Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 65-86

A Comparative Study of Methods for Topic Modeling in Spoken Document Retrieval

Shih-Hsiang Lin[+,*] and Berlin Chen[#,*]

Abstract

Topic modeling for information retrieval (IR) has attracted significant attention and demonstrated good performance in a wide variety of tasks over the years. In this paper, we first present a comprehensive comparison of various topic modeling approaches, including the so-called document topic models (DTM) and word topic models (WTM), for Chinese spoken document retrieval (SDR). Moreover, different granularities of index features, including words, subword units, and their combinations, are also exploited to work in conjunction with various extensions of topic modeling presented in this paper, so as to alleviate SDR performance degradation caused by speech recognition errors. All of the experiments were performed on the TDT Chinese collection.

Keywords: Information Retrieval, Document Topic Models, Word Topic Models, Spoken Document Retrieval.

1. Introduction

Due to the advances in computer technology and the proliferation of Internet activity, huge volumes of multimedia data, such as text files, broadcast radio and television programs, lectures, and digital archives, are continuously growing and filling networks. Development of intelligent and efficient information retrieval techniques to provide people with easy access to all kinds of information is now becoming more and more emphasized. Meanwhile, with the rapid evolution of speech recognition technology, substantial efforts and very encouraging results on spoken document retrieval (SDR) also have been demonstrated in the recent past. Although most retrieval systems participating in the TREC-SDR evaluations claimed that speech recognition errors do not seem to cause much adverse effect on SDR performance

[+] Voice Division Research Center, Delta Electronics
E-mail: shlin@csie.ntnu.edu.tw
[#] Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: berlin@csie.ntnu.edu.tw
[*] Corresponding authors.

when merely using imperfect recognition transcripts derived from one-best recognition results from a speech recognizer (Garofolo *et al.*, 2000; Chelba *et al.*, 2008), this is probably attributed to the fact that the TREC-style test queries tend to be quite long and contain different words describing similar concepts that can help the queries match their relevant spoken documents. Furthermore, a query word (or phrase) may occur repeatedly (more than once) within a relevant spoken document, and it is not always the case that all of the occurrences of the word would be misrecognized totally as other words. We, however, believe that SDR would still present a challenge in situations where the queries are relatively short and there exists severe deviation in word usage between the queries and spoken documents.

Among several promising information retrieval approaches, statistical language modeling (LM) (Ponte & Croft, 1998), aiming to capture the regularity in human natural language and quantify the acceptability of a given word sequence, has continuously been a focus of active research in the last decade (Miller *et al*., 1999; Hofmann, 2001). The basic idea is that each individual document in the collection is treated as a probabilistic language model for generating a given query. A document is deemed to be relevant to a query if its corresponding document language model generates the query with higher likelihood. In practice, the relevance measure for the LM approach is usually computed by two different matching strategies, namely, *literal term matching* and *concept matching* (Lee & Chen, 2005). The unigram language model (ULM) is perhaps the most representative example for literal term matching strategy (Miller *et al*., 1999). In the ULM approach, each document is interpreted as a generative model composed of a mixture of unigram (multinomial) distributions for observing a query, while the query is regarded as observations, expressed as a sequence of indexing words (or terms).

Nevertheless, these approaches would suffer from the problems of word usage diversity, which might make the retrieval performance of the system degrade severely as a given query and its relevant documents are using quite a different set of words. In contrast, the concept matching strategy tries to explore the topic information conveyed in the query and documents. Based on this, the retrieval process is performed. The probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and the latent Dirichlet allocation (LDA) (Blei *et al*., 2003) are often considered to be two basic representatives of this category. They both introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics. More specifically, the relevance between a query and a document is not computed directly based on the frequency of the query words occurring in the document, but instead based on the frequency of these words appearing in the latent topics as well as the likelihood that the document generates those respective topics, which exhibits some sort of concept matching. Further, although there have been many follow-up studies and extensions of PLSA and LDA, it has been shown that more sophisticated (or complicated) topic models, such as the pachinko

allocation model (PAM) and correlated topic model (CTM), do not necessarily offer further retrieval benefits (Zhai, 2008; Blei & Lafferty, 2009). On the other hand, rather than treating each document as a whole as a document topic model (DTM), such as PLSA and LDA, the word topic model (WTM) (Chen, 2009) attempts to discover the long-span co-occurrence dependence "*between words*" through a set of latent topics, while each document in the collection consequently can be represented as a composite WTM model in an efficient way for predicting an observed query. Interested readers can refer to Griffiths *et al*. (2007), Zhai (2008), and Blei and Lafferty (2009) for a thorough and updated overview of the major topic-based language models that have been successfully developed and applied to various IR tasks.

Although most of the above approaches can be equally applied to both text and spoken documents, the latter presents unique difficulties, such as speech recognition errors, problems posed by spontaneous speech, and redundant information. A straightforward remedy, apart from the conventional approaches target at improving recognition accuracy, is to develop more robust representations of spoken documents for spoken document retrieval (SDR). For example, multiple recognition hypotheses, beyond the top scoring ones, are expected to provide alternative representations for the confusing portions of the spoken documents (Chelba *et al*., 2008; Chia *et al*., 2008). Another school of thought attempts to leverage subword units, as well as the combination of words and subword units, for representing the spoken documents, which also has been shown beneficial for SDR. The reason for the fusion of word- and subword-level information is that incorrectly recognized spoken words often include several subword units that are correctly recognized. Hence, the retrieval process based on subword-level representations may take advantage of partial matching (Lin & Chen, 2009).

With the above inspiration in mind, we first compare the structural characteristics of various topic models for Chinese SDR, including PLSA and LDA, as well as WTM. The utility of these models is thoroughly examined using both long and short test queries. Moreover, different granularities of index features, including words, subword units, and their combinations, are also exploited to work in conjunction with various extensions of topic modeling presented in this paper, so as to alleviate SDR performance degradation caused by imperfect recognition transcripts. To our knowledge, there is little literature on leveraging various topic decompositions together with various granularities of index features for topic modeling in SDR.

The rest of this paper is structured as follows. Section 2 elucidates the structural characteristics of the different types of topic models for the retrieval purpose. Section 3 discusses two different extensions of topic modeling. Section 4 describes the spoken document collection used in this paper, as well as the experimental setup. A series of experiments and associated discussions are presented in Section 5. Finally, Section 6 concludes this paper and

suggests possible avenues for future work.

