Leveraging Relevance Cues for Improved Spoken Document Retrieval

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

[1] National Taiwan Normal University, Taiwan
[2] Institute of Information Science, Academia Sinica, Taiwan

berlin@ntnu.edu.tw

Abstract

Spoken document retrieval (SDR) has emerged as an active area of research in the speech processing community. The fundamental problems facing SDR are generally three-fold: 1) a query is often only a vague expression of an underlying information need, 2) there probably would be word usage mismatch between a query and a spoken document even if they are topically related to each other, and 3) the imperfect speech recognition transcript carries wrong information and thus deviates somewhat from representing the true theme of a spoken document. To mitigate the above problems, in this paper, we study a novel use of a relevance language modeling framework for SDR. It not only inherits the merits of several existing techniques but also provides a flexible but systematic way to render the lexical and topical relationships between a query and a spoken document. Moreover, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the various relevance cues. Experiments conducted on the TDT SDR task show promise of the methods deduced from our retrieval framework when compared with a few existing retrieval methods.

Index Terms: spoken document retrieval, language modeling, relevance model, topic model, Kullback-Leibler divergence

1. Introduction

The last ten years have witnessed a large amount of research on spoken document retrieval (SDR) due to the increasing volume of multimedia associated with spoken documents made available to the public. Considerable research efforts have been devoted towards developing robust indexing (or representation) techniques so as to extract probable spoken terms or phrases inherent in a spoken document that could match the query words or phrases literally (the so-called spoken term detection, STD) [1], instead of revolving around the notion of relevance of a spoken document to a query, through the use of existing retrieval models [2]. Nevertheless, 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.

More recently, statistical language modeling (LM) for information retrieval (IR) has enjoyed increasing popularity due to its simplicity and clear probabilistic meaning, as well as state-of-the-art performance. In practice, the relevance measure for the LM approach is usually computed by two different matching strategies, namely, literal term matching and concept matching [3]. The unigram language model (ULM) is the most popular example for literal term matching [4, 5]. 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. Yet, there has been much work to further extend ULM to capture context dependence based on *n*-grams of various orders, or some grammar structures, mostly leading to mild gains or even spoiled results [5].

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 a different set of words. Another family of LM methods attempt to discover the latent topic information inherent in the query and documents, based on which the retrieval is performed; latent Dirichlet allocation (LDA) [6] and its precursor, probabilistic latent semantic analysis (PLSA) [7], are often considered to be two basic formulations 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. 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 [8, 5].

In this paper, we investigate a relevance language modeling framework, leveraging information that is related to the query intent, for improving the query formulation in SDR. It not only inherits the merits of the above several methods but also provides a flexible but systematic way to render the lexical and topical relationships between a query and a spoken document. Moreover, we also exploit multi-levels of index features, including words, syllable-level units and their combination, to work in conjunction with the various relevance cues. Empirical results on the TDT SDR task show the utility of the methods deduced from such a retrieval framework.

2. Language Modeling for SDR

The fundamental formulation of the LM approach to IR (or SDR) is to compute the conditional probability $P(Q \mid D)$, i.e., the likelihood of a query Q generated by each spoken document D [5]. 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 relevance measure $P(Q \mid D)$ can be further decomposed as a product of the probabilities of the query words generated by the document:

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

where $P(q_l \mid D)$ 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), where each document can, respectively, offer a unigram distribution for observing a query word, which is buit on the basis of the words occurring in the document with the maximum likelihood estimator (MLE) and further interpolated with a background unigram language model for the purpose of probability smoothing. 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 [8]. 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 [9].

Another basic formulation of LM for SDR is the Kullback-Leibler (KL)-divergence measure [5]:

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

where both a query and a document is, respectively, modeled as a 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. It turns out that it is easy to show that the KL-divergence measure covers the query-likelihood measure, as shown earlier in (1), as a special case when we merely use the empirical query word distribution derived by MLE to approximate the query model $P(w|Q)$. 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.

Note that the true query model $P(w|Q)$ might not be accurately estimated by MLE, since a query usually consists of only a few words. In order to alleviate this problem, we focus hereafter on exploring relevance cues to improve the query model involved in the KL-divergence measure. The notion of exploring relevance cues, or relevance language modeling, has recently attracted much attention and been applied with empirical success to a number of text IR tasks; however, as far as we are aware, there is still not much research on leveraging relevance cues along with multi-levels of index features for the LM approach to SDR. In this paper, we also make a step further by incorporating latent topic information into such a relevance language modeling framework.

