JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 30, 553-569 (2014)

Enhancing Query Formulation for Spoken Document Retrieval

BERLIN CHEN[1], YI-WEN CHEN[1], KUAN-YU CHEN[2],
HSIN-MIN WANG[2] AND KUEN-TYNG YU[1]
[1]*Department of Computer Science and Engineering*
National Taiwan Normal University
Taipei, 106 Taiwan
[2]*Institute of Information Science*
Academia Sinica
Nankang, 115 Taiwan

The popularity and ubiquity of multimedia associated with spoken documents has spurred a lot of research interest in spoken document retrieval (SDR) in the recent past. Beyond much effort devoted to developing robust indexing and modeling techniques for representing spoken documents, a recent line of thought targets at the improvement of query modeling for better reflecting the user's information need. Pseudo-relevance feedback is by far the most commonly-used paradigm for query reformulation, which assumes that a small amount of top-ranked feedback documents obtained from the initial round of retrieval are relevant and can be utilized for this purpose. Nevertheless, simply taking all of the top-ranked feedback documents obtained from the initial retrieval for query modeling does not always perform well, especially when the top-ranked documents contain much redundant or non-relevant information. In the view of this, we explore in this paper an interesting problem of how to effectively glean useful cues from the top-ranked documents so as to achieve more accurate query modeling. Towards this end, various sources of information cues are considered and integrated into the process of feedback document selection so as to achieve better retrieval effectiveness. Furthermore, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the query and document models. A series of experiments conducted on the TDT (Topic Detection and Tracking) task seem to demonstrate the effectiveness of our query modeling framework for SDR.

Keywords: spoken document retrieval, language modeling, query modeling, pseudo-relevance feedback, speech recognition

1. INTRODUCTION

Over the last two decades, spoken document retrieval (SDR) has become an active area of research and experimentation in the speech processing community. This can be attributed in large part to the advances in automatic speech recognition (ASR) [1, 2] and the ever-increasing volumes of multimedia associated with spoken documents made available to the public, such as broadcast news stories, meeting and lecture recordings, telephone conversations, digital archives, among many others [3-6]. Although most retrieval systems participating 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

Received February 28, 2013; accepted June 15, 2013.
Communicated by Hung-Yu Kao, Tzung-Pei Hong, Takahira Yamaguchi, Yau-Hwang Kuo, and Vincent Shin-Mu Tseng.

from a speech recognizer [7], this is probably due to the fact that the TREC-style test queries tend to be quite long and contain different words describing similar concepts that could help the queries match their relevant spoken documents. Furthermore, a query word (or phrase) might occur repeatedly (more than once) within a relevant spoken document, and it is not always the case that all of the occurrences of the word would be misrecognized totally as other words. Nevertheless, we believe that there are still at least two fundamental challenges facing SDR. On one hand, the imperfect speech recognition transcript carries wrong information and thus would deviate somewhat from representing the true theme of a spoken document. On the other hand, a query is often only a vague expression of an underlying information need, and there probably would be word usage mismatch between a query and a spoken document even if they are topically related to each other.

A significant body of SDR work has been placed on the exploration of robust indexing or modeling techniques to represent spoken documents in order to work around (or mitigate) the problems caused by ASR [5, 8-11]. On the contrary, very limited research has been conducted to look at the other side of the coin, namely, the improvement of query formulation for better reflecting the underlying information need of a user [12]. As for the latter problem, pseudo-relevance feedback [6, 13] is by far the most commonly-used paradigm, which assumes that a small amount of top-ranked spoken documents obtained from the initial round of retrieval are relevant and can be utilized for query reformulation. Subsequently, the SDR system can perform a second round of retrieval with the enhanced query representation to search for more relevant documents. We had recently introduced a new perspective on query modeling [12, 14, 15], saying that it can be approached with pseudo-relevance feedback and the language modeling (LM) retrieval approach [16] leveraging the notion of relevance [17], which seems to show preliminary promise for query reformulation. The success of such query modeling depends largely on the assumption that the set of top-ranked feedback documents obtained from the initial round of retrieval are relevant and can be used to estimate a more accurate query model. However, simply taking all of the top-ranked feedback documents obtained from the initial round of retrieval does not always work well for query modeling (or reformulation), especially when the top-ranked documents contain much redundant or non-relevant information [14, 15].

Our work in this paper continues this general line of research on query formulation for SDR. We explore an interesting problem of how to effectively glean useful cues from the top-ranked feedback documents so as to achieve more accurate query modeling. Towards this end, various sources of information cues are considered and integrated to select representative feedback documents for better retrieval effectiveness. In addition, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the LM-based query and document models. The rest of this paper is organized as follows. In Section 2, we briefly review the basic mathematical formulations of the LM-based retrieval models for SDR, as well as the idea of pseudo-relevance feedback. In Section 3, we describe and explain several cues we explore to select representative feedback documents during pseudo-relevance feedback. After that, the experimental settings and a series of retrieval experiments are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper and outlines possible directions for future work.