2. Topic Models

In this section, we first describe the probabilistic generative framework for information retrieval. We then briefly review the document topic models (DTM), including the probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and the latent Dirichlet model (LDA) (Blei *et al.*, 2003; Wei & Croft, 2006), followed by an introduction to the word topic model (WTM) (Chen, 2009), as well as the word Dirichlet topic model (WDTM).

2.1 Probabilistic Generative Framework

When the language modeling approach is applied to IR, it basically makes use of a probabilistic generative framework for ranking each document D in the collection given a query Q, which can be expressed by $P(D|Q)$. By applying Bayes' theorem, this ranking criterion can be approximated by the likelihood of Q generated by D, *i.e.*, $P(Q|D)$, when we assume that the prior probability of each document $P(D)$ is uniformly distributed. For this idea to work, each document D is treated as a probabilistic language model M_D for generating the query. Furthermore, if the query Q is treated as a sequence of words (or terms), $Q = w_1 w_2 \ldots w_N$, where the query words are assumed to be conditionally independent given the document model M_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 further decomposed as a product of the probabilities of the query words generated by the document:

$$P(Q|D) = \prod_{w_i \in Q} P(w_i|\mathrm{M}_D)^{c(w_i,Q)}, \tag{1}$$

where $c(w_i,Q)$ is the number of times that each distinct word w_i occurs in Q. The document ranking problem has now been reduced to the problem of constructing the document model $P(w_i|\mathrm{M}_D)$.

The simplest way to construct $P(w_i|\mathrm{M}_D)$ is based on literal term matching, or using the unigram language model (ULM), where each document of the collection can respectively offer a unigram distribution for observing a query word, i.e., $P_{\mathrm{ULM}}(w_i|\mathrm{M}_D)$, which is estimated on the basis of the words occurring in the document:

$$P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_D) = \frac{c(w_i,D)}{|D|}, \tag{2}$$

where $c(w_i,D)$ is the number of times that word w_i occurs in the document D and $|D|$ is the number of words in the document. In order to avoid the problem of zero probability, the ULM is usually smoothed by a unigram distribution estimated from a general collection, *i.e.*, $P_{\mathrm{ULM}}(w_i|\mathrm{M}_C)$:

$$\hat{P}_{\mathrm{ULM}}\left(w_i \middle| D\right) = \lambda \cdot P_{\mathrm{ULM}}\left(w_i \middle| \mathrm{M}_D\right) + \left(1-\lambda\right) \cdot P_{\mathrm{ULM}}\left(w_i \middle| \mathrm{M}_C\right), \tag{3}$$

where λ is a weighting parameter. It turns out that a document with more query words occurring in it would tend to receive a higher probability; further, the use of $P_{\mathrm{ULM}}\left(w_i \middle| \mathrm{M}_C\right)$ to some extent can help deemphasize common (non-informative) words but instead put more emphasis on discriminative (or informative) words for the purpose of document ranking (Zhai, 2008). In the following, $P_{\mathrm{ULM}}\left(w_i \middle| \mathrm{M}_D\right)$ and $P_{\mathrm{ULM}}\left(w_i \middle| \mathrm{M}_C\right)$ will be termed the document model and the background model, respectively.

2.2 Document Topic Model (DTM)

As mentioned earlier, there probably would be word usage mismatch between a query and a spoken document, even if they are topically related to each other. Therefore, instead of constructing the document model based on the literal term information, we can exploit probabilistic topic models to represent each spoken document through a latent topic space (Blei *et al.*, 2010). In this spectrum of research, each document D is regarded as a document topic model (DTM), consisting of a set of K shared latent topics $\left\{T_1,\ldots,T_k,\ldots,T_K\right\}$ with document-specific weights $P\left(T_k \middle| \mathrm{M}_D\right)$, where each topic T_k in turn offers a unigram distribution $P\left(w_i \middle| T_k\right)$ for observing an arbitrary word of the language. For example, in the PLSA model, the probability of a word w_i generated by a document D is expressed by:

$$P_{\mathrm{PLSA}}\left(w_i \middle| \mathrm{M}_D\right) = \sum_{k=1}^{K} P\left(w_i \middle| T_k\right) P\left(T_k \middle| \mathrm{M}_D\right). \tag{4}$$

The key idea we wish to illustrate here is that, for PLSA, 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. A document is believed to be more relevant to the query if it has higher weights on some topics and the query words also happen to appear frequently in these topics.

In the practical implementation of PLSA, 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 $P_{\mathrm{PLSA}}\left(w_i \middle| \mathrm{M}_D\right)$ of all words w_i observed in the document collection, or, more specifically, the total likelihood of all documents generated by their own DTM models:

$$\begin{aligned} L_{\mathrm{PLSA}} &= \prod_{D \in \mathbf{D}} P_{\mathrm{PLSA}}\left(D \middle| \mathrm{M}_D\right) \\ &= \prod_{D \in \mathbf{D}} \prod_{w_i \in D} P_{\mathrm{PLSA}}\left(w_i \middle| \mathrm{M}_D\right)^{c\left(w_i, D\right)}. \end{aligned} \tag{5}$$

We can first use the *K*-means algorithm to partition the entire document collection into K topical classes. Hence, the initial topical unigram distribution $P(w_i|T_k)$ for a topical cluster can be estimated according to the underlying statistical characteristics of the document being assigned to it and the probabilities for each document generating the topics, *i.e.*, $P(T_k|\mathrm{M}_D)$, are measured according to its proximity to the centroid of each respective cluster. Then, (5) can be iteratively optimized by the following three expectation-maximization (EM) (Dempster *et al.*, 1977) updating equations:

- **E (Expectation) Step**

$$P(T_k \mid w_i, \mathrm{M}_D) = \frac{P(w_i \mid T_k) P(T_k \mid \mathrm{M}_D)}{\sum_{T_k'} P(w_i \mid T_k') P(T_k' \mid \mathrm{M}_D)}, \tag{6}$$

