Query Modeling for Spoken Document Retrieval

Berlin Chen [#1], Pei-Ning Chen [#1], Kuan-Yu Chen [*2]

[#] *National Taiwan Normal University, Taiwan*

[1] berlin@ntnu.edu.tw

[*] *Institute of Information Science, Academia Sinica, Taiwan*

[2] kychen@iis.sinica.edu.tw

***Abstract*—Spoken document retrieval (SDR) has recently become a more interesting research avenue due to increasing volumes of publicly available multimedia associated with speech information. Many efforts have been devoted to developing elaborate indexing and modeling techniques for representing spoken documents, but only few to improving query formulations for better representing the users' information needs. In view of this, we recently presented a language modeling framework exploring a novel use of relevance information cues for improving query effectiveness. Our work in this paper continues this general line of research in two main aspects. We further explore various ways to glean both relevance and non-relevance cues from the spoken document collection so as to enhance query modeling in an unsupervised fashion. Furthermore, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the various relevance and/or non-relevance cues. Experiments conducted on the TDT (Topic Detection and Tracking) SDR task demonstrate the performance merits of the methods instantiated from our retrieval framework when compared to other existing retrieval methods.**

I. Introduction

In the recent past, spoken document retrieval (SDR) has received a growing amount of interest and activity in the speech processing community. This is due in large part to the advances in automatic speech recognition (ASR) and the ever-increasing volumes of multimedia associated with spoken documents made available to the public, such as radio and TV broadcasts, lecture recordings, meetings and telephone conversations, digital archives, among many others [1, 2]. Unlike research on spoken term detection (STD) that usually embraces the goal of extracting probable spoken terms or phrases inherent in a spoken document that could match the query words or phrases literally [3], research on SDR revolves more around the notion of relevance of a spoken document in response to a query [4]. It is generally agreed upon that a document is relevant if it could address the stated information need of the query, not because it just happens to contain all the words in the query [5]. Nonetheless, most efforts of SDR research have been placed on the exploration of robust indexing or modeling techniques to represent spoken documents [3, 4, 6, 7], but few look at the other side of the coin, that is, the improvement of query modeling for better reflecting the underlying information need of a user.

As to the development of document-ranking algorithms for information retrieval (IR), over the past decade, statistical language modeling (LM) has become an attractive choice due to its simplicity and clear probabilistic meaning, as well as state-of-the-art performance. In practice, the relevance (or similarity) measure for the LM approach is usually computed by two different matching strategies, namely, literal term matching and concept matching [4]. The unigram language model (ULM) is the most popular example for literal term matching [8, 9]. In this category of methods, 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 words (or index terms). Accordingly, documents can be ranked according to their likelihood of generating the query, viz. the query-likelihood. Yet, there has been much work striving to extend ULM to further capture context dependence based on *n*-grams of various orders, or some grammar structures, mostly leading to mild gains or even spoiled results [9].

The above category of methods 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 different sets of words. Another category of LM methods attempt to discover the latent topic information embedded in the query and documents, based on which the retrieval is performed. For example, latent Dirichlet allocation (LDA) [10] and its precursor, probabilistic latent semantic analysis (PLSA) [11], are often considered to be two basic formulations following this line of thought. 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. Despite the fact that there are many follow-up studies and extensions of LDA and PLSA, empirical evidence in the literature indicates that more sophisticated (or complicated) topic models, such as Pachinko allocation model (PAM), do not necessary offer further retrieval benefits [12, 13].

Apart from developing more elaborate document modeling approaches to SDR, we recently introduced a new perspective on improving the query formulation [14]. A relevance language modeling framework that discovers extra information cues that are relevant to the query intent was thus presented. Our work in this paper continues this general line of research by further exploring various ways to glean extra information cues from relevant and/or non-relevant documents

to enhance query modeling in an unsupervised fashion. Furthermore, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the various relevance and/or non-relevance cues.

The remainder of this paper is structured as follows. We briefly review the mathematical formulations of the basic LM methods to SDR in Section II. In Section III, we describe the relevance language modeling framework that can leverage lexical co-occurrence and topic cues extracted from relevant documents for improving query effectiveness, followed by an elucidation of the various methods exploited to leverage non-relevance information cues to enhance query modeling in Section IV. After that, the experimental settings and a series of retrieval experiments are presented in Sections V and VI, respectively. Finally, Section VII concludes our presentation and discusses avenues for future work.