2. LANGUAGE MODELING FOR SPOKEN DOCUMENT RETRIEVAL

2.1 Query-Likelihood Measure

Recently, language modeling (LM) has emerged as a promising approach to building SDR systems [9, 11, 12]. This is due to the fact that the LM approach has inherent clear probabilistic foundation and excellent retrieval performance [16]. 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 (the so-called query-likelihood measure). A spoken document D is deemed to be relevant with respect to the query Q if the corresponding document model is more likely to generate the query. If the query Q is treated as a sequence of words, $Q = w_1, w_2, \ldots, w_L$, 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 [16]:

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

where $P(w_l|D)$ is the likelihood of generating w_l by document D (a.k.a. the document model). The simplest way to construct $P(w_l|D)$ is based on literal term matching [3], or using the unigram language model (ULM). To this end, each document D can, respectively, offer a unigram distribution for observing any given word w, which is parameterized on the basis of the empirical counts of words occurring in the document with the maximum likelihood (ML) estimator [1, 16]:

$$P(w|D) = \frac{c(w, D)}{|D|}, \tag{2}$$

where $c(w, D)$ is the number of times that word w occurs in the document D and $|D|$ is the number of words in the document. The document model is further smoothed by a background unigram language model estimated from a large general collection to model the general properties of the language as well as to avoid the problem of zero probability [16]. However, how to strike the balance between these two probability distributions is actually a matter of judgment, or trial and error.

Furthermore, a family of topic modeling methods has been proposed as an effective complement to the ULM model. Topic models attempt to discover the latent topic information embedded in the query and documents, based on which the retrieval is performed. Among them, probabilistic latent semantic analysis (PLSA) [18] and latent Dirichlet allocation (LDA) [19] are often considered to be the two best-known instantiations. They both introduce a set of latent topic variables $\{T_1, \ldots, T_k, \ldots, T_K\}$ 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. For example, in the PLSA model [16], the probability of a word w generated by a document D is expressed by

$$P(w|D) = \sum_{k=1}^{K} P(w|T_k)P(T_k|D) \tag{3}$$

where $(w|T_k)$ denotes the probability of a certain type of word w occurring in a specific latent topic T_k, and $P(T_k|D)$ is the posterior probability (or weight) of topic T_k conditioned D. The model parameters of PLSA can be estimated using the expectation-maximization (EM) algorithm [20]. On the other hand, LDA, having a formula analogous to PLSA (*cf.* (3)) for SDR, is regarded as an extension to PLSA and has enjoyed much success for various speech and language applications. 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 [12]. Since LDA has a more complex form for model optimization, it is hardly to be solved by exact inference. Several approximate inference algorithms, such as the variational approximation algorithm, the expectation propagation method and the Gibbs sampling algorithm, hence have been proposed for estimating the parameters of LDA. Furthermore, due to the fact that PLSA and LDA offer coarse-grained latent semantic representation 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 a given implementation, there is always good reason to combine them with ULM for better retrieval quality [21, 22].

2.2 Kullback-Leibler (KL)-Divergence Measure

Another basic formulation of LM for SDR is the Kullback-Leibler (KL)-divergence measure [16, 23]:

$$\begin{aligned} -KL(Q\,\|\,D) &= -\sum_{w\in V} P(w\,|\,Q)\log\frac{P(w\,|\,Q)}{P(w\,|\,D)} \\ &\overset{\text{rank}}{=} \sum_{w\in V} P(w\,|\,Q)\log P(w\,|\,D), \end{aligned} \tag{4}$$

where the query and the document are, respectively, framed as a (unigram) language model (*i.e.*, $P(w|Q)$ and $P(w|D)$), $\overset{\text{rank}}{=}$ means equivalent in terms of being used for the purpose of ranking documents, and V denotes the vocabulary. A document D has a smaller value (or probability distance) in terms of $KL(Q\|D)$ is deemed to be more relevant with respect to Q. The retrieval effectiveness of the KL-divergence measure depends primarily on the accurate estimation of the query modeling $P(w|Q)$ and the document model $P(w|D)$. In addition, it is easy to show that the KL-divergence measure will give the same ranking as the ULM model (*cf.* (1)) when the query language model is simply derived with the ML estimator [12]:

$$\begin{aligned} -KL(Q\|D) &\overset{\text{rank}}{=} \sum\nolimits_{w\in V} P(w\,|\,Q)\log P(w\,|\,D) \\ &= \sum\nolimits_{w\in V} \frac{c(w,Q)}{|Q|}\log P(w\,|\,D) \\ &\overset{\text{rank}}{=} \sum\nolimits_{w\in V} c(w,Q)\log P(w\,|\,D) \end{aligned} \tag{5}$$

$$= \log P(Q \mid D)$$
$$\overset{\text{rank}}{=} P(Q \mid D).$$

In Eq. (5), $P(w|Q)$ is simply estimated as $c(w, Q)/|Q|$, where $c(w, Q)$ is the number of times w occurring in Q and $|Q|$ is the total count of words in Q. Accordingly, the KL-divergence measure not only can be thought as a generalization of the query-likelihood measure, but also has the additional merit of being able to accommodate extra information cues to improve the estimation of its component models (especially, the query model) for better document ranking in a systematic manner [12, 16].