- **M (Maximization) Step**

$$\hat{P}(w_i \mid T_k) = \frac{\sum_D c(w_i, D) P(T_k \mid w_i, \mathrm{M}_D)}{\sum_w \sum_D c(w, D) P(T_k \mid w, \mathrm{M}_D)}, \tag{7}$$

$$\hat{P}(T_k \mid \mathrm{M}_D) = \frac{\sum_w c(w, D) P(T_k \mid w, \mathrm{M}_D)}{\sum_{w'} c(w', D)}, \tag{8}$$

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

On the other hand, LDA, having a formula analogous to PLSA for document ranking, is regarded as a generalization of PLSA and has enjoyed considerable success in a wide variety of natural language processing (NLP) tasks. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes the model parameters are fixed and unknown; while LDA places additional *a priori* constraints on the model parameters, *i.e.*, thinking of them as random variables that follow Dirichlet distributions. In other words, the total log-likelihood of all documents generated by LDA models is defined as:

$$L_{\mathrm{LDA}} = \int\int \prod_{z=1}^{K} P(\varphi_z \mid \beta) \prod_{D \in \mathbf{D}} p(\theta_D \mid \alpha) \left(\prod_{i=1}^{|D|} \sum_{k=1}^{K} P(w_i \mid T_k, \varphi_z) P(T_k \mid \theta_D) \right) d\theta d\varphi \tag{9}$$

where θ_d and φ_z are multinomial distributions with Dirichlet parameter α and β, respectively, and $|D|$ is the number of words in the document D. LDA possesses fully consistent generative semantics by treating the topic mixture distribution as a K-parameter hidden random variable rather than a large set of individual parameters that are explicitly linked to the training set (Blei *et al.*, 2003). Compared to PLSA, LDA overcomes the problem of overfitting and the problem of generating new documents incurred by PLSA.

Since LDA has a more complex form for model optimization, which is difficult to be solved by exact inference, several approximate inference algorithms, such as the variational Bayes approximation (Blei *et al.*, 2003), the expectation propagation method (Ypma *et al.*, 2002), and the Gibbs sampling algorithm (Griffiths, 2004), have been proposed in the literature for estimating the model parameters of LDA. In this paper, we adopt the Gibbs sampling algorithm, where θ and φ are marginalized out and only the latent variables T_k are sampled, to infer the model parameters. Then, the probability of a word w_i generated by a document D in the LDA model is expressed by:

$$P_{\mathrm{LDA}}\left(w_i \middle| \hat{\phi}, \hat{\theta}, \mathrm{M}_D\right) = \sum_{k=1}^{K} P\left(w_i \middle| T_k, \hat{\phi}\right) P\left(T_k \middle| \hat{\theta}, \mathrm{M}_D\right), \tag{10}$$

where $\hat{\varphi}$ and $\hat{\theta}$ are the posterior estimates of θ and φ, respectively. We refer the readers to Griffiths and Steyvers (2004) for a better understanding of the detailed inference procedure.

2.3 Word Topic Model (WTM)

Rather than treating each document in the collection as a document topic model, we can regard each word w_j of the language as a word topic model (WTM). To get to this point, all words are assumed to share the same set of latent topic distributions but have different weights over these topics. The WTM model of each word w_j for predicting the occurrence of a particular word w_i can be expressed by:

$$P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right), \tag{11}$$

where $P\left(w_i \middle| T_k\right)$ and $P\left(T_k \middle| \mathrm{M}_{w_j}\right)$ are 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}, respectively. Then, each document naturally can be viewed as a composite WTM, while the relevance measure between a word w_i and a document D can be expressed by:

$$P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_D\right) = \sum_{w_j \in D} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P_{\mathrm{ULM}}\left(w_j \middle| \mathrm{M}_D\right), \tag{12}$$

The resulting composite WTM model for D, in a sense, can be thought of as a kind of language model for translating words in D to w_i.

The model parameters of WTM can be inferred by unsupervised training as well. More precisely, each WTM model M_{w_j} can be trained by concatenating those words occurring in the vicinity of (or a context window of size S around) each occurrence of w_j, which are postulated to be relevant to w_j, to form a relevant observation sequence O_{w_j} for training M_{w_j}. The words in O_{w_j} are also assumed to be conditionally independent, given M_{w_j}.

Therefore, the WTM models of the words in the vocabulary set **w** can be estimated by maximizing the total likelihood of their corresponding relevant observation sequences generated by themselves:

$$L_{\mathrm{WTM}} = \prod_{w_j \in \mathbf{w}} P_{\mathrm{WTM}}\left(O_{w_j} \middle| \mathrm{M}_{w_j}\right) = \prod_{w_j \in \mathbf{w}} \prod_{w_i \in O_{w_j}} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right)^{c\left(w_i, O_{w_j}\right)}, \tag{13}$$

Then, the parameters of each WTM model can be estimated using the following EM updating formulae:

- **E (Expectation) Step**

$$P\left(T_k \mid w_i, \mathrm{M}_{w_j}\right) = \frac{P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)}{\sum_{T_k'} P\left(w_i \mid T_k'\right) P\left(T_k' \mid \mathrm{M}_{w_j}\right)}, \tag{14}$$

- **M (Maximization) Step**

$$\hat{P}\left(w_i \mid T_k\right) = \frac{\sum_{w_j \in \mathbf{w}} c\left(w_i, O_{w_j}\right) P\left(T_k \mid w_i, \mathrm{M}_{w_j}\right)}{\sum_{w_l \in \mathbf{w}} \sum_{w_n \in O_{w_l}} c\left(w_n, O_{w_l}\right) P\left(T_k \mid w_n, \mathrm{M}_{w_l}\right)}, \tag{15}$$

$$\hat{P}\left(T_k \mid \mathrm{M}_{w_j}\right) = \frac{\sum_{w \in O_{wj}} c\left(w, O_{wj}\right) P\left(T_k \mid w, \mathrm{M}_{w_j}\right)}{\sum_{w'} c\left(w', O_{wj}\right)}. \tag{16}$$

Along a similar vein to the LDA model, word Dirichlet topic model (WDTM) can be derived as well. WDTM essentially has the same ranking formula as WTM, except that it further assumes the model parameters are governed by some Dirichlet distributions.

