國立臺灣師範大學
資訊工程研究所碩士論文

指導教授：陳柏琳 博士

探索虛擬關聯回饋技術和鄰近資訊
於語音文件檢索與辨識之改進

Exploring Effective Pseudo-Relevance Feedback
and Proximity Information
for Speech Retrieval and Transcription

研究生：陳憶文 撰

中華民國 一百零二 年 七 月

Exploring Effective Pseudo-Relevance Feedback and Proximity Information for Speech Retrieval and Transcription

Master Thesis

by

Yi-Wen Chen
699470292@ntnu.edu.tw
Department of Computer Science and Information Engineering
National Taiwan Normal University
July, 2013

摘要

虛擬文件檢索(Pseudo-Relevance Feedback)為目前最常見的查詢重建(Query Reformulation)典範。它假設預檢索(Initial-round of Retrieval)排名前端的文件都是相關的，所以可全用於查詢擴展(Query Expansion)。然而，預檢索所獲得的文件中，極可能同時包含重複性資訊(Redundant)和非關聯(Non-relevant)資訊，使得重新建立的查詢不能有良好檢索效能。有鑑於此，本論文探討運用不同資訊以在預檢索獲得的語音文件中挑選適當的關聯文件來建立查詢表示，讓語音文件檢索結果可以更準確。另一方面，關聯模型(Relevance Model)雖然可藉由詞袋(Bag-of-words)假設來簡化模型推導和估測，卻可能因此過度簡化問題，特別是用於語音辨識的語言模型。為了調適關聯模型，本論文有兩個貢獻。其一，本論文提出詞鄰近資訊使用於關聯模型以改善詞袋(Bag-of-words)假設於語音辨識的不適。其二，本論文也進一步探討主題鄰近資訊以強化鄰近關聯模型的架構。實驗結果證明本論文所提出之方法，不論在語音文件檢索還是語音辨識方面皆可有效改善現有方法的效能。

關鍵詞：語音文件檢索、語音辨識、語言模型、虛擬關聯回饋、鄰近資訊

Abstract

Pseudo-relevance feedback is by far the most commonly-used paradigm for query reformulation in spoken document retrieval, which assumes that a small amount of top-ranked feedback documents obtained from the initial retrieval are relevant and can be utilized for query expansion. Nevertheless, simply taking all of the top-ranked feedback documents acquired from the initial retrieval for query modeling does not necessary work well, especially when the top-ranked documents contain much redundant or non-relevant cues. In view of this, we explore different kinds of information cues for selecting helpful feedback documents to further improve information retrieval. On the other hand, relevance model (RM) based on "*bag-of-words*" assumption, which can facilitate the derivation and estimation, may be oversimplified for the task of language modeling in speech recognition. Hence, we also enhance RM in two significant aspects. First, "*bag-of-words*" assumption of RM is relaxed by incorporating word proximity information into RM formulation. Second, topic-based proximity information is additionally explored to further enhance the proximity-based RM framework. Experiments conducted on not only a spoken document retrieval task but also a speech recognition task indicates that our approaches can bring competitive utilities to existing ones.

Keywords: Spoken Document Retrieval, Speech Recognition, Language Modeling, Pseudo-Relevance Feedback, Proximity

Contents

&

List of Figures

List of Tables

1

Introduction

1.1 Motivation

With ongoing multimedia technology evolution, ever-increasing amounts of multimedia whether represented as static texts or audio-visual multimedia has given us tremendous amounts of information. Accompanying exponential proliferation of multimedia related to spoken documents, research on spoken document retrieval (SDR) has received growing amount of interest from researchers and practitioners over the past two decades. The advances of automatic speech recognition (ASR) and the unprecedented volumes of multimedia associated with spoken documents made available to the public, such as broadcast news stories, lecture or meeting recordings, telephone conversations and many others, are the two main reasons [1-3].

1.1.1 Spoken Document Retrieval

Unlike spoken term detection (STD), research on STD usually targets at the probable extraction of spoken terms or phrases inherent in a spoken document that could match

the query words or phrases literally. [2] However, research on SDR pays more attention to the notion of relevance of a spoken document related to a given query [4]. Typically, a document is deemed to be relevant if it could address the stated information need of the query, not just all the query terms overlap alone [5].

Even though merely using imperfect recognition transcripts produced from one-best recognition results, most retrieval systems participated in the TREC-SDR evaluations had claimed that speech recognition errors do not seem to cause very significant deterioration in terms of the retrieval quality [6]. This might partly due to the fact that the queries of TREC tend to be rather long and different in word usage which often describe a similar concept and hence further assist these queries in matching their relevant spoken documents. In addition, the same word in the corpus is not always misrecognized as well as a query word (or phrase) may repeat more than once within a truly relevant spoken document. Accordingly, though SDR apparently looks like a solved problem, there are three fundamental problems we believe it would still require facing in practice:

(I). A query is often a short and vague expression of an underlying information need

(II). Word usage mismatch between a query and a spoken document would probably happens even if these terms are topically related to each other

(III). The imperfect speech recognition transcript carries wrong information which

would drift away somewhat from representing the true theme of a spoken document

Language modeling (LM) approach is by far one of the most popular paradigms in building SDR systems [7-10]. This is attributable to the fact that the neat formulation of LM approach not only embraces impressive retrieval performance but clear probability meaning[11]. In terms of the general measurement of LM approach, the relevance (or similarity) measure between a query and a spoken document is typically computed by two different matching strategies, namely, literal term matching and concept matching. For literal term matching, the most popular instantiation is the unigram language model (ULM) [7-10]. In this class of methods, each document is regarded as a generative model composed of a mixture of unigram (multinomial) distributions for computing the likelihood of generating a query, which is usually expressed as a sequence of words (or index terms) as the document observation. Accordingly, ranking can be done by scoring documents' likelihood of observing the query respectively, that is, the so-called query-likelihood measure. More, the position or the order of term occurred in the document is assumed as unimportance, namely, '*bag-of-words*" assumption. Still, in order to improve ULM, there is a considerable work striving to further glean contextual information with *n*-grams of various orders, or some grammar structures; however, most result in mild gains or mixed results [11].