2.3 Pseudo-Relevance Feedback

In reality, since a query often consists of only a few words, the query model that is meant to represent the user's information need might not be appropriately estimated by the ML estimator. Furthermore, merely matching words between a query and documents might not be an effective approach, as the word overlaps alone could not capture the semantic intent of the query. To cater for this, an LM-based SDR system with the KL-divergence measure can adopt the idea of pseudo-relevance feedback and perform two rounds of retrieval to search for more relevant documents. In the first round of retrieval, an initial query is input into the SDR system to retrieve a number of top-ranked feedback documents. Subsequently, on top of these top-ranked feedback documents, a refined query model is constructed and a second round of retrieval is conducted with this new query model and the KL-divergence measure depicted in Eq. (4). It is usually anticipated that the SDR system can thus retrieve more documents relevant to the query.

However, an LM-based SDR system with the pseudo-relevance feedback process may confront two intrinsic challenges. One is how to purify the top-ranked feedback documents obtained from the first round of retrieval so as to remove redundant and non-relevant information. The other is how to effectively utilize the selected set of representative feedback documents for estimating a more accurate query model. For the latter, there are a number of studies proposing various query modeling techniques directly exploiting the top-ranked feedback text (or spoken) documents, such as the simple mixture model (SMM) [24], the relevance model (RM) [17] and their extensions [12], among others. However, for the former, there is relatively little work done on selecting useful and representative feedback documents from the top-ranked ones for SDR, as far as we are aware. Recently, the so-called "Gapped Top *K*" and "Cluster Centroid" selection methods [25] have been proposed for text information retrieval (IR). "Gapped Top *K*" selects top *K* documents with a ranking gap *J* in between any two top-ranked documents, while "Cluster Centroid" groups the top-ranked documents into *K* clusters and selects one representative document from each cluster to obtain diversified feedback documents. Another more attractive and sophisticated method proposed for text IR is "Active-RDD" [26], which takes into account the relevance, diversity and density cues of the top-ranked documents for feedback document selection. The above three methods have not been extensively studied for SDR.

In this paper, we go a step further by additionally exploring the non-relevance cue during feedback document selection, apart from the relevance, diversity and density cues.

As we will see later, the additional use of the non-relevance cue can further boost the SDR performance. In particular, the resulting feedback document selection method can effectively work in tandem with various query modeling techniques and different granularities of index features.

3. LEVERAGING EXTRA CUES FOR PSEUDO-RELEVANCE FEEDBACK

Our SDR system first takes the initial query and employs the ULM retrieval model to obtain a number of top-ranked documents $\mathbf{D}_{\text{Top}} = \{D_1, D_2, \ldots, D_N\}$. Then in the pseudo-relevance feedback process, the system iteratively selects documents from $\mathbf{D}_{\text{Top}}$ to form a representative set of feedback documents by simultaneously considering the relevance, non-relevance, diversity and density cues. More specifically, each candidate feedback document D is associated with a score that is a linear combination of measures of these cues, expressed as follows:

$$D^* = \arg\max_{D \in \mathbf{D}_{\text{Top}} - \mathbf{D}_{\text{P}}} \Big[(1-\alpha-\beta-\gamma)\cdot M_{Rel}(Q,D) + \alpha \cdot M_{NR}(Q,D) + \beta \cdot M_{Diversity}(D) + \gamma \cdot M_{Density}(D)\Big], \tag{6}$$

where $\mathbf{D}_{\text{P}}$ is the set of already selected feedback documents; $M_{Rel}(Q, D)$, $M_{NR}(Q, D)$, $M_{Diversity}(D)$ and $M_{Density}(D)$ are measures of relevance, non-relevance, diversity and density for each document D in $\mathbf{D}_{\text{Top}}$, respectively; α, β and γ are weighting coefficients. The selection process illustrated in Eq. (6) will be executed iteratively until $\mathbf{D}_{\text{P}}$ contains a predefined number of feedback documents. It is worth mentioning that the method described in Eq. (6) bears a close resemblance in spirit to the maximal marginal relevance (MMR) ranking algorithm [27, 28] which was originally proposed for extractive document summarization. Note also that $M_{Rel}(Q, D)$ is just the similarity (query-likelihood) measure of the ULM retrieval model depicted in Eq. (1). In the following, we will describe how to model the other information cues we explore for a given candidate feedback document.