3. Relevance Language Modeling

3.1. Principle

In relevance language modeling to IR, each query is assumed to be associated with an unknown relevance class R, and documents that are relevant to the information need expressed in the query are samples drawn from R [10, 11, 12]. 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, 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 relevant class and select the word from that document. The joint probability of Q and w being generated by the relevance class R of Q, i.e., $P_{\mathrm{RM}}(Q,w)$, thus can serve as the basis for deriving the enhanced query model $P_{\mathrm{RM}}(w|Q)$.

But in reality, since there is no prior knowledge 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 from the collection to approximate R, denoted by $\mathbf{D}=\{D_1,D_2,\ldots,D_M\}$. Then, as an instantiation, the joint probability of observing Q together with w can be:

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

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:

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

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 frequencies in D_m. The enhanced query model $P_{\mathrm{RM}}(w|Q)$, therefore, can be expressed by:

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

As such, $P_{\mathrm{RM}}(w|Q)$ 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 $P_{\mathrm{RM}}(w|Q)$, the formulation introduced in (5) has been validated to work more effectively and robustly than other variants across different document collections [13].

3.2. Incorporating Latent Topic Information

Not satisfied with merely applying the RM model to SDR, in this paper, we make a step forward to incorporate latent topic information into the RM modeling. For this idea to work, 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:

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

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 expectation-maximization (when with uniform priors) or variational approximation (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 of Q, as shown earlier in (4), is thus decomposed as

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

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.

3.3. Modeling Pairwise Word Associations

Furthermore, instead of using RM to model the association between an entire query Q and a word w, we can alternatively use RM to render the pairwise word association between a word q_l in the query Q and any word w (denoted by PRM):

$$P_{\text{PRM}}(q_l, w) = \sum_{m=1}^{M} P(D_m)P(q_l \mid D_m)P(w \mid D_m). \qquad (8)$$

By performing an algebra similar to (5), we can arrive at the conditional probability $P_{\text{PRM}}(w|q_l)$ of w given q_l. Thus, a "*composite*" probability for the query Q to generate w can be obtained by linearly combining $P_{\text{PRM}}(w|q_l)$ of all the words q_l in the query Q:

$$P_{\text{PRM}}(w|Q) = \frac{1}{L}\sum_{l=1}^{L} P_{\text{PRM}}(w|q_l), \qquad (9)$$

which can be regarded as a kind of LM for translating words q_l in the query to w. By the same token, we can also introduce a set of latent topics $\{T_1, T_2, \ldots, T_K\}$ to describe the word-word co-occurrence relationships in a pseudo-relevant document, and the pairwise word association between a query word q_l and any word w is thus modeled by

$$P_{\text{TPRM}}(q_l, w) = \sum_{m=1}^{M}\sum_{k=1}^{K} P(D_m)P(T_k \mid D_m)P(q_l \mid T_k)P(w \mid T_k). \qquad (10)$$

The query model derived based on (10) is referred to as TPRM.

3.4. Different Granularities of Index Features

In this paper, we also propose to integrate subword-level information into relevance 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.

4. Experimental Setup

We used the Topic Detection and Tracking collection (TDT-2) for this work [14]. 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 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 [2]. Table 1 shows some basic statistics about the TDT-2 collection. Note also that in this paper, the number of latent topics used for constructing TRM, TPRM, 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.

Table 1. *Statistics for 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	4841	153	287
Length of query (in characters)	8	27	13	14
# Relevant documents per test query	2	95	13	29

5. Experimental Results

At the outset, we report on the baseline retrieval results obtained by three basic LM models compared in this paper, including ULM, PLSA and LDA; the results for two classic algebraic models, namely vector space model (VSM) and latent semantic analysis (LSA), are also listed for reference. Consulting Tables 2 and 3 we notice two particularities. One is that these LM models yield comparable or even better performance than the algebraic models when using either the manual transcripts (denoted by TD) or the recognition transcripts (denoted by SD) to represent the spoken documents. Despite that LDA has some known theoretical advantages over PLSA, they tend to perform on par with each other for the SDR task studied here, which is in line with the results recently obtained by Lu et al. [15] for text IR. The other is that PLSA and LDA, aiming at unsupervised analysis of the latent topic information so as to enhance the estimation of the document model involved in (2), work quite well with the syllable features for the SD case. The performance gap between the TD and SD cases can be reduced to a good extent by PLSA and LDA; this reveals that topic modeling and subword-level (or syllable-level) indexing could be complementary of each other for the SDR task. Further, although the word error rate (WER) for the spoken document collection is higher than 35%, it does not lead to catastrophic failures probably due to the reason that recognition errors are overshadowed by a large number of spoken words correctly recognized in the documents.

