Topic Modeling for Spoken Document Retrieval using Word- and Syllable-level Information

Shih-Hsiang Lin and Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University
{shlin, berlin}@csie.ntnu.edu.tw

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 article, we first present a comprehensive comparison among various topic modeling approaches, including the so-called document topic models (DTM) and word topic models (WTM), for Chinese spoken document retrieval (SDR). Moreover, in order to lessen SDR performance degradation when using imperfect recognition transcripts, we also leverage different levels of indexing features for topic modeling, including words, syllable-level units and their combinations. All the experiments are performed on the TDT Chinese collection.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—Indexing Methods

General Terms: Algorithms, Design

Keywords: Information Retrieval, Document Topic Models, Word Topic Models, Spoken Document Retrieval, Speech Recognition

1. INTRODUCTION
Statistical language modeling (LM), 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 for a vast array of speech and language processing tasks. This statistical paradigm was first introduced for building information retrieval (IR) systems in the late 1990s [1-2], indicating very good potential, and has motivated many follow-up studies and extensions [3-4]. Typically, these approaches attempt to build a probabilistic language model explicitly for each individual document in the collection. The basic idea is that a document is deemed to be relevant to a query if its corresponding document language model is more likely to generate the query.

SSCS'09, October 23, 2009, Beijing, China.

In practice, the relevance measure for the LM approaches is usually computed by two different matching strategies, namely, literal term matching and concept matching [5]. The unigram language model (ULM) is the most popular example for literal term matching [2, 4]. In this 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). However, most of these approaches would suffer from the problems of word usage diversity, which might 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 explore the latent topic information conveyed in the query and documents, based on which the retrieval is performed; the probabilistic latent semantic analysis (PLSA) [3] and the latent Dirichlet allocation (LDA) [6] are often considered two basic representatives of this category. They both introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics. 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 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. On the other hand, instead of treating each document as a whole as a document topic model (DTM), such as PLSA and LDA, the word topic model (WTM) [7] 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 [8-10] for a comprehensive 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, or redundant information. A straightforward remedy, apart from the many approaches 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 [11-12]. 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 [13]. The reason for fusion of word- and subword-level information is that, incorrectly recognized spoken words often include several subword units correctly recognized, and the retrieval based on subword-level representations hence may take advantage of partial matching. Nevertheless, most retrieval systems participated in the TREC-SDR evaluations had claimed that speech recognition errors do not seem to cause much adverse effect on SDR performance when merely using imperfect recognition transcripts derived from one-best recognition results [11, 14]. 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 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 short and there exists severe deviation in word usage between the queries and documents.

With the above inspiration in mind, in this article, we first compare the structural characteristics of various topic models for Chinese SDR, including PLSA, LDA, as well as WTM and its extension. The utility of these models is thoroughly verified by using both long and short test queries. Moreover, we also leverage different levels of indexing features, including words, subword units and their combinations so as to alleviate SDR performance degradation caused by imperfect recognition transcripts. To our knowledge, there is little literature on using subword units in topic modeling for SDR. The rest of this article is structured as follows. Section 2 elucidates the structural characteristics of the different types of topic models for the retrieval purpose. Section 3 describes the spoken document collection used in this article, as well as the experimental setup. A series of experiments and associated discussions are presented in Section 4. Finally, Section 5 concludes this article with 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) [3] and the latent Dirichlet model (LDA) [6, 15], followed by an introduction to our recently proposed word topic model (WTM) [7]. We also present an extension of WTM, named word Dirichlet topic model (WDTM) in this study.

2.1 Probabilistic Generative Framework

When language modeling approaches are applied to IR, they basically use a probabilistic generative framework for ranking each document D in the collection given a query Q, which can be expressed by $P(D|Q)$. This ranking criterion can be approximated by the likelihood of Q generated by D, i.e., $P(Q|D)$. To do this, each document is treated as a probabilistic language model for generating the query. 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 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:

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

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 $P(w_i|D)$.

The simplest way to construct $P(w_i|D)$ 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., $P(w_i|\mathrm{M}_D)$, which is estimated on the basis of the words occurring in the document and is further smoothed by a unigram distribution estimated from a general collection, i.e., $P(w_i|\mathrm{M}_C)$, to avoid the problem of zero probability:

$$P_{\mathrm{ULM}}(w|D) = \lambda \cdot P(w_i|\mathrm{M}_D) + (1-\lambda) \cdot P(w_i|\mathrm{M}_C), \tag{2}$$

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. In the following, $P(w_i|\mathrm{M}_D)$ and $P(w_i|\mathrm{M}_C)$ will be termed the document literal term model and the background model, respectively.