3.1 Non-Relevance Measure

For a given query Q, we can estimate a non-relevance model $P(w|NR_Q)$ of it based on the low-ranked documents obtained from the initial round of retrieval, and the non-relevance measure of a candidate feedback document D is thus defined by

$$M_{NR}(D) = KL(NR_Q||D). \tag{7}$$

The additional incorporation of $M_{NR}(D)$ for feedback document selection will prefer those documents that have only a small probability distance to the original query model but also a larger probability distance to the non-relevance model. Since the number of relevant documents with respect to a given query is usually very small compared to that of non-relevant ones in practice, we may assume that the entire spoken document collec-

tion (more specifically, the background language model $P(w|BG)$) could offer an alternative estimate of the non-relevance model $P(w|NR_Q)$.

3.2 Diversity Measure

In the recent past, diversification of retrieval results has gained popularity in the text IR community, since it can be used to complement the conventional document ranking criteria which only consider relevance information and often suffer from returning too many redundant documents. By analogy, in the context of pseudo-relevance feedback, if we use the top-ranked documents that contain too much redundant information to estimate the query model, then the second round of retrieval is prone to return too many "redundant" documents to the user. In order to diversify the selected feedback documents for better query reformulation, we compute the diversity measure of a candidate feedback document with respect to the set $\mathbf{D}_P$ of already selected feedback documents, which is expressed as follows:

$$M_{Diversity}(D) = \min_{D_j \in \mathbf{D}_P} \frac{1}{2} \cdot [KL(D_j \,||\, D) + KL(D \,||\, D_j)]. \tag{8}$$

3.3 Density Measure

Intuitively, the structural information among the top-ranked documents can be taken into account as well during feedback document selection. For this idea to work, we can compute the average negative, symmetric probability distance between a document D and all the other documents D_h in $\mathbf{D}_{\text{Top}}$, which is expressed as follows:

$$M_{Density}(D) = \frac{-1}{|\mathbf{D}_{\text{Top}}| - 1} \cdot \sum_{\substack{D_h \in \mathbf{D}_{\text{Top}} \\ D_h \neq D}} [KL(D_h \,||\, D) + KL(D \,||\, D_h)], \tag{9}$$

where $|\mathbf{D}_{\text{Top}}|$ is the number of documents in $\mathbf{D}_{\text{Top}}$. A document D having a higher value of $M_{Density}(D)$ is deemed to be closer to the other documents in $\mathbf{D}_{\text{Top}}$ and thus to be more representative (and less likely to be an outlier).

4. EXPERIMENTAL SETUP

4.1 Spoken Document Collection

We used the Topic Detection and Tracking collection (TDT-2) [12, 29, 30] for this work. 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. The average word error rate (WER) of the spoken documents is about 35% [30]. 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

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

(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 [6]:

$$\text{mAP} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}} \tag{10}$$

where I 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 jth document that is relevant to query Q_i, counting down from the top of the ranked list.

Table 1 shows some basic statistics about the TDT-2 collection. In order to evaluate the performance of the various feedback document selection methods studied in this paper, the number of the top-ranked documents obtained from the first round of retrieval is set to 25 (*i.e.*, $|\mathbf{D}_{\text{Top}}| = 25$) and the target number of selected feedback document is set to 5 (*i.e.*, $|\mathbf{D}_{\text{P}}| = 5$, unless otherwise stated. Albeit that, it is known that the way to systemically determine the values of the free parameters that the feedback document selection methods, as well as the retrieval models, incorporate is still an open issue and needs further investigation and proper experimentation.

4.2 Query Modeling

In this paper, we employ RM and SMM for query reformulation in combination with the various feedback document selection methods studied in this paper. For a given query $Q = w_1, w_2, \ldots, w_L$, the refined query model based on RM [11] is formulated by

$$P_{\text{RM}}(w|Q) = \frac{\sum_{D_j \in \mathbf{D}_{\text{P}}} P(D_j)P(w \mid D_j)\prod_{l=1}^{L} P(w_l \mid D_j)}{\sum_{D_j \in \mathbf{D}_{\text{P}}} P(D_j)\prod_{l=1}^{L} P(w_l \mid D_j)}, \tag{11}$$

where the probability $P(D_j)$ can be simply kept uniform or determined in accordance with the relevance of D_j to Q, while $P(w|D_j)$ and $P(w_l|D_j)$ are estimated on the grounds of the word occurrence counts in D_j with the ML estimator. The RM model assumes that words w that co-occur with the query Q in the feedback documents will have higher probabilities. We had recently presented an improved version of the RM model by further incorporating a set of latent topics into the modeling of $P(w|D_j)$ and $P(w_l|D_j)$, referred to as the topic-based relevance model (TRM) hereafter. Just as with PLSA and