II. Language Modeling for SDR

The fundamental formulation of the LM approach to SDR, is to compute the conditional probability $P(Q|D)$, i.e., the likelihood of a query Q generated by each spoken document D [9]. A spoken document is deemed to be relevant to a query if the corresponding document model is more likely to generate the query. If the query Q is treated as a sequence of words (or terms), $Q = q_1, q_2, \cdots, q_L$, 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 similarity measure $P(Q|D)$ can be further decomposed as a product of the probabilities of the query words generated by the document:

$$SIM_1(Q, D) = P(Q|D) = \prod_{l=1}^{L} P(q_l|D), \quad (1)$$

where $P(q_l|D)$ is the likelihood of D generating q_l (a.k.a. the document model). Here, we consider two variants for constructing the document model for each document D. One is to use the unigram language model (ULM). To this end, each document can, respectively, offer a unigram distribution for observing a query word, which is built on the basis of the words occurring in the document with the maximum likelihood (ML) estimator. The document model is further smoothed by a background unigram language model estimated from a large general collection to avoid the problem of zero probability. The other is to employ a probabilistic topic model, such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), which calculates the query-likelihood based on the frequency of q_l occurring in a given latent topic as well as the likelihood that D generates the respective topic. However, PLSA and LDA offer coarse-grained latent semantic representations about the information need at the expense of losing the power to distinguish the fine-grained difference in the meanings of semantically-related words. In implementation, there is always good reason to combine them with ULM for better retrieval quality [4, 15].

Another basic formulation of LM for SDR is the Kullback-Leibler (KL)-divergence measure [9, 16]:

$$SIM_2(Q, D) = -KL(Q\|D) = -\sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|D)}, \quad (2)$$

where both a query and a document is, respectively, regarded as a unigram language model (i.e., $P(w|Q)$ and $P(w|D)$) for predicting any word w in the vocabulary V. A document D has a smaller value (or probability distance) in terms of $KL(Q\|D)$ is deemed to be more relevant to Q. The retrieval effectiveness of the KL-divergence measure depends primarily on the accurate estimation of the query model $P(w|Q)$ and the document model $P(w|D)$. Further, as it turns out, the KL-diverge measure will give the same ranking as the query-likelihood measure shown earlier in (1), when the query model $P(w|Q)$ is simply derived with the ML estimator by counting the number of occurrences of w in Q. However, the KL-divergence measure has the merit of being able to accommodate extra information cues to improve the estimate of its component models (e.g., the query model) for better document ranking in a systematic way.

Due to that a query usually consists of only a few words, the true query model $P(w|Q)$ might not be accurately estimated by the ML estimator. We hence look to mitigate this problem by exploring extra cues to improve the query model involved in the KL-divergence measure. The notion of leveraging relevance cues, or relevance language modeling, has recently attracted much attention and been applied with empirical success to a number of text IR tasks [17, 18]; however, as far as we are aware, there is still not much research on leveraging relevance cues for the LM approach to SDR [14]. In this paper, we take a step forward by incorporating non-relevance cues into the similarity measure to improve the retrieval effectiveness of a given query.

III. Query Modeling with Relevance Information

A. The Relevance Model (RM)

A simple yet effective strategy to improve the query formulation in the KL-divergence measure is to explore extra relevance information pertaining to the query with the relevance model (RM) [17, 18]. For this idea to work, each query Q is assumed to be associated with an unknown relevance class R_Q, and documents that are relevant to the information need expressed in the query are samples drawn from R_Q. The document ranking problem then can be reduced to the problem of finding a mechanism to determine the relevance model (RM) or, more specifically, the probability $P_{\mathrm{RM}}(w)$ of observing words w in the documents relevant to a particular information need. The relevance model $P_{\mathrm{RM}}(w)$, as a multinomial view of R_Q, can be defined as the probability distribution which gives the probability that we would observe a word if we were to randomly select a document from the relevance class and select the word from that document. The joint probability of Q and w being generated by the

relevance class R_Q of Q, viz. $P_{\text{RM}}(Q,w)$, thus can serve as the building block for deriving the enhanced query model $\widetilde{P}(w|Q)$.