2.4 Analytic Comparisons between DTM and WTM

DTM (PLSA or LDA) and WTM (WTM or WDTM) can be analyzed from several perspectives. First, DTM models the co-occurrence relationship between words and documents, while WTM models the co-occurrence relationship between words in the collection. More explicitly, we may compare DTM and WTM through nonnegative (or probabilistic) matrix factorizations, as depicted in Figure 1. For DTM models, each column of Matrix **A** denotes the probability vector of a document in the collection, which offers a probability for every word occurring in the document. For WTM models, each column of Matrix **B** is the probability vector of a word's vicinity, which offers a probability for observing every other word occurring in its vicinity. Both Matrices **A** and **B** can be decomposed into two matrices standing for the topic mixture components and the topic mixture weights, respectively.

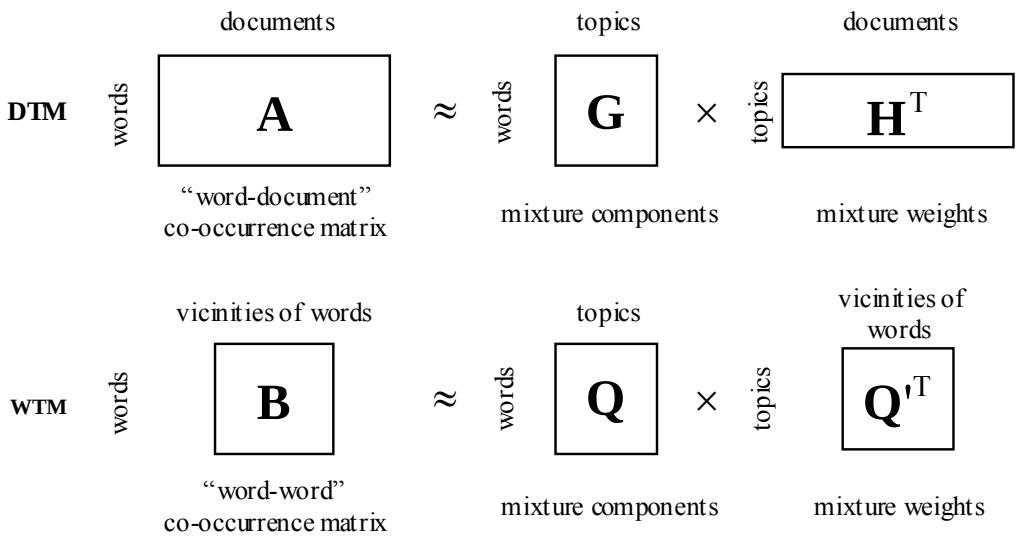

Figure 1. A schematic illustration for the matrix factorizations of DTM and WTM.

Furthermore, the topic mixture weights of DTM for a new document have to be estimated online using EM or other more sophisticated algorithms, which would be time-consuming; on the contrary, the topic mixture weights of WTM for a new document D can be obtained on the basis of the topic mixture weights of all words involved in the document without using a complex inference procedure.

Finally, if the context window for modeling 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 $1 < K < V$). Thus, with some appropriate values of S and K being chosen, we can show that WTM seems to be a good method of approximating the bigram or skip-bigram models for sparse data (Chen, 2009).

3. Extensions of Topic Modeling

3.1 Hybrid of DTM and WTM

As mentioned in the previous section, DTM and WTM are different from each other in their fundamental premises to determine a hidden topical decomposition of the document collection through the exploration of the topical information underlying the "word-document" or "word-word" co-occurrence relationships, respectively. Thus, we may fuse the results of the two different topical decompositions from DTM and WTM together for better ranking of spoken documents.

One possible method is to train each of these two models individually and linearly combine their respective document-ranking scores in the log-likelihood domain subsequently (called "Individual Topics" hereafter). Nevertheless, this approach could not arrive at the same set of topic components (*i.e.*, $P(w_i|T_k)$, $k = 1,\ldots,K$) that are potentially associated with the spoken document collection. Alternatively, we may seek to conduct a single (or unique) topical decomposition of the spoken document collection by simultaneously exploiting these two types of co-occurrence relationships (called "Shared Topics" hereafter). This approach tries to estimate the DTM and WTM model parameters by jointly maximizing the total likelihood of words occurring in the spoken documents and the total likelihood of the words occurring in the vicinities of arbitrary words in the vocabulary. A pictorial representation for the probabilistic matrix decomposition of the spoken document collection with this approach is illustrated in Figure 2, where each column of the left hand side matrix denotes either the probability vector of a document in the collection, which offers a probability for every word occurring in the document (*i.e.*, DTM), or the probability vector of the vicinity of a word in the vocabulary, which offers a probability for observing every other word occurring in the vicinity (*i.e.*, WTM). Then, this matrix can be decomposed into two matrices standing for the topic mixture components (*i.e.*, $\mathbf{F}$) and the topic mixture weights (*i.e.*, $\mathbf{H}$ and $\mathbf{Q}'$), respectively.

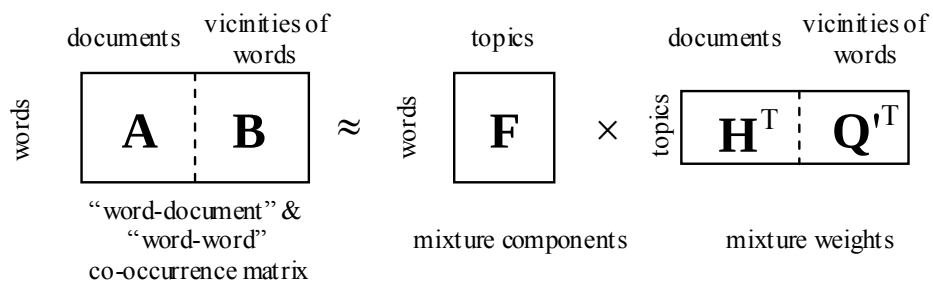

Figure 2. A schematic illustration for the matrix factorization of hybrids of DTM and WTM.

3.2 Topic Modeling with Subword-level Units

In this paper, we also investigate leveraging subword-level information cues for topic modeling in Chinese SDR. To do this, syllable pairs are taken as basic units for indexing instead of words. In the following paragraphs, we will elucidate the reasons for using syllable-level features for the retrieval purpose before describing how they can be integrated into the DTM and WTM models.

Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio if the differences in tones are disregarded. On the other hand, an inventory of about 13,000 characters provides full textual coverage of written Chinese. Each word is composed of one or more characters, and each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are generated easily by combining a few characters. Such new words also include many proper nouns, like personal names, organization names, and domain-specific terms. The construction of words from characters is often quite flexible. One phenomenon is that different words describing the same or similar concepts can be constructed of slightly different characters. Another phenomenon is that a longer word can be arbitrarily abbreviated into a shorter word. Moreover, there is a many-to-many mapping between characters and syllables; a foreign word can be translated into different Chinese words based on its pronunciation, while different translations usually have some syllables in common, or may have exactly the same syllables. Statistical evidence also shows that, in the Chinese language, about 91% of the top 5,000 most frequently used polysyllabic words are bi-syllabic, i.e., they are pronounced as a segment of two syllables. Therefore, such syllable segments (or syllable pairs) definitely carry a plurality of linguistic information and make great sense to be used as important index terms.

The characteristics of the Chinese language mentioned above lead to some special considerations for SDR. Word-level index features possess more semantic information than syllable-level ones; thus, word-based retrieval enhances the precision. On the other hand, syllable-level index features are more robust against the Chinese word tokenization ambiguity, Chinese homophone ambiguity, open vocabulary problem, and speech recognition errors; therefore, the syllable-level information would enhance the recall. Accordingly, there is good reason to fuse the information obtained from index features of different levels. It has been shown that using syllable pairs as the index terms is very effective for Chinese SDR, and the retrieval performance can be further improved by incorporating the information from word-level index features.

In this paper, both the manual transcript and the recognition transcript of each spoken document, in the form of a word stream, were automatically converted into a stream of overlapping syllable pairs. Then, all of the distinct syllable pairs occurring in the spoken document collection were identified to form an indexing vocabulary of syllable pairs. Topic modeling with the syllable-level information can be fulfilled in two ways. One is to simply use syllable pairs, as a replacement for words, to represent the spoken documents and to construct the associated probabilistic latent topic distributions for DTM and WTM accordingly. The other is to jointly utilize both words and syllable pairs, as two types of index terms, to represent the spoken documents, as well as to construct the associated probabilistic latent topic

distributions. To this end, each spoken document is represented virtually with a spliced text stream, consisting of both words and syllable pairs. Figure 3 takes DTM as an example to graphically illustrate such an attempt, which is expected to discover correlated topic patterns of the spoken document collection when using both word- and syllable-level index features simultaneously.

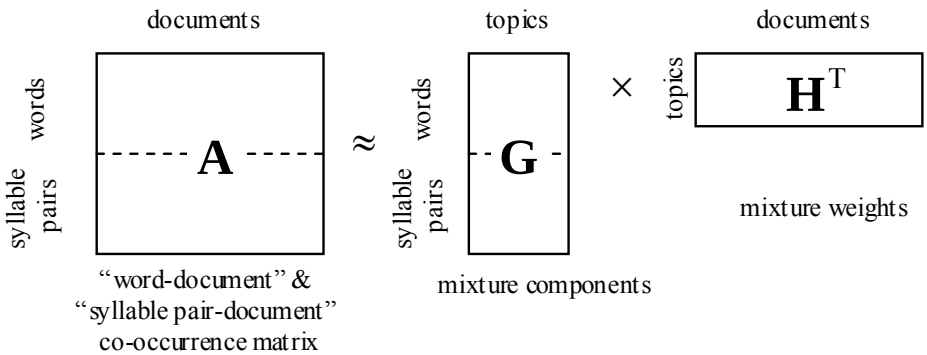

Figure 3. A schematic illustration for the matrix factorization of DTM, jointly using words and syllable pairs as the index terms.

4. Experimental Setup

4.1 Corpus and Evaluation Metric

We used the Topic Detection and Tracking (TDT-2) collection for the SDR task (LDC, 2000). 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 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 (CLSDR) in the Mandarin English Information (MEI) Project (Meng *et al.*, 2004). In this paper, we used the Mandarin Chinese collections of the TDT corpora for the retrospective retrieval task, such that the statistics for the entire document collection was obtainable. Chinese text news stories from Xinhua News Agency were compiled to form the test queries (or query exemplars). More specifically, in the following experiments, we will either use a whole text news story as "*long*" query or merely extract the title field from a text news story to form a relatively "*short*" query.

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 merely serve as the relevance judgments for performance evaluation and will not be utilized in the training of topic models (*cf.* Section 2). Table 1 shows some basic statistics about the corpus used in this paper. The Dragon large-vocabulary continuous speech

recognizer provided Chinese word transcripts for our Mandarin audio collections. To assess the performance level of the recognizer, we spot-checked a fraction of the spoken document collection set (about 40 hours), and obtained error rates of 35.38% (in word), 17.69% (in character), and 13.00% (in syllable). Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with 24,000 words extracted from Dragon's word recognition output, and used the augmented LDC lexicon (about 51,000 words) to tokenize the manual transcripts for computing error rates. We also used this augmented LDC lexicon to tokenize the text queries in the retrieval experiments.

Table 1. Statistics for TDT-2 Collections Used for Spoken Document Retrieval

# Spoken documents	2,265 stories 46.03 hours of audio			
# Distinct test queries	16 Xinhua text stories (Topics 20001~20096)			
	Min.	Max.	Med.	Mean
Document length (in characters)	23	4841	153	287
Length of long query (in characters)	183	2623	329	533
Length of short query (in characters)	8	27	13	14
# Relevant documents per test query	2	95	13	29

The retrieval results are expressed in terms of non-interpolated mean average precision (mAP) following the TREC evaluation (Harman, 1995), which is computed by the following equation:

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}}, \tag{17}$$

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.