LDA (*cf.* (3)), TRM introduces two sets of probability distributions, *i.e.*, $P(w|T_k)$ and $P(T_k|D_j)$, to describe the "*word-document*" relationship:

$$P_{\mathrm{TRM}}(w|Q) = \frac{\sum_{D_j \in \mathbf{D}_P} \sum_{k=1}^{K} P(D_j) P(T_k \mid D_j) P(w \mid T_k) \prod_{l=1}^{L} P(w_l \mid T_k)}{\sum_{D_j \in \mathbf{D}_P} \sum_{k=1}^{K} P(D_j) P(T_k \mid D_j) \prod_{l=1}^{L} P(w_l \mid T_k)}. \tag{12}$$

The model parameters $P(w|T_k)$ and $P(T_k|D_j)$ can be estimated by maximizing the total log-likelihood of the spoken document collection in terms of the unigram of all document words observed. 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 [12].

On the other hand, SMM [24] assumes words in the set of feedback documents $\mathbf{D}_\mathrm{P}$ are drawn from two models: (1) the feedback model $P(w|FB)$ and (2) the background model $P(w|BG)$. The feedback model $P(w|FB)$ is estimated by maximizing the log-likelihood of the set of feedback documents $\mathbf{D}_\mathrm{P}$ expressed as follows, using the EM algorithm:

$$LL_{\mathbf{D}_\mathrm{P}} = \sum_{D_j \in \mathbf{D}_\mathrm{P}} \sum_{w \in V} c(w, D_j) \log[\lambda \cdot P(w|FB) + (1-\lambda) \cdot P(w \mid BG)], \tag{13}$$

where $c(w, D_j)$ is the occurrence count of w in D_j and λ is the interpolation parameter used to control the degree of reliance on $P(w|FB)$ rather than on $P(w|BG)$. The maximization of Eq. (13) can be conducted iteratively via the following two EM update equations:

$$P^{(m)}(FB|w) = \frac{\lambda \cdot P^{(m)}(w|FB)}{\lambda \cdot P^{(m)}(w|FB) + (1-\lambda) \cdot P(w|BG)} \tag{14}$$

and

$$P^{(m+1)}(w \mid FB) = \frac{\sum_{D_j \in \mathbf{D}_\mathrm{P}} c(w, D_j) \cdot P^{(m)}(FB|w)}{\sum_{w} \sum_{D_j \in \mathbf{D}_\mathrm{P}} c(w, D_j) \cdot P^{(m)}(FB|w)}, \tag{15}$$

where m denotes the mth iteration of the EM algorithm and $c(w, D_j)$ is the number of times w occurring in D_j. The resulting feedback model can be linearly combined with or used to replace the original query model. A schematic illustration of the SDR process is shown in Fig. 1.

4.3 Subword-level Index Units

In an effort to alleviate SDR performance degradation caused by imperfect speech recognition, we also utilize different levels of index features for construct the query and document models involved in the KL-divergence measure, including words, syllable-level units, and their combination. To do this, syllable pairs are taken as the basic units for indexing besides words. The recognition transcript of each spoken document, in form of a word stream, was 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 a vocabulary of syllable pairs for indexing. We can simply use syllable pairs, in replace of words, to represent the spoken documents and test queries, and subsequently construct the associated language model distributions.

Furthermore, it is well acknowledged that word-level indexing features possess more semantic information than subword-level features; hence, word-based retrieval enhances precision. On the other hand, subword-level indexing features behave more robustly against the homophone ambiguity, open vocabulary problem, and speech recognition errors; hence, subword-based retrieval enhances recall. Accordingly, there is good reason to fuse the information obtained from indexing the features of different levels [12, 31].

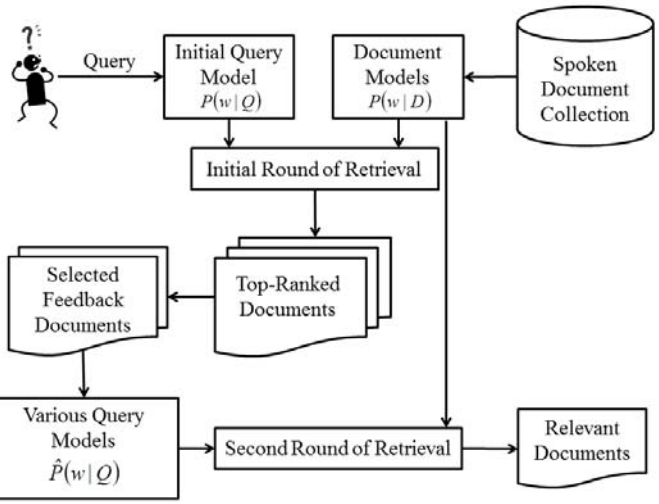

Fig. 1. A schematic illustration of the SDR process.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1 Baseline Experiments