Since the aforementioned class of methods follow the thought of literal term matching, these methods inevitably would confront the problems of word usage diversity, which might lead to retrieval performance degradation for the differential in word usage between a given query and its corresponding relevant documents. In consequence of that, a family of topic modeling methods has been proposed. Topic models attempt at depicting the latent topic cues hidden in the query and documents. For instance, latent Dirichlet allocation (LDA) [12] and its precursor, probabilistic latent semantic analysis (PLSA) [13], are often treated as two typical examples of concept matching. Both of them employ a set of latent topic variables to portray the co-occurrence relationship between a word and a document. Thus, the relevance measure between a query and a document is instead estimated the frequency of query words occurred in the possible latent topics and the probability that the document observes the respective topics as well, which demonstrates some sort of concept matching. Although there are many follow-up researches devoting to extend LDA and PLSA, empirical results imply that more sophisticated (or complicated) topic models, such as Pachinko allocation model (PAM), can not provide further benefits for retrieval [14,15].

Although most of the aforementioned retrieval methods can be applied to not only text but also spoken documents without adaptation, the latter ones still suffer from

unique difficulties, such as speech recognition errors, or redundant information. Apart from many conventional researches that focus on boosting recognition accuracy, an intuitive idea is to directly develop more robust representations for spoken documents. For instance, aside from the top scoring ones, multiple recognition hypotheses can be constructed to derive alternative representations for the unclear part of the spoken documents [1,2,10]. Another line of research leverages subword-level index features or the combination of word- and subword-level index features to stand for the spoken documents, which also has been demonstrated beneficial to SDR. This might attribute to the fact that the incorrectly recognized spoken words often comprise several correctly recognized subword-level units. As a result, the retrieval process based on subword-level indexing of spoken documents may gain from partial matching [9,16,17].

In order to better represent spoken documents, a large body of SDR research has been devoted to exploring more robust indexing or modeling techniques [4,9,10,16,17], however, very limited work has been placed on the other side of the coin, that is, the possible improvement of query modeling for better reflecting the underlying information need of a user [18]. As for the latter problem, we had recently given a new picture of query modeling [18], which can be worked with pseudo-relevance feedback [5] to leverage the notion of relevance [19] and exhibits preliminary promise for query

reformulation. It is worth mentioning that the relevant notion is built on the assumption that the small amount of top-ranked feedback documents obtained from the initial round of retrieval are relevant which almost dominate the success of such query modeling and can be used to estimate a more precise query model for further retrieve more relevant documents, namely, the so-called pseudo-relevance feedback. Nevertheless, simply exploiting all of the top-ranked documents for query modeling (or reformulation), does not necessarily promise for a good performance, especially when the top-ranked documents contain much redundant or n on-relevant information.

1.1.2 Speech Recognition

In automatic speech recognition (ASR) system, the language modeling (LM) also plays a crucial role, which assists in constraining acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output. Due to its simplicity and predictive power, the n-gram model remains the predominant language model. A growing number of novel and ingenious LM techniques have been developed to complement or to replace the n-gram model. A more recent school of thought is to build a language model by leverage information cues extracted from pseudo-relevance feedback (PRF) to complement the n-gram model. For example, relevance modeling (RM) formulates the language model based on the notion of relevance, which can be approximated by PRF. RM that explores the relevance

information inherent between the search history and an upcoming word has exhibited preliminary promise for dynamic language model adaptation. Consequently, how to further explore useful cues from PRF for better estimating relevance modeling is an interesting research issue.

1.2 Contribution

In view of above mentioned problems, our research develops into two parts. First of all, with the above background, in this study we turn our attention to a more challenging problem of how to additionally glean useful cues from the top-ranked feedback documents to achieve more accurate query modeling. Towards this end, several kinds of information cues are considered and integrated to select representative and useful feedback documents for better query reformulation which leads to better retrieval performance.

Furthermore, we also investigate representing the query and documents with different granularities of index features to work in conjunction with the various information cues selection criteria for pseudo-relevance feedback. Finally, the utility of the methods deduced from our framework is verified by extensive comparisons with several existing active feedback methods for pseudo-relevance feedback.

On the other hand, in terms of speech recognition, this thesis follows this general line of research to build language models on top of the notion of relevance modeling

(RM) and has two significant contributions. First, the so-called "*bag-of-word*s" assumption of RM is relaxed by further incorporating loosely word order information and word proximity evidence into the RM formulation. Second, topic-based proximity information is additionally explored to further enhance the proximity-based RM language model.

1.3 Outline of the Thesis

The remainder of this paper is structured as follows. Chapter 2 briefly reviews the theoretical underpinnings of the LM approach not only for SDR but also for speech recognition to give readers a picture about how language modeling figures prominently in these two different fields of research. Plus, a concise introduction is given to existing related variations of pseudo-relevance feedback and proximity information modeling techniques, following by shedding light on the basic foundation of the RM modeling framework that can leverage lexical co-occurrence in a systematic way for language modeling in speech recognition. In Chapter 3, we describe and explain several cues we explore to select representative feedback documents during pseudo-relevance feedback. Subsequently, an elucidation of integrating proximity information cues into the formulation of RM for speech recognition. After that, the experimental settings and a series of retrieval experiments are presented in Chapter 4. Finally, Chapter 5 draws a conclusion from our study and suggests avenues for future

work.

2

Related Work

In this chapter, we provide a survey of the literature on the language modeling for spoken document retrieval and for speech recognition, respectively. We first present an overview of the retrieval modeling approaches for spoken document retrieval, and then we review major work up to date for pseudo-relevance feedback. Then, a brief introduction to language modeling for speech recognition is provided, following by the basic foundation of the relevance modeling for speech recognition.

2.1 Language Modeling for Spoken Document Retrieval

Language modeling (LM), providing proper quantitative scores to sequences of words or tokens by employing a statistical mechanism, has been an interesting yet challenging problem in speech and language processing community for a long time [20,21]. For instance, it can be applied to facilitate the acoustic analysis in speech recognition, lead the search through the vast space of candidate word strings, and quantify the acceptability of the final output from the speech recognizer. This statistical

paradigm was first proposed for solving IR problems by [7,8], demonstrating very good potential, and was latter also introduced to the field of SDR [4,9,10]. Language modeling (LM) approach is a recent trend in building SDR systems [4,10,18]. It can be attributed to both of the sound theoretical underpinnings and impressive empirical performance exhibited by the LM approach.