4.2 Model Implementation

Topic models, such as DTM and WTM, introduce a set of latent topics to cluster concept-related words and match a query with a document at the level of these word clusters. Although document ranking based merely on DTM or WTM tends to increase recall, using just one of them is liable to hurt the precision for SDR. Specifically, they offer coarse-grained concept clues about the document collection at the expense of losing discriminative power

among concept-related words in finer granularity. Therefore, in this paper, when either DTM or WTM was employed in evaluating the relevance between a query Q and a document D, we additionally incorporated the unigram probabilities of a query word (or term) occurring in the document $P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_D)$ and a general text corpus $P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_C)$ with the topic model $P_{\mathrm{Topic}}(w_i \mid \mathrm{M}_D)$ (either DTM or WTM), for probability smoothing and better performance. For example, the probability of a query word generated by one specific topic model of a document (*cf.* (4), (10), and (12)) was modified as follows:

$$\begin{aligned} P(w_i \mid D) &= \alpha \cdot \left[\beta \cdot P_{\mathrm{Topic}}(w_i \mid \mathrm{M}_D) + (1-\beta) \cdot P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_D) \right] \\ &\quad + (1-\alpha) \cdot P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_C) \end{aligned} \tag{18}$$

where $P_{\mathrm{Topic}}(w_i \mid \mathrm{M}_D)$ can be the probability of a word w_i generated by PLSA or LDA (*cf.* (4) or (10)) or WTM (*cf.* (12)); the values of the interpolation weights α and β can be empirically set or further optimized by other optimization techniques (Zhai, 2008). A detailed account of this issue will be given in Section 5.2. On the other hand, the Gibbs sampling algorithm (Griffiths, 2004) is used to infer the parameters of LDA and WDTM.

5. Experimental Results

5.1 Baseline Experiments

The baseline retrieval results obtained by the ULM model are shown in Table 2. The retrieval results, assuming manual transcripts for the spoken documents to be retrieved (denoted TD, text documents) are known, are listed for reference and are compared to the results when only erroneous recognition transcripts generated by speech recognition are available (denoted SD, spoken documents). As can be seen, the performance gap between the TD and SD cases was about 7% absolute in terms of mAP when using either long or short queries, although the word error rate (WER) for the spoken document collection was higher than 35%. On the other hand, retrieval using short queries degraded the performance approximately 45% relative to retrieval using long queries. This is due to the fact that a long query usually contains a variety of words describing similar concepts. Even though some of these words might not be correctly transcribed in the relevant spoken documents, they, in the ensemble, still provide plenty of clues for literal term matching. From now on, unless otherwise stated, we will only report the retrieval results for the SD case.

Table 2. Baseline retrieval results (in mAP) achieved by ULM.

Query Type	TD	SD
Long	0.639	0.562
Short	0.370	0.293

5.2 Experiments on DTM and WTM

In the next set of experiments, we assessed the utility of various topic models for SDR, including PLSA, LDA, and WTM, as well as WDTM. The corresponding retrieval results are shown in Table 3. It is worth mentioning that all of these topic models were trained without supervision and had the same number of latent topics, which was set to 32 in this study. A detailed analysis for the impact of the model complexity of PLSA and WTM on SDR performance can be found in Chen (2009). On the other hand, both WTM and WDTM had the same context window size S set to 21. Since this project set out to investigate the effectiveness of various topic models for SDR, the interpolation weights α and β defined in (18) were optimized for each respective topic model with a two-dimensional grid search over the range from 0 to 1 and in increments of 0.1. Consulting Table 3, we find that all of these topic models give moderate but consistent improvement over the baseline ULM model when long queries are evaluated. One possible explanation is that the information need already might have been stated fully in a long query, whereas additional incorporation of the topical information into the document language model does not seem to offer many extra clues for document ranking. On the contrary, the retrieval performance receives great boosts from the additional use of the topical information when the queries are short. This implies that incorporating the topical information with the literal term information for document modeling is especially useful when the query is inadequate to address the information need.

Table 3. Spoken document retrieval results achieved by various topic models.

Method	Long Query	Short Query
ULM	0.562	0.293
PLSA	0.569	0.374
LDA	0.590	0.407
WTM	0.573	0.351
WDTM	0.574	0.377
LDA+WDTM (Individual Topics)	0.592	0.418
LDA+WDTM (Shared Topics)	0.595	0.415

We then turned our attention to compare the following topic models. 1) LDA outperforms PLSA, and WDTM outperforms WTM. This finding supports the argument that constraining the latent topic distributions with Dirichlet priors will lead to better model estimation. 2) LDA is the best among these topic models. As compared to the baseline ULM model, it yielded about 5% and 39% relative improvements for long and short queries, respectively. Moreover, we investigated the effectiveness of the fusion of DTM and WTM to the retrieval performance (*cf.*, the last two rows of Table 3). Here, we took LDA and WDTM

as the training example since they achieved better retrieval performance in the previous experiment. It is also worth mentioning that the row "LDA+WDTM (Individual Topics)" shown in Table 3 indicates that each topic model was trained individually and their respective document-ranking scores were combined in the log-likelihood domain. On the contrary, the row "LDA+WDTM (Shared Topics)" in Table 3 denotes the hybrid of DTM and WTM in both model training and testing (*cf.* Section 3.1). As is evident, the fusion of LDA and WDTM (*i.e.*, with either individual sets of topics or a shared set of topics) is beneficial to the retrieval performance. This provides an additional 1% absolute improvement for the case of using short queries, as compared to that using LDA alone. Nevertheless, the joint exploration of "word-document" and "word-word" latent topic information (*i.e.*, with a shared set of topics) in the training phrase does not provide any added benefit compared to the results obtained by training LDA and WDTM individually (*i.e.*, with individual sets of topics). This is an interesting phenomenon and awaits further exploration. Readers may refer to Chen, *et al.* (2010) for an attempt that applies a similar idea to the speech recognition task.

To go a step further, we attempted to investigate the more subtle interaction effects among the topic model $P_{\text{Topic}}\left(w_i|\text{M}_D\right)$, the document model $P_{\text{ULM}}\left(w_i|\text{M}_D\right)$, and the background model $P_{\text{ULM}}\left(w_i|\text{M}_C\right)$ in (18) by varying the values of the interpolation weights α and β. Here, LDA was taken as an example topic model since it exhibits the best performance among the topic models compared in this paper. The retrieval results are graphically illustrated in Figure 4, where the horizontal and vertical axes denote the values of α and β, respectively. As seen in the results revealed in Figure 4, additional incorporation of $P_{\text{ULM}}\left(w_i|\text{M}_D\right)$ and $P_{\text{ULM}}\left(w_i|\text{M}_C\right)$ into LDA is beneficial for retrieval. In an extreme case, when both the values of α and β are set to one, as shown in the top right corner of Figure 4, the retrieval model is based merely on the topical information, which has poor retrieval performance, especially for the case using long queries. One possible reason is that a long query may contain several common non-informative words and using the topical information alone will let the query become biased away from representing the true theme of the information need, probably due to these non-informative words. This argument again can be verified by examining the rightmost columns of Figure 4, where using the background model $P_{\text{ULM}}\left(w_i|\text{M}_C\right)$ can absorb the contributions of the common (or non-informative) words made to document ranking, thus giving better retrieval performance.