At the outset, we compare the performance of RM, TRM and SMM when the top-ranked (*i.e.*, top 25) documents obtained from the initial round of retrieval is used for constructing the refined query models. The corresponding results are shown in Table 2, where the results of ULM, PLSA and LDA are listed as well for reference. As mentioned earlier in Section 2.1, PLSA and LDA are two state-of-the-art (more sophisticated) LM-based retrieval models [21, 22], both of which incorporate a set of latent topics for representing (spoken) documents and determine the ranking of spoken documents with respect to a query based on the query-likelihood measure. It is also worth mentioning that ULM, PLSA and LDA perform retrieval only with the initial query. Inspection of Table 2 reveals three noteworthy points. First, the performance gap between the retrieval using manual transcripts (denoted by TD) and the recognition transcripts (denoted by SD) is about 0.05 in terms of MAP, such degradation is apparently less pronounced as com-

Table 2. Retrieval results (in MAP) achieved by various retrieval models.

	ULM	PLSA	LDA	RM	TRM	SMM
TD	0.371	0.418	0.401	0.421	0.456	0.415
SD	0.323	0.435	0.341	0.369	0.397	0.361

pared to the WER of spoken documents [12]. Second, RM and SMM tend to perform on par with each other, and they deliver substantial improvements over ULM (and perform comparably to PLSA and LDA). Third, TRM exhibits superior performance over RM and SMM, which confirms the merits of leveraging topical information for query modeling.

5.2 Experiments on Feedback Document Selection

In the second set of experiments, we evaluate the utility of the various feedback document selection methods investigated in this paper, including "Gapped Top *K*" (denoted by "Gapped" for short), "Cluster Centroid" (denoted by "Cluster" for short), "Active-RDD" and our purposed method (*cf.* Sections 2 and 3), in concert with some of the above retrieval (query) models (the number of selected feedback documents is set to 5). The corresponding results are shown in Table 3, whereas the results of simply using the top *N* (*N* = 5, 10, 15, 20, 25 or 30) documents obtained from the initial round of retrieval to construct the refined query models are listed in Table 4 for comparison. A closer look at these results reveals three things. First, using either "Active-RDD" or our proposed method to select feedback documents seems to outperform that simply using the top *N* (*N* = 5, 10, 15, 20, 25 or 30) documents (*cf.* Table 4) obtained from the initial round of retrieval as the feedback documents by a big margin, indicating that appropriate feedback document selection is critical to the success of query reformulation. Second, our proposed method delivers better performance gains over "Active-RDD" for all cases, which exhibits the advantage of using the non-relevance cue for feedback document selection. Third, "Gapped Top *K*" and "Cluster Centroid" both result in performance that appears to be much inferior to that of "Active-RDD" and our proposed method.

Table 3. Retrieval results (in MAP) achieved by various combinations of retrieval models and feedback document selection methods.

		RM	TRM	SMM
TD	Gapped	0.414	0.452	0.406
	Cluster	0.396	0.441	0.380
	Active-RDD	0.471	0.492	0.457
	Our Method	0.491	0.507	0.490
	Our Method + TW	0.523	0.522	0.496
SD	Gapped	0.357	0.391	0.333
	Cluster	0.378	0.395	0.325
	Active-RDD	0.437	0.461	0.403
	Our Method	0.448	0.475	0.424
	Our Method + TW	0.485	0.494	0.435

Table 4. Retrieval results (in MAP) achieved when simply using the top 5, 10, 15, 25 or 30 documents obtained from the initial round of retrieval for constructing various query models.

		RM	TRM	SMM
TD	Top 5	0.405	0.440	0.438
	Top 10	0.417	0.452	0.483
	Top 15	0.421	0.455	0.468
	Top 25	0.421	0.456	0.415
	Top 30	0.421	0.457	0.411
SD	Top 5	0.369	0.396	0.399
	Top 10	0.372	0.398	0.398
	Top 15	0.370	0.397	0.367
	Top 25	0.369	0.397	0.361
	Top 30	0.369	0.396	0.360

Table 5. Retrieval results (in MAP) for the SD case, achieved by using words, syllable-level units, and their combination for construct the query and document models.