2.1.1 Retrieval Modeling Approaches

(I). Unigram Language Model

The basic 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)[11]. A spoken document is deemed to be relevant to a query if the corresponding document model is more likely to generate the query. 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 under the assumption that given the document D these query words are conditionally independent and no concern of word order (i.e., the so-called "*bag-of-words*" assumption). Thus, the similarity measure between a query and a document $P(Q|D)$ 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 $P(q_l|D)$ represents the probability of D generating q_l (a.k.a. the document model). The document model is constructed by two variants for each document D. in this study. One is to use the unigram language model (ULM). Toward this end, each document can, respectively, offer a unigram distribution for observing a query word, which is based on the empirical counts of words occurring in the document with the maximum likelihood (ML) estimator [11,20]. 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. However, how to strike the balance between these two probability distributions is actually a matter of judgment, or trial and error. The other is to employ a 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 a document D generates the corresponding topic. Nevertheless, both of PLSA and LDA merely provide coarse-grained latent semantic representations for the user information need, which is essentially unable to distinguish the fine-grained difference between the semantically-related words. In a given implementation, combining them with ULM to obtain better retrieval quality is always good [4,22]. For instance, as the equation (2) shown below, ULM can be linear combined with PLSA (referred to equation (3)),

where their retrieval results with optimal parameters tuning conducted on TDT-2 are demonstrated respectively on Table 1. The experiment results show the beneficial to further combine with PLSA.

$$P(Q|D)=\prod_{l=1}^{L}[\lambda\cdot P(q_l|D)+(1-\lambda)\cdot P_{PLSA}(q_l\mid D)], \tag{2}$$

$$P_{PLSA}(q_l\mid D)=\sum_{k=1}^{K}P(q_l\mid T_k)P(T_k\mid D), \tag{3}$$

Table 1 : Retrieval results (in mAP) achieved by ULM and PLSA.

Dev.	ULM	PLSA
TD	0.371	0.418
SD	0.323	0.345

(II). Kull-Leibler Divergence Measure

For SDR, another fundamental LM is the Kullback-Leibler (KL)-divergence measure [11,23,24]:

$$\begin{aligned} SIM_2(Q,D) &= -KL(Q\|D) \\ &= -\sum_{w\in V}P(w\mid Q)\log\frac{P(w\mid Q)}{P(w\mid D)} \\ &= \sum_{w\in V}P(w\mid Q)\log P(w\mid D)-\sum_{w\in V}P(w\mid Q)\log P(w\mid Q) \\ &\overset{\text{rank}}{=} \sum_{w\in V}P(w\mid Q)\log P(w\mid D), \end{aligned} \tag{4}$$

where the part $\sum_{w\in V}P(w\mid Q)\log P(w\mid Q)$ in the equation can be directly ignored since for a given query, the query entropy is identical for all the documents thus has no effect on ranking documents; $\overset{\text{rank}}{=}$ means equivalent ranking results. Compared with (1) where a query Q is regarded as a sample drawn from the language model of a possibly relevant document D, (4) entails designing the models of the query Q and the

document D, which are conventionally formed as a (unigram) language model (denoted by $P(w|Q)$ and $P(w|D)$), respectively, for observing any word w in the vocabulary V. In practice, the degree of relevance of the document D is measured by the value of $KL(Q\|D)$ (or probability distance), that is, the smaller the value of KL-divergence the more relevant this document is. The retrieval effectiveness of the KL-divergence measure depends largely on the accurate estimation of the query model $P(w|Q)$ and the document model $P(w|D)$. Moreover, it is easy to prove that the KL-divergence measure will generate the same ranking as the query-likelihood measure when the query model $P(w|Q)$ is simply estimated with the ML estimator [25]:

$$\begin{aligned} SIM_2(Q,D) &\overset{\text{rank}}{=} \sum_{w\in V} P(w|Q)\log P(w|D) \\ &= \frac{c(w,Q)}{|Q|}\log P(w|D) \\ &\overset{\text{rank}}{=} c(w,Q)\log P(w|D) \\ &= \log P(Q|D) \\ &\overset{\text{rank}}{=} P(Q|D). \end{aligned} \tag{5}$$

As (5), $P(w|Q)$ is simply derived as $\frac{c(w,Q)}{|Q|}$, where $c(w,Q)$ is the frequency of w occurring in Q and $|Q|$ is the total number of words a query Q has. Accordingly, the KL-divergence measure can be viewed as a generalization of the query-likelihood measure, which embraces additional merit of being able to accommodate extra information cues to better estimate its component models (especially, the query model) for obtaining document ranking in a systematic way.

Due to the fact that a query usually consists of only a few words, the query model

$P(w|Q)$ that is deem to reflect the user's information need might not be well-estimated merely by the ML estimator. In order to alleviate this problem, a conventional approach is to explore extra cues to strengthen the query model in the KL-divergence measure, that is, the so-called query expansion.

2.1.2 Pseudo-Relevance Feedback

As the problem mentioned above, a query often consists of only a few words, which is usually short for represent the user's information need. Therefore, estimating a query model by the ML estimator might not be appropriate. Furthermore, merely matching words between a query and documents might not be an effective approach, as the word overlaps alone could not show the semantic intent of the query. To cater for this, a conventional strategy is to adopt the idea of pseudo-relevance feedback which performs two rounds of retrieval so as to retrieve more relevant documents as Figure 1. In the first round of retrieval, a user given query is exploited for a SDR system to retrieve a small number of top-ranked feedback documents as pseudo-relevant documents. Subsequently, a refined query model is reformulated by leveraging these top-ranked feedback documents to add possible query terms and to reweight the query terms for improving query representation. After that, a second round of retrieval is ready with this new and better estimated query model to conduct with the KL-divergence measure depicted in (4) again. It is generally anticipated that the SDR

system can thus retrieve more relevant documents.