2.2 Document Topic Model (DTM)

Each document D is regarded as a document topic model (DTM), consisting of a set of K shared latent topics $\{T_1,\ldots,T_k,\ldots,T_K\}$ with document-specific weights $P(T_k|\mathrm{M}_D)$, where each topic T_k in turn offers a unigram distribution $P(w_i|T_k)$ 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

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

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 $P_{\mathrm{PLSA}}(w_i|\mathrm{M}_D)$ 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, using the Expectation-Maximization (EM) training algorithm:

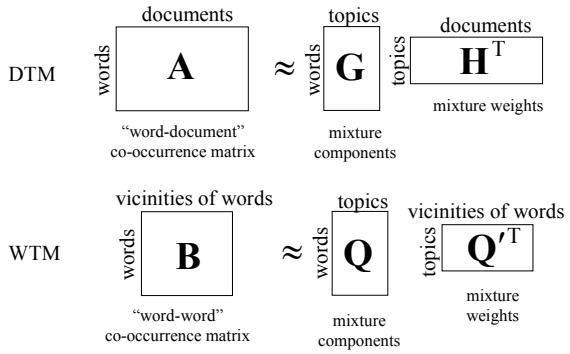

Figure 1. A schematic illustration for the matrix factorizations of DTM and WTM.

$$\log L_{\mathbf{D}} = \sum_{D \in \mathbf{D}} \log P_{\text{PLSA}}\left(D \middle| \mathrm{M}_D\right) = \sum_{D \in \mathbf{D}} \sum_{w_i \in D} c\left(w_i, D\right) \log P_{\text{PLSA}}\left(w_i \middle| \mathrm{M}_D\right). \quad (4)$$

On the other hand, LDA, having a formula analogous to PLSA (cf. Eq. (3)) for document ranking, is regarded as an extension to PLSA and has enjoyed much success for various text IR 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 some Dirichlet distributions. Since LDA has a more complex form for model optimization, which is hardly to be solved by exact inference. Several approximate inference algorithms, such as the variational Bayes approximation [6], the expectation propagation method [16] and the Gibbs sampling algorithm [17], hence have been proposed for estimating the parameters of LDA.

2.3 Word Topic Model (WTM)

Instead of 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 a 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

$$P_{\text{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right), \quad (4)$$

where $P\left(w_i \middle| T_k\right)$ and $P\left(T_k \middle| \mathrm{M}_{w_j}\right)$, respectively, 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}. 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

$$P_{\text{WTM}}\left(w_i \middle| \mathrm{M}_D\right) = \sum_{w_j \in D} P_{\text{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P\left(w_j \middle| D\right), \quad (5)$$

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 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, 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 $\mathbf{w}$ can be estimated by maximizing the total log-likelihood of their corresponding relevant observation sequences respectively generated by themselves:

$$\log L_{\mathbf{w}} = \sum_{w_j \in \mathbf{w}} \log P_{\text{WTM}}\left(O_{w_j} \middle| \mathrm{M}_{w_j}\right) = \sum_{w_j \in \mathbf{w}} \sum_{w_i \in O_{w_j}} c\left(w_i, O_{w_j}\right) \log P_{\text{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right). \quad (6)$$

Along a similar vein, in this article we propose a new topic model, named word Dirichlet topic model (WDTM). 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 Comparison 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, each column of the 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, 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 respectively standing for the topic mixture components and the topic mixture weights.

Second, 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 simply obtained on the basis of the topic mixture weights of all words involved in the document without using any 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 way to approximate the bigram or skip-bigram models for sparse data.

2.5 Topic Models with Subword-level Units

In this article, we also propose to leverage subword-level information in topic modeling for Chinese SDR. To do this, syllable pairs are taken as the basic units for indexing besides

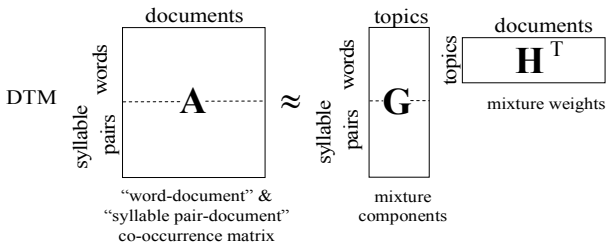

Figure2. A schematic illustration for the matrix factorization of DTM, jointly using words and syllable pairs as the indexing terms.

words. Below we will first elucidate the reasons that motivate us to use syllable-level features for the retrieval purpose, and then detail 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 easily generated 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 very often quite flexible. One phenomenon is that different words describing the same or similar concepts can be constructed by 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 indexing terms.