	RM	TRM	SMM
Word	0.485	0.494	0.435
Syllable	0.507	0.510	0.484
Word+Syllable	0.531	0.521	0.505

5.3 Experiments on IDF-Based Term Weighting

In the third set of experiments, we explore to emphasize the roles of those words occurring in the feedback documents that have higher descriptive capabilities in the estimation of the refined query models. To this end, when estimating the refined query models, the occurrence count of a given word in a feedback document is multiplied (or weighted) by its corresponding inverse document frequency (IDF). IDF, indicating how predictive a word is, typically is expressed as a function of the inverse logarithm of the number of documents that contain the word [6]. It is evident from Table 3 that utilizing such an IDF-based weighting scheme (denoted by TW for short) can further improve the retrieval performance, in combination with the various retrieval modeling techniques (*cf.* Rows "Our Method" vs. "Our Method+TW" in Table 3). Furthermore, as compared to the baseline results of ULM and LDA shown in Table 2, it corroborates that more elaborate query modeling is of paramount importance to an LM-based SDR system.

5.4 Fusion of Different Levels of Indexing Features

In the final set of experiments, we investigate how the word- and syllable-level index features complement each other in representing both the test queries and spoken documents. The results for the SD case are shown in Table 5, as a function of different query modeling techniques being used (*i.e.*, RM, TRM and SMM). One thing to note is that these query modeling techniques are implemented in combination with our feedback document selection method and IDF-based term weighting method (denoted by "Our Method + TW"), as discussed previously in Section 5.3. As can been seen from Table 5, the results for the various query modeling techniques, in general, have consistent trends

with that of the previous experiments. In particular, there are two noteworthy points to these results. First, the subword-level (syllable-level) index features seem to show competitive or even better performance than the word-level index features when being used for retrieving spoken documents on top of imperfect recognition transcripts (*i.e.*, for the SD case). Second, not surprisingly, compared to the results of using either the word- or syllable-level index features in isolation, fusion of these two levels of index features can inherit their advantages so as to achieve better performance. It, therefore, implies that fusion of different granularities of index features works well for SDR, in concert with the presented feedback document selection method and IDF-based term weighting method scheme.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a novel framework to incorporate various sources of information cues into the process of feedback document selection for enhanced LM-based query formulation (modeling) in SDR. The retrieval methods deduced from this framework have also been validated by extensively comparisons with several existing methods. The corresponding experimental results demonstrate the applicability of our methods for SDR. They also reveal that effective query formulation is critical to the success of an SDR system. On the other hand, there exist diverse characteristics of (spoken) documents and test queries, such as the length and word usage for each document and each query, and the number of relevant documents for each query, among others, which would also significantly affect the ultimate performance of pseudo-relevance feedback (PRF) and query formulation in the LM-based retrieval framework. Although a detailed and comprehensive analysis of the aforementioned factors seems to be beyond the scope of this paper that was set out to develop effective modeling techniques for PRF and query formulation, this is left to our future work. In addition, we plan to explore more extra information cues and different model (and parameter) optimization criteria [28, 32, 33] for feedback document selection. We are interested, as well, in investigating more robust indexing and sophisticated modeling techniques [5, 9, 19] to represent spoken documents, in the hope that they can bring additional gains when being used in association with the various query modeling techniques and feedback document selection methods for larger-scale SDR tasks. Among other things, we would like to make use of this LM-based framework for speech recognition and summarization applications [34, 35].

ACKNOWLEDGEMENT

This research is supported in part by the “Aim for the Top University Project” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, and by the Ministry of Science and Technology, Taiwan, under Grants NSC 101-2221-E-003-024-MY3, NSC 102-2221-E-003-014-, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 and NSC 103-2911-I-003-301.

REFERENCES

1. F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, MA, 1999.

2. S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda, "Fundamental technologies in modern speech recognition," *IEEE Signal Processing Magazine*, Vol. 29, 2012, pp. 16-17.
3. L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, Vol. 22, 2005, pp. 42-60.
4. M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, Vol. 25, 2008, pp. 150-152.
5. C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, Vol. 25, 2008, pp. 39-49.
6. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, Pearson Education Limited, Edinburg, England, 2011.
7. J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the 8th Text Retrieval Conference*, 2000, pp. 107-129.
8. V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 631-638.
9. B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM Transactions on Asian Language Information Processing*, Vol. 8, 2009, pp. 2:1-2:27.
10. S. Parlak and M. Saraçlar, "Performance analysis and improvement of Turkish broadcast news retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, 2012, pp. 731-743.
11. T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "Statistical lattice-based spoken document retrieval," *ACM Transactions on Information Systems*, Vol. 28, 2010, pp. 2:1-2:30.
12. B. Chen, K.-Y. Chen, P.-N. Chen, and Y.-W. Chen, "Spoken document retrieval with unsupervised query modeling techniques," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, 2012, pp. 2602-2612.
13. J. Rocchio, "Relevance feedback in information retrieval," in G. Salton, ed., *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313-323, 1971, Prentice Hall.
14. Y.-W. Chen, K.-Y. Chen, and B. Chen, "Empirical comparisons of various pseudo-relevant document selection methods for improved spoken document retrieval," in *Proceedings of the Conference on Technologies and Applications of Artificial Intelligence*, 2012, pp. 140-147.
15. Y.-W. Chen, K.-Y. Chen, H.-M. Wang, and B. Chen, "Effective pseudo-relevance feedback for spoken document retrieval," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 8535-8539.
16. C. X. Zhai, "Statistical language models for information retrieval: A critical review," *Foundations and Trends in Information Retrieval*, Vol. 2, 2008, pp. 137-213.
17. V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001, pp. 120-127.
18. T. Hoffmann, "Unsupervised learning by probabilistic latent semantic analysis,"