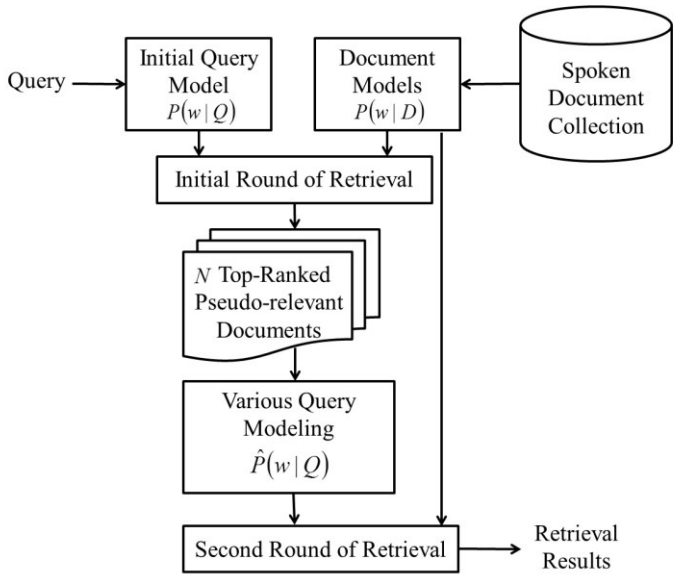

Figure 1: A schematic illustration of the Pseudo-Relevance Feedback process

Nevertheless, there are two basic challenges may necessarily encounter for LM-based SDR system implementing the pseudo-relevance feedback process. First, after initial round of retrieval, how to purify the obtained top-ranked feedback documents so as to preserve pure relevant information but remove redundant and non-relevant information, that is, to extract useful information from top-ranked documents for query expansion. This problem is quite important since the general state-of-the-art query modeling techniques exploit the whole feedback documents as a unit without thoroughly checking the information within. Second, if the top-ranked documents can guarantee some degree of relevance and usefulness for query modeling, how to effectively leverage these selected feedback documents or information for

estimating a more precise query model. As the latter, it might attribute to its effective in retrieval performance, researches focus on query modeling work with pseudo-relevance feedback are many, such as the simple mixture model (SMM) [26], the relevance model (RM) [19] and their extensions [18] , among others. Nevertheless, as far as we know, little work has been placed on the other side of coin, namely filtering helpful and representative feedback documents from the pseudo-relevance feedback for SDR query expansion. Researches in text information retrieval (IR) recently introduced several interesting algorithm to avoid redundant information and select more diversified feedback documents for query representation, such as "Gapped Top *K*" and "Cluster Centroid" selection methods [27] and many others. Based on different points of view, these related approaches are briefly reviewed as follows.

(I). Local Context Analysis

Local Context Analysis, a variation of pseudo relevance feedback, is proposed to avoid expansion term selection from non-relevant passages of the assumed-relevant documents. Therefore, the essence of local context analysis is to exploit top-ranked passages instead of top-ranked documents of the initial round of retrieval. It is addressed that local context analysis is at least as effective as pseudo-relevance feedback [28,29].

(II). Flexible Pseudo-Relevance Feedback

Pseudo Relevance Feedback (PRF) or Local Feedback (LF) is a well-known technique for improving average retrieval performance without human-judged relevant documents, however, a closer look often reveals that around one-third of search requests is actually ruined, which often results in worse retrieval performance than that of the initial retrieval by the original query [30,31]. Even though, its potential to improve average retrieval results completely automatically is known to be effective not only for monolingual retrieval [28] but also for cross-language retrieval [29], the user probably would not be glad if the automatic query expansion after a long time waiting finally spoils the user given query and of course retrieval performance.

In order to improve pseudo-relevance feedback, some researches make an attempt to vary the number of expansion terms [32]. Some researchers try to make it more reliable by proposing flexible local feedback (FLF) or flexible pseudo-relevance feedback (FPRF), which estimates not only the best number of expansion terms for each query but also the optimal number of top assumed-relevant documents [30] illustrated as Figure 2. In terms of FPRF, years before 2005, existing FPRF approaches determine optimal number of top assumed-relevant documents and query expansion terms under the assumption of the small amount of top-ranked documents are relevant. In 2005, approaches like Selective Sampling and Selective Sampling with Memory

Resetting [33] explore query word set within top-ranked documents to skip some top-ranked document for avoiding some redundant documents. However, since this study has not attempted to study optimal number of top assumed-relevant documents as well as optimal number of query expansion terms for each single query, our approach does not pertain to this category, however, this school of thought might be a good avenue for future work.

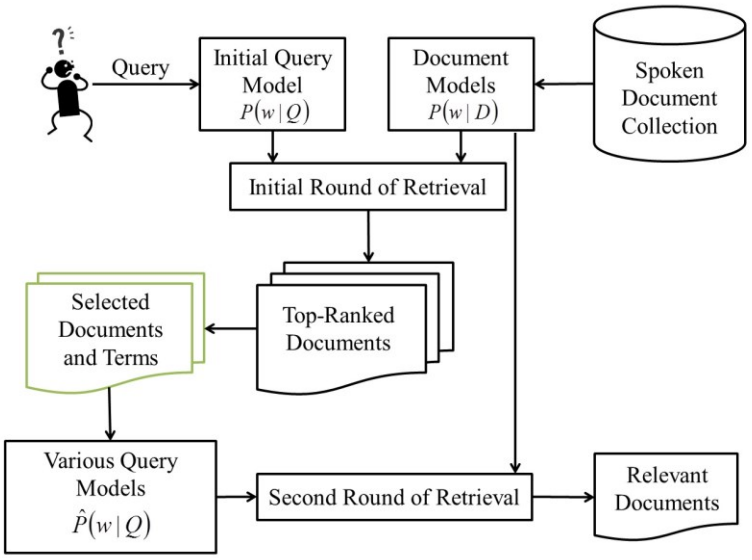

Figure 2 : A schematic illustration of the Flexible Pseudo-Relevance Feedback process

(III).Active Pseudo-Relevance Feedback

Shen and Zhai (2005) proposed the notion of active feedback, which tackles this problem from the view point of maximum learning benefit, that is, how to extract useful subset of documents from feedback documents so that the retrieval system can obtain maximum learning benefits. For active feedback, they develop two practical

algorithms to avoid redundant information among the top-ranked documents and to search for more diversity documents instead, including the Gapped Top K, and K Cluster Centroid algorithms [27]. The whole procedure of this category is illustrated in Figure 3, where the blue documents represent the selected documents as well as the primary goal of active pseudo-relevance feedback.