But in reality, since we usually do not have any information about the ideal set of relevant documents in the collection for each query, we may conduct an initial round of retrieval (or a local feedback-like procedure) that poses Q to an IR system to obtain a top-ranked list of M pseudo relevant documents $\mathbf{D}_{\text{Top}} = \{D_1, D_2, \ldots, D_M\}$ from the collection to approximate R_Q. Consequently, the joint probability of observing Q together with w can be:

$$P_{\text{RM}}(Q,w) = \sum_{m=1}^{M} P(D_m) P(q_1, q_2, \ldots q_L, w \mid D_m), \tag{3}$$

where $P(D_m)$ is the probability that we would randomly select D_m and $P(q_1, q_2, \ldots q_L, w \mid D_m)$ is the joint probability of simultaneously observing Q and w in D_m. If we further assume that words are conditionally independent given D_m and their order is of no importance (i.e., the "*bag-of-words*" assumption), then the joint probability can be decomposed as a product of unigram probabilities of words generated by D_m:

$$P_{\text{RM}}(Q,w) = \sum_{m=1}^{M} P(D_m) P(w \mid D_m) \prod_{l=1}^{L} P(q_l \mid D_m). \tag{4}$$

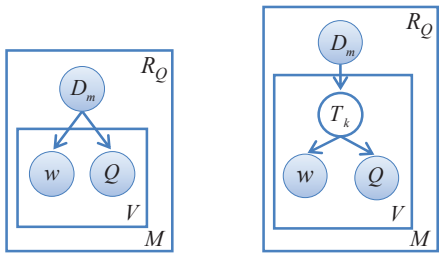
The probability $P(D_m)$ can be simply kept uniform or determined in accordance with the relevance of D_m to Q, while $P(w \mid D_m)$ and $P(q_l \mid D_m)$ are estimated based on the word occurrence counts in D_m. The enhanced query model $\widetilde{P}(w|Q)$, therefore, can be expressed by:

$$\begin{aligned} \widetilde{P}(w|Q) &= \frac{P_{\text{RM}}(Q,w)}{P_{\text{RM}}(Q)} \\ &= \frac{\sum_{m=1}^{M} P(D_m) P(w \mid D_m) \prod_{l=1}^{L} P(q_l \mid D_m)}{\sum_{m=1}^{M} P(D_m) \prod_{l=1}^{L} P(q_l \mid D_m)}. \end{aligned} \tag{5}$$

As such, $\widetilde{P}(w|Q)$ can be linearly combined with or used to replace $P(w|Q)$ in the KL-divergence measure to distinguish relevant documents from irrelevant ones. Although there have been previous efforts on exploiting different ways to derive $\widetilde{P}(w|Q)$, the formulation introduced in (5) has been validated to work more effectively and robustly than other variants across different document collections [19].

B. Incorporating Topic Cues into RM

Not content to merely apply the RM model to SDR, we may make a step forward to incorporate latent topic information into the RM modeling [20]. In doing so, the pseudo-relevant documents obtained by local feedback are assumed to share a set of pre-defined latent topic variables $\{T_1, T_2, \ldots, T_K\}$ describing the "*word-document*" co-occurrence characteristics. Therefore, the probability that a word w is sampled from a pseudo-relevant document D_m is not estimated directly based on the frequency of the word occurring in the document, but rather based on the frequency of the word in the latent topics as well as the likelihood that the document generates the respective topics:



(a) (b)

Fig. 1. The graphical model representations for (a) RM and (b) TRM. V is the number of distinct word tokens and M is the number of documents in the relevance class R_Q.

$$\widetilde{P}(w \mid D_m) = \sum_{k=1}^{K} P(w \mid T_k) P(T_k \mid D_m). \tag{6}$$