In the second set of experiments, we evaluate the utility of the various RM models investigated in this paper, including RM, TRM, PRM and TPRM. All these four models are trained without supervision, as described in Section 3. In contrast to PLSA and LDA, these RM models focus on the orthogonal problem of robustly estimating the query model involved in (2). The corresponding results are shown in Tables 4 and 5, from which several observations can be made. At first glance, it seems that all these RM models can achieve considerable improvements over the baseline ULM model. Second, for the SD case, document modeling (i.e., PLSA and LDA) seems to benefit more from the alternative use of syllable features than query modeling (i.e., the various RM models), probably since that the former suffer more from imperfect speech recognition transcripts. Third, TRM (with either the uniform priors or the Dirichlet priors) tends to be consistently superior to the other three RM variants in most conditions and demonstrates performance competitive to PLSA and LDA. Fourth, RM assumes that the entire query (or all the query words) $Q = q_1, q_2, \cdots, q_L$ and the word w are sampled from the same relevant document, while PRM relaxes this assumption by considering individually the pairwise co-occurrence relationship between any query word q_l and the word w in each pseudo-relevant document. The latter turns out not to have better retrieval results than the former, which suggests that discovering the co-occurrence relationship between the entire query and any given word in the pseudo-relevant documents is paramount to the success of relevance modeling. An analogous reasoning can also be applied to the comparison of TRM and TPRM.

In the last set of experiments, as an illustration, we consider the paring of TRM with PLSA through the use of different levels of index features (i.e., word-level features, syllable-level features and their combination), both of which conspire to enhance the estimation of the query and document models employed in (2) simultaneously. The corresponding results are shown in Table 6, demonstrating that the marriage of TRM with PLSA can offer additional gains and yield the best retrieval effectiveness among the methods compared in this paper.

6. Conclusions

In this paper, we have investigated a relevance language modeling framework for SDR, which suggests a promising avenue for the integration of relevance, topic and co-occurrence information. 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 two directions: one is utilizing speech summarization techniques to help better estimate the query and document models [16]; the other is training the query and document models in a lightly-supervised manner through the exploration of users' click-through data [17].

7. Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under Grants 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.

Table 2. *Retrieval results achieved by various baseline retrieval models using word–level index features.*

	ULM	PLSA	LDA	VSM	LSA
TD	0.372	0.418	0.401	0.339	0.360
SD	0.323	0.345	0.341	0.275	0.352

Table 3. *Retrieval results achieved by various baseline retrieval models using syllable–level index features.*

	ULM	PLSA	LDA	VSM	LSA
TD	0.349	0.438	0.442	0.308	0.331
SD	0.330	0.419	0.413	0.257	0.338

Table 4. *Retrieval results achieved by the various RM models using word–level index features.*

	RM	PRM	TRM		TPRM	
			Uniform	Dirichlet	Uniform	Dirichlet
TD	0.402	0.400	0.442	0.440	0.399	0.397
SD	0.364	0.366	0.394	0.384	0.364	0.362

Table 5. *Retrieval results achieved by the various RM models using syllable–level index features.*

	RM	PRM	TRM		TPRM	
			Uniform	Dirichlet	Uniform	Dirichlet
TD	0.408	0.400	0.423	0.419	0.403	0.399
SD	0.378	0.371	0.383	0.382	0.372	0.368

Table 6. *Retrieval results achieved by pairing TRM with PLSA using different levels of index features.*

	TRM (Uniform) + PLSA		
	Word	Syllable	Combination
TD	0.456	0.447	0.471
SD	0.392	0.441	0.462

8. References

[1] C. Chelba et al., "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 25(3), pp. 39–49, 2008.

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

[3] 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, March 2009.

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

[5] C. X. Zhai, *Statistical Language Models for Information Retrieval: A Critical Review*, Foundations and Trends in Information Retrieval, 2 (3), 137–213, 2008.

[6] D. M. Blei et al., "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, pp. 993–1022, January 2003.

[7] T. Hoffmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning,* 42, pp. 177–196, 2001.

[8] D. Blei and J. Lafferty, "Topic models," in A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

[9] B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," in *Proc. ICASSP 2009*.

[10] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proc. SIGIR 2001*.

[11] V. Lavrenko, *A Generative Theory of Relevance*, Springer, 2009.

[12] K. Y. Chen and B. Chen, "Relevance language modeling for speech recognition," in *Proc. ICASSP 2011*.

[13] Y. Lv and C. X. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proc. CIKM 2009*.

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

[15] Y. Lu et al., "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Information Retrieval*, 2010.

[16] B. Chen and S. -H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[17] T. Joachims and F. Radlinski, "Search engines that learn from implicit feedback," *IEEE Trans. Computer*, 40(8), pp. 34–40, 2007.