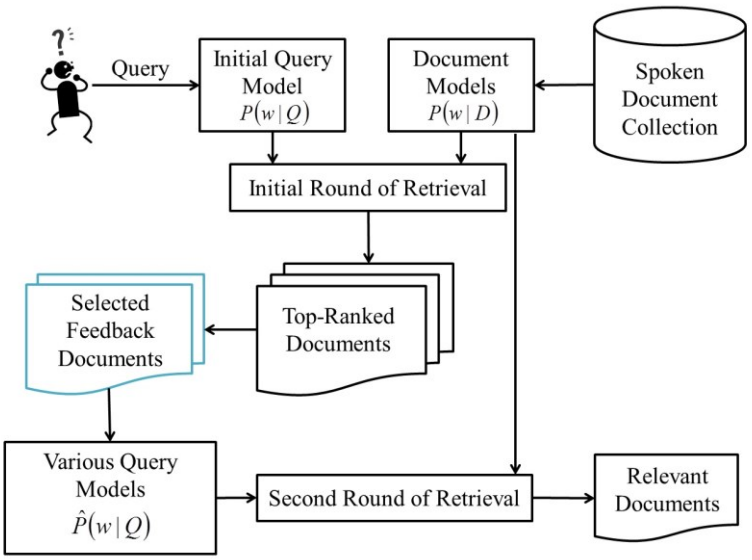

Figure 3 : A schematic illustration of the Active Pseudo-Relevance Feedback process

- Gapped Top K

Supposed there are K documents need to be selected from top N documents for relevance feedback and the mathematical equation can be shown as following:

$$N = (\mathrm{G}+1)K \tag{6}$$

where $G+1$ is a small positive integer. To consider both relevance and diversity of a candidate document, one possible way is to cluster top N documents into K

groups based on its corresponding relevance scores and choose the document with the highest relevance score from each group. By this way, the first group will be the top $G+1$ documents on ranking and the second group will be next $G+1$ documents, and so on. If we visualize this method like below, one can quickly figure out why authors named this method "Gapped Top K". Also, this method can perform the same as traditionally pseudo-relevance feedback, which simply using Top K documents of initial rank of first retrieval, when G is equal to 0 and K is set to N.

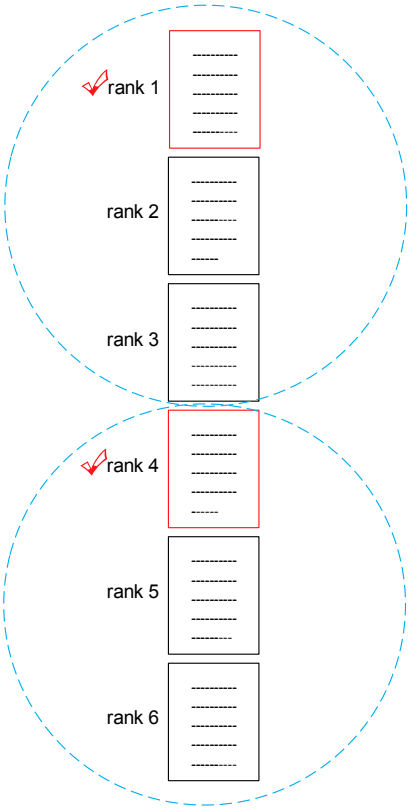

Figure 4: The Gapped Top K algorithm. For example, if selecting 2 documents from the top 6 documents of initial round of retrieval, that is, $N=6$, $K=2$, and $G=2$

Figure 4 demonstrates an example for the Gapped Top K algorithm, where the top-ranked 6 documents are ranked based on their relevance score and then clustered into two groups by the relevance score of each document. Then the document with a red check on its left hand side is the chosen documents which embrace the highest relevance score for each group. In other words, the algorithm picks one document and skips two documents (which are the "gap" between chosen ones)regularly until the pre-defined number of selected documents is met.

- Maximal Marginal Relevance

For Maximal Marginal Relevance (MMR), which is a greedy algorithm, documents are ranked based on relevance and non-redundancy cues [34,35]. Specifically, the documents are selected iteratively one by one which further excludes some already selected documents as well as optimize the MMR function below.

$$s(d|D) = \alpha \cdot r(d) + (1-\alpha) \cdot \max_{d' \in D} sim(d, d') \tag{7}$$

where $r(d)$ is a relevance scoring function, $sim(d, d')$ stands for a similarity function and α represents a weighting parameter for balancing relevance and non-redundancy. It is worth noting that when $\alpha = 1$ MMR can also reduce to conventional Top K method as a special case.

- Cluster Centroid

Apart from the aforementioned methods, a more straightforward method to model diversity is to directly divide the top N documents ranked by relevance score into K clusters and gather a representative document from each cluster to construct a better feedback document subset. The diversity of the chosen documents rests on clustering and selecting only one document of each cluster for representation. One can choose a representative document for each cluster by different measures. One of the intuitive approaches is to choose the document with the highest relevance score. Alternatively, choosing the centroid document maximizes the similarity between the documents in the same cluster for better representation.

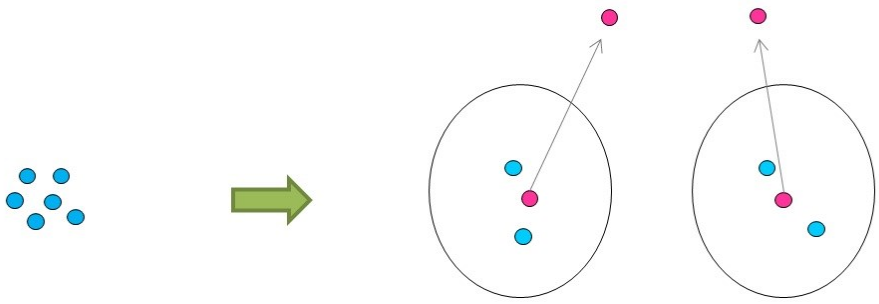

Figure 5 : A Cluster Centroid example

As an instance, each blue dot on the left side in Figure 5 indicates a top-ranked document. The dots in a circle on the right hand side of Figure 5 form a cluster, where

the red dot stands for the selected documents. Then selecting the feedback documents by Cluster Centroid is to extract a representative document for each cluster, which is the ultimate goal of Cluster Centroid. In this case, Figure 5 illustrates the selection process of Cluster Centroid when the number of the top-ranked documents N is 6 and the predefined number of selected documents K is 2.