The characteristics of the Chinese language mentioned above hence lead to some special considerations for the spoken document retrieval task. Word-level indexing features possess more semantic information than syllable-level ones; thus, word-based retrieval enhances the precision. On the other hand, syllable-level indexing features are more robust against the Chinese word tokenization ambiguity, Chinese homophone ambiguity, open vocabulary problem, and speech recognition errors; thus, the syllable-level information would enhance the recall. Accordingly, there is good reason to fuse the information obtained from indexing features of different levels. It has been shown that using syllable pairs as the indexing terms, in conjunction with the vector space model (VSM), is very effective for Chinese SDR, and the retrieval performance can be further improved by incorporating the information from word-level indexing features [13].

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

In this article, both the manual transcript and the recognition transcript of each spoken document, in form of a word stream, were automatically converted into a stream of overlapping syllable pairs. Then, all the distinct syllable pairs occurring in the spoken document collection were then 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, in replace of words, to represent the spoken documents, and construct the associated probabilistic latent topic distributions for DTM and WTM accordingly. The other is to jointly utilize both words and syllable pairs, two types of indexing terms, to represent the spoken documents, as well as to construct the associated probabilistic latent topic distributions. To this end, each spoken document virtually is represented with a spliced text stream, consisting of both words and syllable pairs. Figure 2 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 indexing features simultaneously.

3. EXPERIMENTAL SETUP

3.1 Corpus and Evaluation Metric

We used the Topic Detection and Tracking (TDT-2) collection [18] for this work. 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 [19]. In this article, 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. The 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 "*short*" query.

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,

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

Query Type	TD	SD
Long	0.639	0.562
Short	0.370	0.293

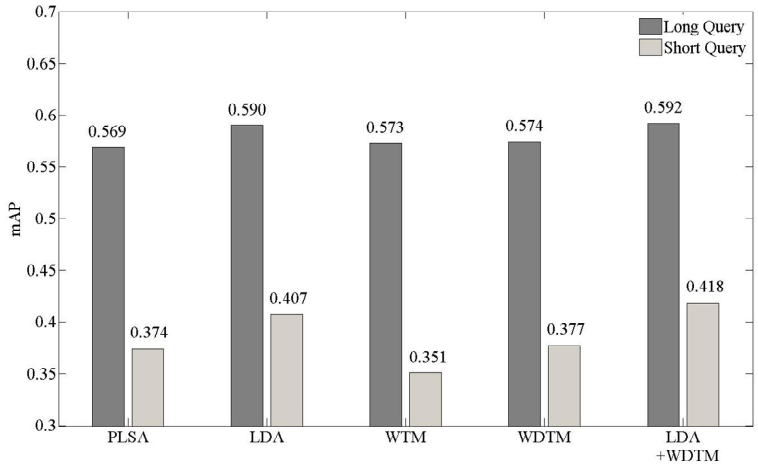

Figure 3. Retrieval results achieved by various topic models.

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 article. To assess the performance level of the recognizer, we spot-checked a fraction of spoken document collection set (about 39.90 h), and obtained error rates of 35.38% (word), 17.69% (character) and 13.00% (syllable).

The retrieval results are expressed in terms of non-interpolated mean average precision (mAP) following the TREC evaluation [20], which is computed by the following equation:

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}}, \tag{7}$$

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.

3.2 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 accordingly. Though document ranking based merely on DTM or WTM tends to increase recall, either 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 the discriminative power among concept-related words in finer granularity. Therefore, in this article, when either DTM or WTM is employed in evaluating the relevance between a query Q and a document D, we additionally incorporate the unigram probabilities of a query word (or term) occurring in the document $P(w_i \mid \mathrm{M}_D)$ and a general text corpus $P(w_i|\mathrm{M}_C)$ with the topic model $P_{\text{Topic}}(w_i|\mathrm{M}_D)$, for probability smoothing and better performance. For example, the probability of a query word generated by one specific topic model of a document (cf. Eqs. (3) and (5)) is modified as follows:

$$\begin{aligned} P(w|D) = \alpha \cdot \left[\beta \cdot P_{\text{Topic}}(w_i|\mathrm{M}_D) + (1-\beta)\cdot P(w_i|\mathrm{M}_D)\right] \\ + (1-\alpha)\cdot P(w_i|\mathrm{M}_C) \end{aligned} \tag{8}$$

where $P_{\text{Topic}}(w_i|\mathrm{M}_D)$ can be the probability of a word w_i generated by PLSA (cf. Eq. (3)) or WTM (cf. Eq. (5)); the values of the interpolation weights α and β can be empirically set or further optimized by other optimization techniques [4, 8]. A detailed account of this issue will be given in Section 4.2. On the other hand, the Gibbs sampling algorithm [17] is used to inference the parameters of the LDA and WDTM models.

4. EXPERIMENTAL RESULTS

4.1 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 also listed for reference, 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 is 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 is higher than 35%. On the other hand, retrieval using short queries degrades the performance approximately 45% relative when compared to retrieval using long queries. This is due to the fact that a long query usually contains more different words describing the 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

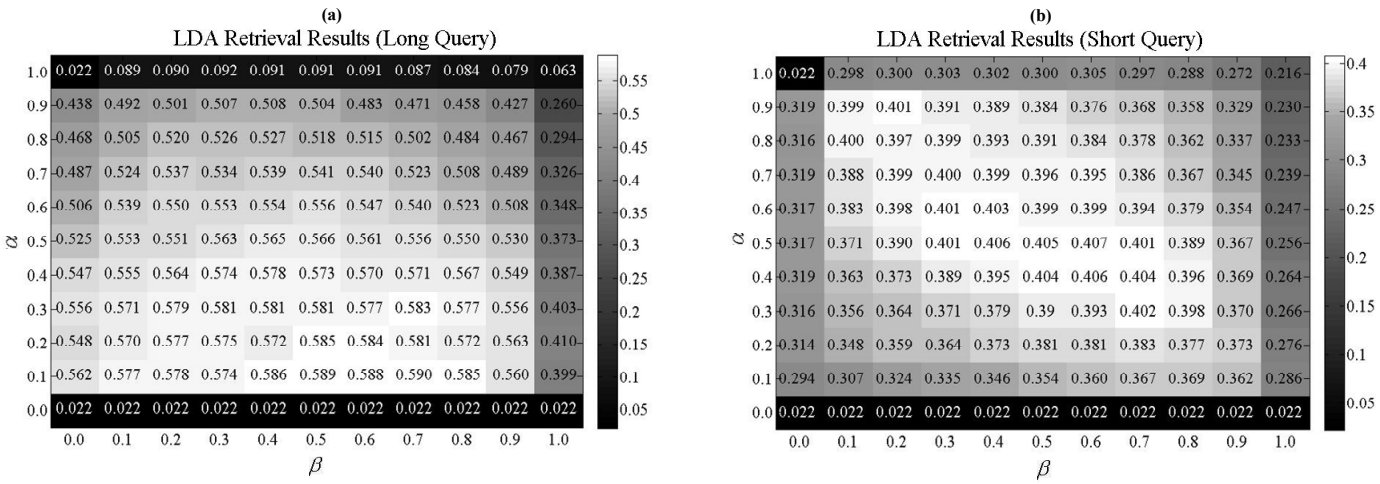

Figure 4. Detailed retrieval results achieved by LDA with respect to different types of queries.

matching. From now on, unless otherwise stated, we will only report the retrieval results for the SD case.

4.2 Experiments on DTM and WTM

In the next set of experiments, we assess the utility of various topic models for SDR, including PLSA, LDA, WTM, as well as WDTM. The corresponding retrieval results are shown in Figure 3. It is worth mentioning that all these topic models are trained without supervision and have the same number of latent topics which is 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 [21]. On the other hand, Both WTM and WDTM have the same context window size S set to 21 [7]. Since this article is set out to investigate the effectiveness of various topic models for SDR, the interpolation weights α and β defined in Eq. (8) hence are further 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. Refer to Figure 3, all these topic models give moderate but consistent improvements over the baseline ULM model when long queries are evaluated. One possible explanation is that the information need might have been already fully stated 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.