Machine Learning, Vol. 42, 2001, pp. 177-196.
19. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
20. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, Vol. 39, 1977, pp. 1-38.
21. X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 178-185.
22. Y. Lu, Q. Mei, and C. X. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Information Retrieval*, Vol. 14, 2011, pp. 178-203.
23. S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, 1951, pp. 79-86.
24. C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of ACM SIGIR Conference on Information and Knowledge Management*, 2001, pp. 403-410.
25. X. Shen and C. Zhai, "Active feedback in ad hoc information retrieval," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 55-66.
26. Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *Proceedings of European Conference on IR Research*, 2007, pp. 245-257.
27. J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 335-336.
28. B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, 2012, pp. 199-210.
29. LDC, "Project topic detection and tracking," *Linguistic Data Consortium*, 2000.
30. H. Meng, B. Chen, S. Khudanpur, G. A. Levow, W. K. Lo, D. Oard, P. Schone, K. Tang, H. M. Wang, and J. Wang, "Mandarin–English information (MEI): investigating translingual speech retrieval," *Computer Speech and Language*, Vol. 18, 2004, pp. 163-179.
31. K.-Y. Chen, H.-M. Wang, and B. Chen, "Spoken document retrieval leveraging unsupervised and supervised topic modeling techniques," *IEICE Transactions on Information and Systems*, Vol. E95-D, 2012, pp. 1195-1205.
32. Y. Lv, C. X. Zhai, and W. Chen, "A boosting approach to improving pseudo-relevance feedback," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 165-174.
33. K. S. Lee and W. B. Croft, "A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback," *Information Processing and Management*, Vol. 40, 2013, pp. 792-806.
34. B. Chen and K.-Y. Chen, "Leveraging relevance cues for language modeling in speech recognition," *Information Processing and Management*, Vol. 40, 2013, pp. 807-816.

35. B. Chen, S.-H. Lin, Y.-M. Chang, and J.-W. Liu, "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing and Management*, Vol. 49, 2013, pp. 1-12.

Berlin Chen (陳柏琳) received the B.S. and M.S. degrees in Computer Science and Information Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1994 and 1996, respectively, and the Ph.D. degree in Computer Science and Information Egineering from National Taiwan University, Taipei, in 2001.

In 2002, he joined the Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal Unirsity, Taipei. He is currently a Professor in the Department of Computer Science and Information Engineering of the same university. His research interests generally lie in the areas of speech and natural language processing, information retrieval, and artificial intelligence; he is the author/coauthor of over 100 academic publications. Dr. Chen is a member of IEEE, ISCA and ACLCLP.

Yi-Wen Chen (陳憶文) received the B.A. and M.S. degrees in Applied Foreign Languages from National Formosa University, Yunlin, Taiwan, and Computer Science and Information Engineering from National Taiwan Normal University, Taipei, Taiwan, in 2008 and 2013, respectively. Her research interests are in the fields of spoken document retrieval, language modeling and natural language processing.

Kuan-Yu Chen (陳冠宇) received the B.S. and M.S. degrees in Computer Science and Information Engineering from National Taiwan Normal University, Taipei, Taiwan, in 2007 and 2010, respectively. Currently, he is a Ph.D. candidate of the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

In November 2010, he joined the Speech, Language, and Music Processing Laboratory, Institute of Information Science, Academia Sinica, Taiwan, as a Research Assistant. His research interests are in the fields of language modeling, speech recognition, information retrieval and natural language processing.

Hsin-Min Wang (王新民) received the B.S. and Ph.D. degrees in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively.

In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is now a Research Fellow. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, and pattern recognition.

Dr. Wang received the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He currently serves as the President of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) and as an editorial board member of the International Journal of Computational Linguistics and Chinese Language Processing and Journal of Information Science and Engineering. He is a life member of ACLCLP and Institute of Information and Computing Machinery (IICM), a senior member of IEEE, and a member of International Speech Communication Association (ISCA) and ACM.

Kuen-Tyng Yu (余坤庭) is currently a Ph.D. student of the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. His research interests are in the fields of computer-aided language learning, information retrieval and natural language processing.