

FURNAS, G. W., DEERWESTER, S., DUMAIS, S. T., LANDAUER, T. K., HARSHMAN, R., STREETER, L. A., AND LOCHBAUM, K. E. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'88)*, 465–480.

GAO, J., SUZUKI, H., AND YUAN, W. 2006. An empirical study on language model adaptation. *ACM Trans. Asian Lang. Inform. Process. 5*, 3, 209–227.

GAROFOLO, J., AUZANNE, G., AND VOORHEES, E. 2000. The TREC spoken document retrieval track: A success story. In *Proceedings of the 8th Text Retrieval Conference (TREC'00)*, 107–129.

GILDEA, D. AND HOFFMANN, T. 1999. Topic-based language models using EM. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'99)*, 2167–2170.

HARMAN, D. 1995. Overview of the 4th Text Retrieval Conference. In *Proceedings of the 4th Text Retrieval Conference (TREC'95)*, 1–23.

HOFFMANN, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 50–57.

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

JELINEK, F. AND MERCER, R. L. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop Pattern Recognition in Practice (PRP'80)*, 381–397.

JELINEK, F., MERIALDO, B., ROUKOS, S., AND STRAUSS, M. 1991. A dynamic language model for speech recognition. In *Proceedings of the Speech and Natural Language DARPA Workshop (DARPA'91)*, 293–295.

JOACHIMS, T. AND RADLINSKI, F. 2007. Search engines that learn from implicit feedback. *IEEE Trans. Comput. 40*, 8, 34–40.

JORDAN, M., Ed. 1999. *Learning in Graphical Models*. Cambridge, MA: MIT Press.

KUO, J.-W. AND CHEN, B. 2005. Minimum word error based discriminative training of language models. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'05)*, 1277–1280.

LAFFERTY, L. AND ZHAI, C. 2001. Document language models and risk minimization for information retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'01)*, 111–119.

LDC. 2000. Project topic detection and tracking. Linguistic Data Consortium. Retrieved from http://www.ldc.upenn.edu/Projects/TDT/.

LEE, L.-S. AND CHEN, B. 2005. Spoken document understanding and organization. *IEEE Signal Processing Magazine 22*, 5, 42–60.

LEE, H.-S. AND CHEN, B. 2008. Linear discriminant feature extraction using weighted classification Confusion information. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'08)*, 2254–2257.

LIU, X. AND CROFT, W. B. 2005. Statistical language modeling for information retrieval. In *Annual Review of Information Science and Technology 39*, Chapter 1, 3–31.

MENG, H., CHEN, B., KHUDANPUR, S., LEVOW, G. A., LO, W. K., OARD, D., SCHONE, P., TANG, K., WANG, H. M., AND WANG, J. 2004. Mandarin–English information (MEI): Investigating translingual speech retrieval. *Comput. Speech and Lang. 18*, 2, 163–179.

MILLER, D. R. H., LEEK, T., AND SCHWARTZ, R. 1999. A hidden Markov model information retrieval system. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 214–221.

MRVA, D. AND WOODLAND, P. C. 2004. A PLSA-based language model for conversational telephone speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'04)*, 2257–2260.

PALLETT, D., FISHER, W., AND FISCUS, J. 1990. Tools for the analysis of benchmark speech recognition tests. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90), 1*, 97–100.

PONTE, J. M. AND CROFT, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'98)*, 275–281.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE 77*, 2, 257–286.

RENALS, S., ABBERLEY, D., KIRBY, D., AND ROBINSON, T. 2000. Indexing and retrieval of broadcast news. *Speech Comm. 32*, 5–20.

ROSENFELD, R. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE 88*, 8, 1270–1278.

SALTON, G. and MCGILL, M. J. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

SAUL, L. AND PEREIRA, F. 1997. Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, 81–89.

SONG, F. AND CROFT, W. B. 1999. A general language model for information retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'99)*, 316–321.

SRINIVASAN, S. AND PETKOVIC, D. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'00)*, 81–87.

STEYVERS, M. AND GRIFFITHS, T. 2007. Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch Eds., *Handbook of Latent Semantic Analysis.* Erlbaum, Hillsdale, NJ.

STOLCKE, A. 2000. SRI language modeling toolkit. http://www.speech.sri.com/projects/srilm/.

WEI, X. AND CROFT, W. B. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'06)*, 178–185.

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

ZHAN, P., WEGMANN, S., AND GILLICK, L. 1999. Dragon Systems' 1998 broadcast news transcription system for Mandarin. In *Proceedings of the DARPA Broadcast News Workshop (DARPA'99)*. 183–186.

Received April 2008; revised August 2008; accepted November 2008