We then compare among these topic models: LDA outperforms PLSA while WDTM outperforms WTM. This finding supports the argument that constraining the latent topic distributions with Dirichlet priors will lead to better model estimation. Moreover, LDA is the best among these topic models. As compared to the baseline ULM model, it yields about 5% and 39% relative improvements for long and short queries, respectively. On the other hand, fusion of LDA and WDTM (denoted by LDA + WDTM) provide an additional gain of about 1% for the case of using short queries. To do this, we linearly combined the document ranking scores of LDA and WDTM in the log-likelihood domain [4].

To go a step further, we attempt to investigate the more subtle interaction effects among the topic model $P_{\text{Topic}}(w_i|\text{M}_D)$, the document literal term model $P(w_i|\text{M}_D)$ and the background model $P(w_i|\text{M}_C)$ in Eq. (8) by varying the values of the interpolation weights α and β. Here LDA is taken as an example topic model since it exhibits the best performance among the topic models compared in this article. The retrieval results are graphically illustrated in Figure 4 where the horizontal and vertical axes denote the values of α and β, respectively. As revealed by Figure 4, additionally incorporating of $P(w_i|\text{M}_D)$ and $P(w_i|\text{M}_C)$ into LDA is beneficial for retrieval. In an extreme case, when both the values of α and β are set equal to one, as shown in the top right corner of Figure 4, it leads to a retrieval model based merely on the topical information, which has poor retrieval performance especially for the case of 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 be 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 right most columns of Figure 4 that using the background model $P(w_i|\text{M}_C)$ can absorb the contributions of the common and non-informative words made to document ranking, and thus give better retrieval performance.

Looking at each row of Figure 4, we see that smoothing LDA with the document literal term model $P(w_i|\text{M}_D)$ is also useful. This is attributed to the fact that discriminative (or informative) words will occur repeatedly in a document; $P(w_i|\text{M}_D)$ hence gives more emphasis on these words. On the other hand, Figure 4 also reflects that smoothing LDA with the background model $P(w_i|\text{M}_C)$ is necessary when the query is long, but it does not seem to be helpful for the case of using a short query. This is mainly because the information need stated by the short query is already in a concise manner, and the importance of the role that $P(w_i|\text{M}_C)$ plays to filter out or deemphasize non-informative words is reduced.

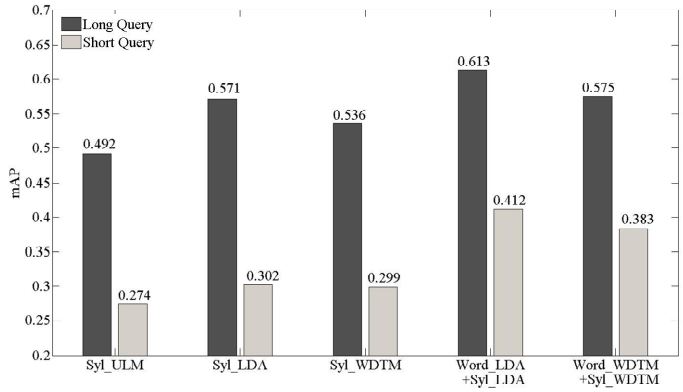

Figure 5. Retrieval results achieved by LDA and WDTM, respectively, using syllable-pairs as well as the combination of words and syllable pairs.

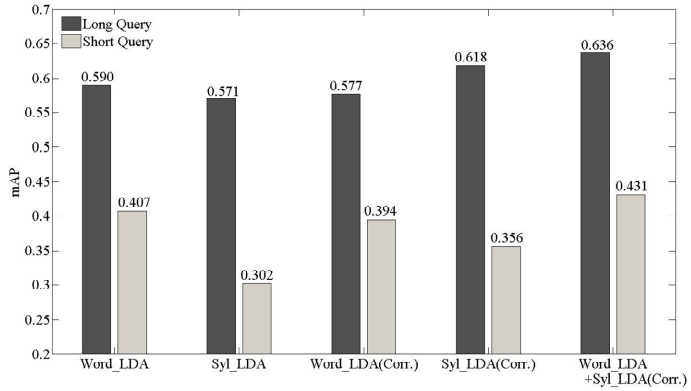

Figure 6. Retrieval results achieved by correlated LDA, using words (Word_LDA(Corr.)), syllable pairs (Syl_LDA(Corr.)) and their combination (Word_LDA + Syl_LDA(Corr.)).