- Active-RDD

Recently, another more attractive and interesting approach to select helpful documents is Active-RDD (indicating Active Learning to achieve Relevance, Diversity and Density), which is originally proposed for text IR and here introduced to SDR. This algorithm as its name means investigates relevance, diversity and density measure to select a better set of feedback documents for query expansion. Its objective function is expressed as below:

$$D^* = \underset{D \in \mathbf{D}_{\mathrm{Top}} - \mathbf{D}_{\mathrm{P}}}{\arg\max} \left[(1-\alpha-\beta) \cdot M_{Rel}(Q,D) + \alpha \cdot M_{Diversity}(D) + \beta \cdot M_{Density}(D) \right], \tag{8}$$

where D^* stand for a better set of feedback documents; $\mathbf{D}_{\mathrm{Top}}$ indicates the top-ranked documents from initial retrial; $\mathbf{D}_{\mathrm{P}}$ represents the already selected feedback documents; $M_{Rel}(Q,D)$, $M_{Diversity}(D)$ and $M_{Density}(D)$ denote the measure of relevance, diversity and density respectively; the weighting parameters are α and β. The detail implementation of these three measures can refer to chapter 3. Active-RDD the most

resembling method to ours in this study selects a document from the top-ranked document at a time, which simultaneously considers cues of relevance, diversity, and density of the document with respect to the already selected documents set.

2.1.3 Query Modeling

In order to investigate various feedback document selection methods as well as proximity information cues studied in this study, this section introduce query models, such as relevance model (RM), topic-based relevance model (TRM) and simple mixture model (SMM) for query reformulation.

(I). Relevance Model

In the KL-divergence measure, one of straightforward but effective techniques to enhance the query formulation is to leverage extra relevance cues related to the query by the so-called relevance model (RM) [19,36,37]. To realize this idea, each query Q is assumed to be associated with an unknown relevance class R_Q. Under the assumption of relevance class R_Q, relevant documents (which satisfy the stated information need of a query) of cause are also assumed to be samples drawn from R_Q. To strengthen the original query formulation with relevance information extracted from R_Q is anticipated to exhibits potential for better discriminating the content of relevant document from the content of non-relevant documents. Hence, retrieval issue

can be solved by discovering a strategy to formulate the relevance model (RM), that is, the probability model $P_{\mathrm{RM}}(w)$. The relevance model $P_{\mathrm{RM}}(w)$, which is the probability of observing a word w of a document related to the stated information need, as a multinomial observation of R_Q, can be interpreted as randomly selecting a document from the relevance class and observing a word from it.

Estimating the enhanced query model depends largely on the ideal relevance class; however, there is no prior knowledge about how to find the ideal relevance class in reality. A common strategy is to employ pseudo-relevance feedback process. As introduced above, the process typically performs two rounds of retrieval, first round of retrieval conducted by user given query, second round of retrieval rely on the newly formulated query based on the small amount top-ranked documents of initial retrieval. Hence, the relevance model can also adopt pseudo-relevance feedback and leverage top-ranked documents from initial retrieval as pseudo-relevant documents to approximate the ideal relevance class and further estimate the enhanced query model on top of these documents. It is worth mentioning that the initial retrieval in this study if not otherwise note is implemented with the KL-divergence measure, where the query model $P(w|Q)$ is estimated with the ML estimator $P(w|Q)$ (*cf.* (5)) to obtain a top-ranked list of M pseudo-relevant documents $\mathbf{D}_{\mathrm{Top}} = \{D_1, D_2, \ldots, D_M\}$ from the spoken document collection. An enhanced query model $P_{\mathrm{RM}}(w|Q)$ is then constructed

by these top-ranked documents. In addition to query modeling, $P_{\mathrm{RM}}(w|Q)$ can further combine with or replace the original query $P(w|Q)$ to form a better estimated (or enhanced) query model so as to identify more relevant documents. After query modeling, the final stage of pseudo-relevance feedback is to retrieval the final ranked list with new constructed relevance model.

To be more specific, based on the top-ranked list of M pseudo-relevant documents, the joint probability $P_{\mathrm{RM}}(Q,w)$ of Q and w being observed together in the relevance class R_Q of Q is formulated as follows:

$$P_{\mathrm{RM}}(Q,w)=\sum_{m=1}^{M}P(D_m)P(q_1,q_2,\ldots q_L,w\mid D_m), \tag{9}$$

where $P(D_m)$ is the probability of the document D_m to be randomly selected and $P(q_1,q_2,\ldots q_L,w\mid D_m)$ is the joint probability of co-occurrence of Q and w in D_m, where essentially assumes higher probability to the words co-occurred with query terms in D_m. To further assume that words are conditionally independent given D_m and word order is not importance (i.e., the "*bag-of-words*" assumption), the joint probability can then be decomposed as below:

$$P_{\mathrm{RM}}(Q,w)=\sum_{m=1}^{M}P(D_m)P(w\mid D_m)\prod_{l=1}^{L}P(q_l\mid D_m). \tag{10}$$

The probability of a pseudo-relevant document $P(D_m)$ can be simply set uniform or decided by the relevance degree of D_m to Q. $P(w\mid D_m)$ and $P(q_l\mid D_m)$ are determined by ML estimation, which is based on the word occurrence counts in D_m.

As the result, the enhanced query model $P_{\mathrm{RM}}(w|Q)$ can be expressed as

$$P_{\mathrm{RM}}(w|Q)=\frac{P_{\mathrm{RM}}(Q,w)}{P_{\mathrm{RM}}(Q)}=\frac{\sum_{m=1}^{M}P(D_m)P(w\mid D_m)\prod_{l=1}^{L}P(q_l\mid D_m)}{\sum_{m=1}^{M}P(D_m)\prod_{l=1}^{L}P(q_l\mid D_m)}. \tag{11}$$

Even though there has been explored different ways to derive relevance model $P_{\mathrm{RM}}(w|Q)$, the equation shown above (11) is validated to be more effectively and robustly than the other variants across different collections[37] .