Looking at each row of Figure 4, we see that smoothing LDA with the document model $P_{\text{ULM}}\left(w_i|\text{M}_D\right)$ is also useful. This is attributed to the fact that discriminative (or informative) words will occur repeatedly in a specific document; $P_{\text{ULM}}\left(w_i|\text{M}_D\right)$ gives more emphasis on these words. On the other hand, Figure 4 also reflects that smoothing LDA with the background model $P_{\text{ULM}}\left(w_i|\text{M}_C\right)$ is necessary when the query is long, but it does not seem to be helpful for the case of using a relatively short query. This is mainly because the

information need stated by the short query is already concise, and the importance of the role that $P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_C)$ plays in filtering out or deemphasizing common (or non-informative) words is less pronounced.

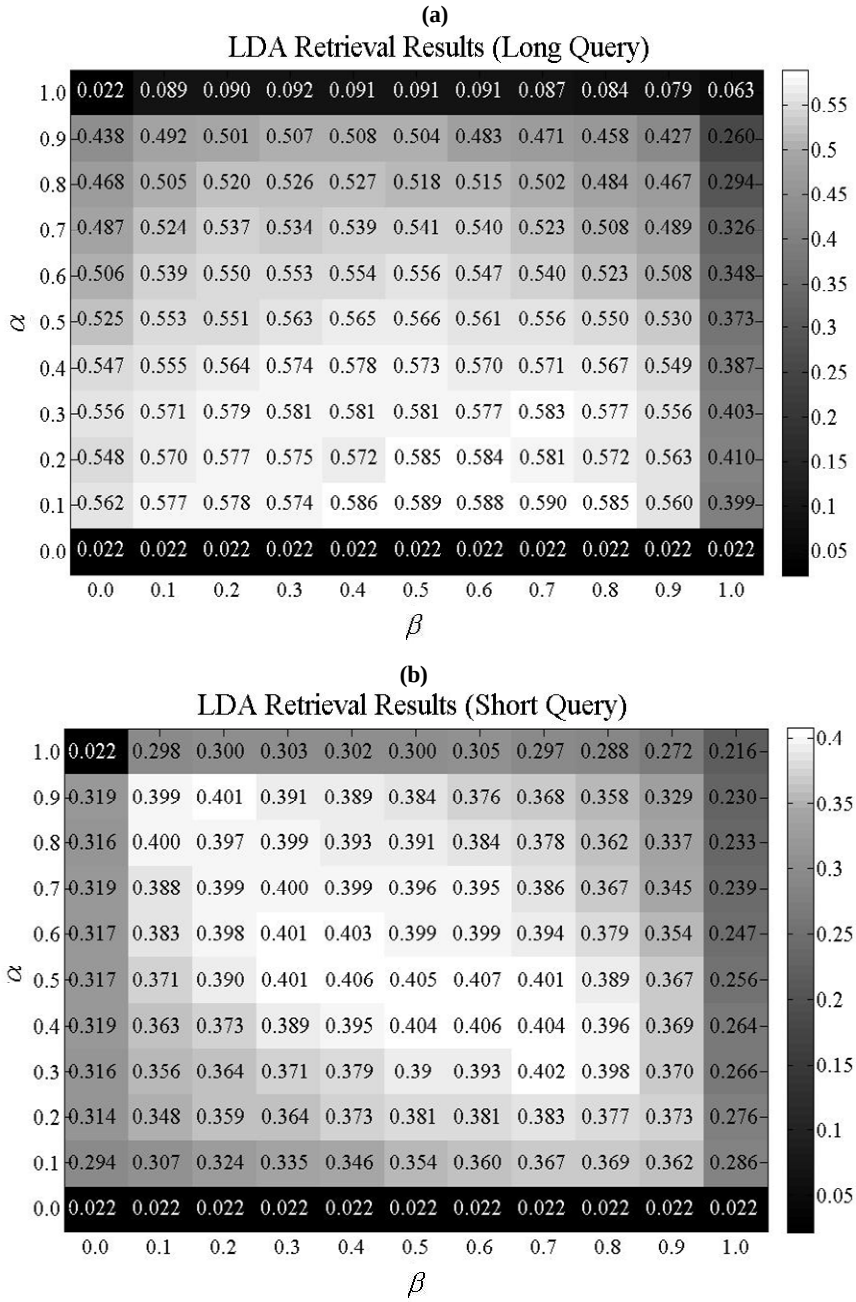

Figure 4. Detailed spoken document retrieval results achieved by LDA with respect to different types of queries.

5.3 Experiments on using Subword-level Index features

In the fourth set of experiments, we evaluated the performance of the topic models when syllable pairs were utilized instead as the index terms. Here, we took LDA and WDTM as the example topic models, and the corresponding models are denoted by Syl_LDA and Syl_WDTM, respectively. The fusion of words and syllable pairs for topic modeling was investigated as well. Notice that Word_LDA denotes LDA using words as the index terms, which was termed LDA in the previous sections.

The retrieval results of Syl_LDA and Syl_WDTM are shown in Table 4, where the results achieved by ULM and using syllable pairs as the index terms (denoted by Syl_ULM) are also depicted for comparison. Several observations can be made from Table 4. First, the topic models (Syl_LDA and Syl_WDTM) again are superior to the unigram language model when the syllable-level information is used in place of the word-level information (denoted by Syl_ULM). Syl_LDA results in absolute improvements of about 8% and 3% over Syl_ULM when evaluated using the long and short queries, respectively. Second, the topic models with the syllable-level information perform worse than those with the word-level information. This may be due simply to the fact that syllable pairs are not as good as words in representing the semantic content of the queries and the documents. Third, the fusion of the word- and syllable-information for topic modeling (each topic model was trained individually beforehand) demonstrates much better retrieval results (*cf.* the last two rows of Table 4) as compared to that of the topic models with merely the word-level information (*cf.* Table 3).

Table 4. Spoken document retrieval results achieved by LDA and WDTM, respectively, using syllable pairs along with the combination of words and syllable pairs.