4.3 Experiments on using Subword-level Indexing Features

In the third set of experiments, we evaluate the performance of the topic models when syllable pairs are instead utilized as the indexing terms. Here we take LDA and WDTM as the example topic models, and the corresponding models are denoted by Syl_LDA and Syl_WDTM, respectively. Fusion of words and syllable pairs for topic modeling is investigated as well. Note that Word_LDA denotes LDA using words as the indexing terms, and was termed LDA in the previous sections.

The retrieval results of Syl_LDA and Syl_WDTM are shown in Figure 5, where the results achieved by ULM and using syllable pairs as the indexing terms (denoted by Syl_ULM) are also depicted for comparison. Several observations can be made from Figure 5. First, the topic models (Syl_WDTM and Syl_LDA) are again 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 that with the word-level information. This may be simply due 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 combinations of the word- and syllable-information for topic modeling demonstrate much better retrieval results (cf. the right-most two sets of bars in Figure 5) as compared to that of the topic models with merely the word-level information (cf. Figure 3).

Finally, we examine 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 take the LDA model as an example to study the effectiveness of such an attempt and the associated results are shown in Figure 6. The results reveal that when only syllable pairs are used as the indexing 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 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 be interfered by syllable pairs when we attempt to splice these two distinct indexing 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. Figures 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. However, in the case of using short queries, Syl_LDA(Corr.) does not perform as well as LDA that uses words as the indexing terms to construct the latent topic distributions (denoted by Word_LDA). We conjecture 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 combine 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 have in mind that these results were obtained by using the erroneous speech transcripts of the spoken documents (i.e., the SD case). It 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.

5. CONCLUSIONS

In this article, we have thoroughly investigated two categories of topic models, including the document topic models (DTM) and the word topic models (WTM), for SDR. Moreover, we have leveraged different levels of indexing 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 [7], 2) investigating discriminative training of topic models [4], and 3) integrating the topic models with the other more elaborate representations of the speech recognition output [10, 11, 22] for larger-scale SDR tasks.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-003-015-MY3, NSC95-2221-E-003-014-MY3, and NSC97-2631-S-003-003.

7. REFERENCES

[1] Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. In Proc. the ACM SIGIR Conference on R&D in Information Retrieval, 275-281.

[2] Miller, D. R. H., Leek, T., and Schwartz, R. 1999. A hidden Markov model information retrieval system. In Proc. ACM SIGIR Conference on R&D in Information Retrieval, 214-221.

[3] Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196.

[4] Chen, B., Wang, H. M. and Lee, L. S. 2004. A discriminative HMM/n-gram-based retrieval approach for Mandarin spoken documents. ACM Transactions on Asian Language Information Processing 3, 2, 128-145.

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

[6] Blei, D.M., Ng, A.Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993-1022.

[7] Chen, B. 2009. Latent topic modeling of word co-occurrence information for spoken document retrieval. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 3961-3964.

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

[9] Griffiths, T. L., Steyvers, M. and Tenenbaum, J. B. 2007. Topics in semantic representation. Psychological Review 114, 211-244.

[10] Yi, X. and Allan, J. 2009. A Comparative Study of Utilizing Topic Models for Information Retrieval. In Proc. the European Conference on Information Retrieval, 29-41.

[11] Chelba, C., Hazen, T. J., and Sarclar, M. 2008. Retrieval and browsing of spoken content. IEEE Signal Processing Magazine. 25, 3, 39-49.

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

[13] 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. on Speech and Audio Processing 10, 5, 303-314.

[14] Garofolo, J., Auzanne, G., and Voorhees, E. 2000. The TREC spoken document retrieval track: A success story. In Proc. the 8th Text REtrieval Conference. NIST, 107-129.

[15] Wei, X., and Croft, W. B. 2006. LDA-based document models for ad-hoc retrieval. In Proc. the ACM SIGIR Conference on R&D in Information Retrieval, 178-185.

[16] Ypma, J., Basten, T. and Lafferty, J. 2002. Expectation-propagation for the generative aspect model. In Proc. Conference on Uncertainty in Artificial Intelligence, 352–359.

[17] Griffiths, T. L. and Steyvers, M 2004. Finding scientific topics. In Proc. of the National Academy of Sciences, 5228-5235.

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

[19] 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. Computer Speech and Language 18, 2, 163-179.

[20] Harman D. 1995. Overview of the Fourth Text Retrieval Conference (TREC-4). In Proc. the Fourth Text Retrieval Conference, 1-23.

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

[22] Lin, S.H., Chen, B., 2009. Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures. In Proc. the Annual Conference of the International Speech Communication Association.