(II). Topic-based Relevance Model

Apart from RM, in this study we also consider the performance evaluation of topic-based relevance model, which leverages latent topic information for the modeling of RM. To this end, a set of pre-defined latent topic variables $\{T_1,T_2,\ldots,T_K\}$ is assumed to describe the "*word-document*" co-occurrence characteristics among the pseudo-relevant documents obtained by the initial round of retrieval. Consequently, the word probability observed from a pseudo-relevant document D_m is no longer estimated directly by the frequency of the word occurring in a document, but instead based on likelihood of the document generates the topic and the probability of the word observed in the respective latent topics as well:

$$\widetilde{P}(w\mid D_m)=\sum_{k=1}^{K}P(w\mid T_k)P(T_k\mid D_m). \tag{12}$$

The joint probability of Q and w being simultaneously occurred in the relevance class R_Q of Q, as shown earlier in (10), is thus decomposed as

$$P_{\mathrm{TRM}}(Q,w)=\sum_{m=1}^{M}\sum_{k=1}^{K}P(D_m)P(T_k\mid D_m)P(w\mid T_k)\prod_{l=1}^{L}P(q_l\mid T_k). \tag{13}$$

This is topic-based relevance model (TRM), which employ a set of latent variable to reinterpret the probability a word is observed in a pseudo-relevant document. In contrast to RM, TRM assumes that the word distribution across a set of latent topics obtained from all spoken document in the collection may carry useful global topic information for relevance modeling.

In order to obtain the probabilities $P(w\mid T_k)$ and $P(T_k\mid D_m)$**,** we can employ PLSA or LDA so that the topical probability can be estimated by maximizing the total log-likelihood $\log L_{\mathbf{D}}$ of the spoken document collection $\mathbf{D}$, which can be further derived leveraging inference algorithms like the expectation-maximization (EM) algorithm [38] with uniform priors, or the variational approximation algorithm [39] with Dirichlet priors. To be more specific, we take EM algorithm for example. The objective function for driving $P(w\mid T_k)$ and $P(T_k\mid D_m)$ can be defined as below::

$$\log L_{\mathbf{D}}=\sum_{D\in\mathbf{D}}\sum_{w_i\in D}c(w_i,D)\log\widetilde{P}(w_i\mid D), \tag{14}$$

where $c(w_i,D)$ stands for the frequency count of w_i occurring in D. Then, the objective function (14) can be maximized by three iteratively updating equations as following:

$$P(w_i\mid T_k)=\frac{\sum_{D\in\mathbf{D}}c(w_i,D)P(T_k\mid w_i,D)}{\sum_{D\in\mathbf{D}}\sum_{w_j\in D}c(w_j,D)P(T_k\mid w_j,D)}, \tag{15}$$

$$P(T_k \mid D) = \frac{\sum_{w_i \in D} c(w_i, D) P(T_k \mid w_i, D)}{\sum_{w_j \in D} c(w_j, D)}, \tag{16}$$

$$P(T_k \mid w_i, D) = \frac{P(w_i \mid T_k) P(T_k \mid D)}{\sum_{l=1}^{K} P(w_i \mid T_l) P(T_l \mid D)}, \tag{17}$$

where $P(T_k \mid w_i, D)$ represents the probability of observing the latent topic T_k when a word w_i and a document D are given. To get a closer look, the probability $P(T_k \mid w_i, D)$ is estimated by $P(w_i \mid T_k)$ and $P(T_k \mid D)$ obtained from the previous training iteration.

(III).Simple Mixture Model

A school of thought to derive a feedback query model is to assumes words in the set of feedback documents $\mathbf{D}_{\mathrm{P}}$ are generated from two models: 1) the feedback model $P(w \mid FB)$ and 2) the background model $P(w \mid BG)$**,** namely simple mixture model (SMM) [26]. For feedback model $P(w \mid FB)$, it is estimated by the log-likelihood of a set of feedback documents $\mathbf{D}_{\mathrm{P}}$ expressed as follows, which can be maximized via the EM algorithm:

$$LL_{\mathbf{D}_{\mathrm{P}}} = \sum_{D_j \in \mathbf{D}_{\mathrm{P}}} \sum_{w \in V} c(w, D_j) \log\left[\lambda \cdot P(w \mid FB) + (1-\lambda) \cdot P(w \mid BG)\right], \tag{18}$$

where $c(w, D_j)$ is the number of times w occurring in a feedback document D_j and λ is the waiting parameter which can be used to estimate possible amount of background information (modeled as background model $P(w \mid BG)$) in feedback documents. In order to obtain feedback model, the objective function (18) can also be

maximized by the following EM algorithm via iterative maximization steps:

$$P^{(m)}(FB|w) = \frac{\lambda \cdot P^{(m)}(w|FB)}{\lambda \cdot P^{(m)}(w|FB) + (1-\lambda) \cdot P(w|BG)} \tag{19}$$

and

$$P^{(m+1)}(w|FB) = \frac{\sum_{D_j \in \mathbf{D}_P} c(w, D_j) \cdot P^{(m)}(FB|w)}{\sum_{w} \sum_{D_j \in \mathbf{D}_P} c(w, D_j) \cdot P^{(m)}(FB|w)}, \tag{20}$$

where m indicates the m-th iteration of the EM algorithm and $c(w, D_j)$ is the frequency count of w occurring in the feedback document D_j. After EM training, the feedback model $P(w \mid FB)$ can be used to support or replace the original query model. A schematic illustration of the SDR process is shown in Figure 1.

2.2 Language Modeling for Speech Recognition

In any large vocabulary continuous speech recognition (LVCSR) system, language modeling (LM) plays a critical and indispensable role [20]. It might attributable to the fact that LM has ability to assist the acoustic analysis, guide the search through a vast space filled with possible candidate word strings, and judge the quality or acceptability of the best output transcript for the speech recognizer.

2.2.1 N-gram Language Model

Due to inherent simplicity and predictive power, the *n*-gram language model [40,41] based on a statistical modeling paradigm is still the most commonly-used LM in speech recognition. The n-gram language models the regularity between the

immediately preceding *n*-1 words and a newly decoded word w_i.

$$P(W) = P(w_1)\prod_{i=2}^{m} P(w_i \mid w_{i-N+1},...,w_{i-1}) \tag{21}$$

where the $P(W)$ stand for the probability to generate a sequence of word W. However, the *n*-gram language model, good at modeling the local contextual cues or lexical regularity of a language, has inevitably faced the problems on two fronts. First, it is brittle across domains. The performance of n-gram language model is sensitive to topics of test data different from its training corpus. That is, its training corpus will affect or limit its performance directly. Second, it misses the information (either semantic or syntactic information) carried in the history beyond the immediately preceding *n*-1 words of a newly decoded word.