Method	Long Query	Short Query
Syl_ULM	0.492	0.274
Syl_LDA	0.571	0.302
Syl_WDTM	0.536	0.299
Word_LDA+Syl_LDA	0.613	0.412
Word_WDTM+Syl_WDTM	0.575	0.383

Finally, we examined the contributions made by modeling the correlated topic patterns of the spoken document collection when jointly using words and syllable pairs in the construction of the latent topic distributions. We took the LDA model as an example to study the effectiveness of such an attempt, and the associated results are shown in Table 5. The results reveal that, when only syllable pairs are used as the index terms for the final document ranking, modeling the correlated topic patterns, namely, jointly using words and syllable pairs

in the construction of the latent topic distributions for LDA (denoted by Syl_LDA (Corr.)) is better than that only using syllable pairs to construct the latent topic distributions (denoted by Syl_LDA). On the other hand, such an attempt slightly hurts the performance of LDA using words for the final document ranking (denoted by Word_LDA (Corr.)). This phenomenon seems to be reasonable because the semantic meanings carried by words would probably see interference from syllable pairs when we attempt to splice these two distinct index term streams together for constructing the latent topic distributions of LDA. It can be observed that Syl_LDA (Corr.) significantly outperforms all other topic models in the case of using long queries (*cf.* Tables 3, 4, and 5). This demonstrates the potential benefit of using the syllable-level information in topic modeling for SDR if we can carefully delineate the syllable-level information. Nevertheless, in the case of using short queries, Syl_LDA (Corr.) does not perform as well as LDA using words as the index terms to construct the latent topic distributions (denoted by Word_LDA). We conjecture that one possible reason is that the topical information inherent in a short query cannot be unambiguously depicted with limited syllable pairs. In order to mitigate this deficiency, we combined Word_LDA with Syl_LDA (Corr.) to form a new retrieval model (denoted by Word_LDA + Syl_LDA (Corr.)), which yields the best results of 0.636 and 0.431 for long and short queries, respectively. One should keep in mind that these results were obtained using the erroneous speech transcripts of the spoken documents (*i.e.*, the SD case). This also reveals that Word_LDA + Syl_LDA (Corr.) can make retrieval using the speech transcripts achieve almost the same performance as ULM using the manual transcripts (*i.e.*, the TD case) when the queries are long, and can perform even better than the latter for short queries.

Table 5. Spoken document Retrieval results achieved by correlated LDA, using words (Word_LDA(Corr.)), syllable pairs (Syl_LDA(Corr.)), and their combination (Word_LDA + Syl_LDA(Corr.)).

Method	Long Query	Short Query
Word_LDA (Corr.)	0.577	0.349
Syl_LDA (Corr.)	0.618	0.356
Word_LDA+Syl_LDA (Corr.)	0.636	0.431

6. Conclusions

In this paper, we have investigated the utility of two categories of topic models, namely, the document topic models (DTM) and the word topic models (WTM), for SDR. Moreover, we have leveraged different levels of index features for topic modeling, including words, syllable pairs, and their combinations, so as to prevent the performance degradation facing most SDR tasks. The proposed models indeed demonstrated significant performance improvements over

the baseline model on the Mandarin SDR task. Our future research directions include: 1) training the topic models in a lightly supervised manner through the exploration of users' click-through data, 2) investigating discriminative training of topic models, 3) integrating the topic models with the other more elaborate representations of the speech recognition output (Yi and Allan, 2009; Chelba *et al.*, 2008) for larger-scale SDR tasks, and 4) utilizing speech summarization techniques to help estimate better document models and topic models.

Acknowledgement

This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 101-2221-E-003 -024 -MY3, NSC 99-2221-E-003-017-MY3, NSC 98-2221-E-003-011-MY3, NSC 100-2515-S-003-003, and NSC 99-2631-S-003-002.

Reference

Blei, D.M., Ng, A.Y., & Jordan, M. I., (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D. & Lafferty, J., (2009). Topic models. In A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Blei, D., Carin, L., & Dunson, D., (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55-65.

Chelba, C., Hazen, T. J., & Sarclar, M., (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, 25(3), 39-49.

Chen, B., (2009). Word topic models for spoken document retrieval and transcription. *ACM Transactions on Asian Language Information Processing*, 8(1), Article 2.

Chia, T. K., Sim, K. C, Li, H. Z. & Ng, H. T., (2008). A lattice-based approach to query-by-example spoken document retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 363-370.

Dempster, A. P., Laird, N. M., & Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39 (1): 1-38.

Garofolo, J., Auzanne, G., & Voorhees, E., (2000). The TREC spoken document retrieval track: A success story. In *Proceeding the 8th Text REtrieval Conference*. NIST, 107-129.

Griffiths, T. L. & Steyvers, M., (2004). Finding scientific topics. In *Proceeding of the National Academy of Sciences*, 5228-5235.

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B., (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244.

Harman D., (1995). Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proceeding the Fourth Text Retrieval Conference*, 1-23.

Hofmann, T., (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.

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

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

Meng, H., Chen, B., Khudanpur, S., Levow, G. A., Lo, W. K., Oard, D., Schone, P., Tang, K., Wang, H. M., & Wang, J., (2004). Mandarin-English information (MEI): investigating translingual speech retrieval. *Computer Speech and Language*, 18(2), 163-179.

Miller, D. R. H., Leek, T., & Schwartz, R., (1999). A hidden Markov model information retrieval system. In *Proceeding ACM SIGIR Conference on R&D in Information Retrieval*, 214-221.

Ponte, J. M. & Croft, W. B., (1998). A language modeling approach to information retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 275-281.

Wei, X., & Croft, W. B., (2006). LDA-based document models for ad-hoc retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 178-185.

Lin, S. H. & Chen B., (2009). Topic modeling for spoken document retrieval using word- and syllable-level information. In *Proceedings of the third workshop on Searching spontaneous conversational speech*, 3-10.

Chen, K. Y., Chiu, H. S. & Chen B., (2010). Latent topic modeling of word vicinity information for speech recognition. In *Proceeding of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5394-5397.

Ypma, J., Basten, T. & Lafferty, J., (2002). Expectation-propagation for the generative aspect model. In *Proceeding Conference on Uncertainty in Artificial Intelligence*, 352-359.

Zhai, C. X., (2008). Statistical language models for information retrieval (Synthesis Lectures Series on Human Language Technologies). Morgan & Claypool Publishers.