As with PLSA and LDA, the probabilities $P(w \mid T_k)$ and $P(T_k \mid D_m)$ presented here can be estimated using inference algorithms like the expectation-maximization (EM) algorithm when with uniform priors, or the variational approximation algorithm when with Dirichlet priors, on the whole spoken document collection. The joint probability of Q and w being simultaneously observed in the relevance class R_Q of Q, as shown earlier in (4), is thus decomposed as

$$\begin{aligned} &P_{\text{TRM}}(Q,w) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m) P(T_k \mid D_m) P(w \mid T_k) \prod_{l=1}^{L} P(q_l \mid T_k). \end{aligned} \tag{7}$$

We term (7) the topic-based relevance model (TRM) hereafter. In contrast to RM, TRM assumes that the additional cues of how words are distributed across a set of latent topics, gleaned from all spoken documents in the collection, can carry useful global topic structure for relevance modeling. The RM and TRM models are, respectively, depicted as a probabilistic graphical model in Figure 1.

IV. Leveraging Non-relevance Information

In addition to using the various RM models mentioned above to gather the relevance information in the initial round of retrieval, we hypothesize that the low-ranked (or pseudo non-relevant) documents $\mathbf{D}_{\text{Low}} = \{D'_1, D'_2, \ldots, D'_L\}$ can provide useful cues as well to boost the retrieval effectiveness of a given query. For this idea to work, we attempt to estimate a non-relevance model $P(w|NR_Q)$ for each test query Q based on those selected pseudo non-relevant documents. The similarity measure between Q and any D thus can be computed as follows:

$$SIM_3(Q,D) = SIM_2(Q,D) + \alpha \cdot KL(NR_Q \| D), \tag{8}$$

where the parameter α is used to adjust the relative contributions of $SIM_2(Q,D)$ and $KL(NR_Q \| D)$ to the final similarity measure $SIM_3(Q,D)$. Clearly, $SIM_3(Q,D)$ prefers those documents whose document models have not only a

smaller probability distance to the query model but also have a larger probability distance to the non-relevance model (NR). In implementation, the non-relevance model, viz. $\widetilde{P}(w|NR_Q)$, is estimated simply based on the number of times that w appears in $\mathbf{D}_{\text{Low}}$ with the ML estimator and then interpolated with a background unigram language model for probability smoothing:

$$\widetilde{P}(w|NR_Q) = \lambda \cdot P(w|NR_Q) + (1-\lambda) \cdot P(w|BG), \quad (9)$$

where $P(w|NR_Q)$ is the ML-estimated word distribution and $P(w|BG)$ is the background model; λ is a parameter which controls the contribution of $P(w|NR_Q)$. Alternatively, $P(w|NR_Q)$ in (9) can also be further optimized with the EM algorithm, leading to the following two update formulas:

$$P^{(m)}(NR_Q|w) = \frac{\lambda \cdot P^{(m)}(w|NR_Q)}{\lambda \cdot P^{(m)}(w|NR_Q) + (1-\lambda) \cdot P(w|BG)} \quad (10)$$

and

$$P^{(m+1)}(w|NR_Q) = \frac{\sum_{D' \in \mathbf{D}_{\text{Low}}} c(w,D') \cdot P^{(m)}(NR_Q|w)}{\sum_{w} \sum_{D' \in \mathbf{D}_{\text{Low}}} c(w,D') \cdot P^{(m)}(NR_Q|w)}, \quad (11)$$

where m denotes the m-th iteration of the EM algorithm and $c(w,D')$ is the number of times w occurring in D'. This will enable more specific words (viz. words in $\mathbf{D}_{\text{Low}}$ that are not well-explained by the background model) to receive more probability mass, thereby leading to a more discriminative non-relevance model for better retrieval performance.

It should be noted that, unlike previous studies that use simulated query conditions [21] or human-judged non-relevant documents [22] to construct the non-relevance model, in this paper, we investigate an unsupervised way to estimate the non-relevance model. Besides, we further study whether the relevance and non-relevance cues of a test query can conspire to enhance the SDR performance when using either word- or subword-level index terms, or their combination.

V. Experimental Setup

We used the Topic Detection and Tracking collection (TDT-2) for this work [23]. The Mandarin news stories from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. This task is especially useful for news monitoring and tracking. The average word error rate obtained for the spoken documents is about 35%. The retrieval results, assuming that manual transcripts for the spoken documents to be retrieved (denoted TD, text documents, in the tables below) are known, are also shown for reference, compared to the results when only the erroneous transcripts by speech recognition are available (denoted SD, spoken documents, in the tables below). The retrieval results are expressed in terms of non-interpolated mean average precision (mAP) following the TREC evaluation [5]:

TABLE I
Statistics for the TDT-2 Collection.

# 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	4,841	153	287
Length of query (in characters)	8	27	13	14
# Relevant documents per test query	2	95	13	29

$$\text{mAP} = \frac{1}{E} \sum_{i=1}^{E} \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{j}{r_{i,j}} \quad (12)$$

where E 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. Table I shows some basic statistics about the TDT-2 collection. Note also that in this paper, the number of latent topics used for constructing TRM, PLSA and LDA is set to 32, while the number of pseudo-relevant documents retrieved from the local feedback-like procedure for the various RM models is 15, albeit that these constants can be further fine-tuned for the spoken document collection through proper experimentation.

In this paper, we also propose to integrate subword-level information into query modeling for SDR. To do this, syllable pairs are taken as the basic units for indexing besides words. 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 identified to form a vocabulary of syllable pairs for indexing. We can simply use syllable pairs, in replace of words, to represent the spoken documents, and construct the associated component models of the retrieval framework accordingly. Further, it is generally expected that the fusion of different levels of index features would further improve the retrieval performance.

VI. Experimental Results

In this section, we begin by comparing the performance of various LM methods, working in conjunction with the word-level index features. They include ULM, PLSA and LDA, as well as the two variants of RM (viz. RM and TRM) introduced in Section III. Several observations can be made from Table II. First, PLSA and LDA consistently yield improvements of about 2% absolute over ULM no matter whether the manual transcripts (TD) or the recognition transcripts (denoted by SD) are being used. Second, RM and

TRM, which target at improving query modeling in the KL-divergence measure, deliver quite competitive results when compared to PLSA and LDA; the latter two focus on exploring latent topic information for more elaborate document modeling. Third, the additional exploration of topic cues in relevance modeling of the test query can further boost the performance (TRM vs. RM). Forth, retrieval using the recognition transcripts apparently falls short of that using the manual transcripts. Nevertheless, despite that speech recognition results in a word error rate higher than 35%, retrieval using the recognition transcripts does not cause severe performance degradation.

In the second set of experiments, we attempt to evaluate the utility of leveraging the low-ranked (pseudo non-relevant) documents to discover the non-relevance cues with respect to the test query. Meanwhile, the RM method is employed to explore the relevance cues of the query from the top-ranked (pseudo relevant) documents as well. Table III shows the corresponding results as a function of different numbers of low-ranked documents (counting from the bottom of the ranked list of spoken documents returned by the initial round of retrieval) being employed in the construction of $\widetilde{P}\left(w\middle|NR_Q\right)$. Observing Table III we notice two particularities. One is that the using the EM algorithm to infer $\widetilde{P}\left(w\middle|NR_Q\right)$ seems to be better than that with the ML estimator. The other is that the inclusion of more low-ranked documents for constructing $\widetilde{P}\left(w\middle|NR_Q\right)$ tends to achieve slightly better performance than that with fewer low-ranked documents. Thus, we also study the feasibility to use the entire spoken document collection to estimate $\widetilde{P}\left(w\middle|NR_Q\right)$ instead of using the low-ranked documents of a given test query. We expect that the entire spoken document collection can offer an alternative estimate of the non-relevance model, since the number of relevant documents with respect to a given query is usually very small in practice. It also has the additional advantage of estimating the non-relevance model beforehand (prior to query time) and thus reduces the effort of on-line query modeling. The corresponding results shown in the rightmost column of Table III (denoted by ALL) indeed confirm our expectation. Further, as we compare them to the results of RM shown in Table I (for the SD case), it reveals that the additional exploration of non-relevance information (viz. RM+NR) can benefit SDR to a significant extent, thereby corroborating the important role it plays in query modeling.

In the final set of experiments, we investigate the joint exploration of relevance and non-relevance cues for query modeling, in conjunction with different levels of index features (viz. word-level features, syllable-level features and their combination). As evidenced in the rightmost column of Table IV, such joint exploration of relevance and non-relevance cues for query modeling is quite effective across the various index features being used.

TABLE II
Retrieval results (in mAP) achieved by various LM-based retrieval models, using the word-level index features.

	ULM	PLSA	LDA	RM	TRM
TD	0.372	0.418	0.401	0.402	0.442
SD	0.323	0.345	0.341	0.364	0.394

TABLE III
Retrieval results (in mAP) achieved by pairing RM with NR, using the word-level index features.

		100	500	1,000	1,500	ALL
SD	ML	0.383	0.384	0.385	0.385	0.386
	EM	0.385	0.387	0.387	0.391	0.392

TABLE IV
Retrieval results (in mAP) achieved by pairing RM with NR, and TRM with NR, using different levels of index features.

		Word	Syllable	Combination
SD	RM	0.364	0.378	0.396
	TRM	0.394	0.383	0.412
	RM+NR(ALL)	0.392	0.405	0.426
	TRM+NR(ALL)	0.402	0.415	0.441

VII. CONCLUSION

In this paper, we have investigated a novel framework to explore both relevance and non-relevance cues for improved query modeling, which suggests a promising avenue for the LM approach to SDR. The utility of the methods deduced from such a framework have also been validated by extensively comparisons with several widely used retrieval methods. The experimental results indeed demonstrate the applicability of our methods. As to future work, we envisage three directions: 1) investigating more elaborate training algorithms to enhance the discriminative capabilities of the query and document models involved in the presented framework [24], 2) further confirming our observations on larger-scale experiments, and 3) applying the presented methods to speech summarization [25].

ACKNOWLEDGMENT

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 98-2221-E-003-011-MY3, NSC 99-2221-E-003-017-MY3, NSC 100-2515-S-003-003, and NSC 100-2631-S-003 -006.

REFERENCES

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

[2] M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, 25(3), pp. 150–152, 2008.

[3] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 25 (3), pp. 39–49, 2008.

[4] B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM Transactions on Asian Language Information Processing*, 8 (1), pp. 2:1–2:27, 2009.

[5] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, ACM Press, 2011.

[6] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "Statistical lattice-based spoken document retrieval," *ACM Transactions on Information Systems*, 28 (1), pp. 2:1–2:30, 2010.

[7] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proc. SIGIR 2007*.
[8] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. SIGIR 1998*.
[9] C. X. Zhai, *Statistical Language Models for Information Retrieval: A Critical Review*, Foundations and Trends in Information Retrieval, 2 (3), 137–213, 2008.
[10] D. M. Blei et al., "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, pp. 993–1022, 2003.
[11] T. Hoffmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning,* 42, pp. 177–196, 2001.
[12] D. Blei and J. Lafferty, "Topic models," in A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.
[13] X. Yi and J. Allan, "A Comparative Study of Utilizing Topic Models for Information Retrieval," in *Proc. ECIR 2009*.
[14] P.-N. Chen, K.-Y. Chen, B. Chen, "Leveraging relevance cues for improved spoken document retrieval," in *Proc. Interspeech 2011*.
[15] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. SIGIR 2006*.
[16] S.-H. Lin, Y.-M. Yeh, B. Chen, "Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 19(4), pp. 871-882, 2011.
[17] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proc. SIGIR 2001*.
[18] V. Lavrenko, A Generative Theory of Relevance, Springer, 2009.
[19] Y. Lv and C. X. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proc. CIKM 2009*.
[20] K.-Y. Chen and B. Chen, "Relevance language modeling for speech recognition," in *Proc. ICASSP 2011*.
[21] X. Wang, H. Fang, and C. X. Zhai, "A study of methods for negative relevance feedback," in *Proc. SIGIR 2008*.
[22] N. Soskin, O. Kurland and C. Domshlak, "Navigating in the dark: Modeling uncertainty in ad hoc retrieval using multiple relevance models," in *Proc. ICTIR 2009*.
[23] LDC. 2000. Project topic detection and tracking. Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/TDT/.
[24] J. W. Kuo, B. Chen, "Minimum word error based discriminative training of language models," in *Proc. Interspeech 2005*.
[25] B. Chen and S. -H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.