2.2.2 Topic-based Language Models

Consequent to the fact of that, a number of latent topic modeling approaches, which were originally formulated for information retrieval (IR) [5], have been introduced to dynamic language model adaptation and investigated to complement the *n*-gram models with varying degrees of success [11,42,43], such as latent Dirichlet allocation (LDA) [13]and its precursor, probabilistic latent semantic analysis (PLSA) [12]. Both of LDA and PLSA exploits a set of latent topic variables to portray the "*word-document*" co-occurrence relationship. Similar to topic model in information retrieval, the relationship between an upcoming word and its preceding search history

(regarded as a document in SDR) is reinterpreted by a set of predefined latent topics. That is, the search history predicts the subsequent decoded word is based on the likelihood that the search history generates the topics as well as the probability of the word observed in the respective latent topic. The main difference between LDA and PLSA is the inference of model parameters: The model parameters in PLSA are assumed to be fixed and unknown, whereas the model parameters in LDA are assumed to follow Dirichlet distributions.

2.2.3 Trigger-based Language Model

Apart from the topic models mentioned above, there are some other researchers have developed a number of complement approaches for the *n*-gram models, such as the trigger-based language model (TBLM) [44]. As the named for the language model, the concept of word trigger pairs are considered and formulated for language modeling. To shed light on TBLM, word trigger pairs can be automatically generated to describe the co-occurrence relationship between preceding history sequence $w_1,...,w_{i-1}$ and the upcoming word w_i as following:

$$P(w_i \mid w_1,...,w_{i-1}) = \frac{1}{i-1}\sum_{j=1}^{i-1} P(w_i \mid w_j) \tag{22}$$

The word trigger pairs are estimated by the prepared adaptation corpus, where the function to decide whether having a close or trigger-pair relationship between two words can be designed by mutual information inverse document frequency. As a

language model for speech recognition, the triggers often exist in the preceding history words. Thus, TBLM can capture the associations between the words in the search history and an upcoming word.

2.2.4 Recurrent Neural Network Language Model vs. Discriminative Language Model

In recent year, the recurrent neural network language model (RNNLM) [45] and the discriminative language model (DLM) [46] have received considerable interests from not only researchers but practitioners. The former attempts to map both of the preceding history and a upcoming decoded word into a continuous space and leverage a recursive fashion to derive the probability of a decoded word observed after the history sequence. In contrast to RNNLM, DLM tries to effectively discriminate correct decoded word from incorrect recognition hypotheses via a rich set of lexical and/or syntactic features as well as a wide variety of training algorithms for achieving better recognition results instead of solely relying on the distribution of training data.

2.2.5 Relevance Modeling

In addition to the above LM, a more recent school of thought is to leverage the notion of relevance to construct language models for speech recognition, namely relevance

modeling (RM) [47]. The notion of relevance modeling, which is originally developed in information retrieval, has recently attracted much attention and been successfully applied to many IR tasks. Nevertheless, as far as we're concerned, the investigation on exploring the effectiveness of relevance modeling for language modeling in speech recognition is still little [47].

In speech recognition, the role of language modeling can be simply interpreted as estimating the conditional probability $P(w|H)$, in which H is a search history, usually expressed as a sequence of words $H = h_1, h_2, \ldots, h_L$, and w indicates a possible decoded words (i.e., an upcoming word) [20,40,41]. In contrast to RM in SDR, each search history H (which can be interpreted as a query in SDR)has further assumed to be associated to a relevance class R_H, which can assist in predicting its immediately subsequent words w. As same as RM in SDR, the decoded word w is deem to be relevant to H if w is drawn from the same relevance class R_H of H and has higher probability to co-occurred with H. To this end, the joint probability of H and w being observed from R_H, i.e., $P_{\mathrm{RM}}(H,w)$, can thus be used to derive the conditional probability $P(w|H)$ for speech recognition [47].

Still, as RM in SDR, since there is no prior knowledge about the ideal relevance class R_H for each search history H, one possible strategy is to leverage a local feedback-like procedure, namely pseudo-relevance feedback, which takes H as a

query and can make an initial round of retrieval to obtain a top-ranked list of M pseudo-relevant documents from a contemporaneous (or in-domain) corpus to approximate R_H, denoted as $\mathbf{D}_H = \{D_1, D_2, \ldots, D_M\}$. Accordingly, the joint probability of simultaneously observing H and w can be defined as

$$P_{\mathrm{RM}}(H, w) = \sum_{m=1}^{M} P(D_m) P(H, w \mid D_m), \tag{23}$$

where $P(D_m)$ is the probability of D_m is randomly selected D_m from R_H and $P(H, w \mid D_m)$ (or $P(h_1, h_2, \ldots h_L, w \mid D_m)$) is the joint probability of observing H together with w in D_m. If the joint probability is further assumed that words are conditionally independent given D_m and word order is of no importance (i.e., the so-called "*bag-of-words*" assumption), equation (23) can then be decomposed as a product of unigram probabilities of words observed from D_m:

$$\begin{aligned} &P_{\mathrm{RM}}(H, w) \\ &\quad = \sum_{m=1}^{M} P(D_m) P(w \mid D_m) \prod_{l=1}^{L} P(h_l \mid D_m). \end{aligned} \tag{24}$$

The probability $P(D_m)$ can be simply set uniform or weighted referred to the relevance of D_m to H. Both of $P(w \mid D_m)$ and $P(h_l \mid D_m)$ are calculated based on the word occurrence frequencies in a pseudo-relevant document and integrated with the Bayesian or Jelinek-Mercer smoothing method [11]. As a result, the conditional probability $P(w \mid H)$ can be expressed as

$$\begin{aligned} P_{\mathrm{RM}}(w \mid H) &= \frac{P_{\mathrm{RM}}(H, w)}{P_{\mathrm{RM}}(H)} \\ &= \frac{\sum_{m=1}^{M} P(D_m) P(w \mid D_m) \prod_{l=1}^{L} P(h_l \mid D_m)}{\sum_{m=1}^{M} P(D_m) \prod_{l=1}^{L} P(h_l \mid D_m)}. \end{aligned} \tag{25}$$

If the probability of language model can be realized in the logarithmic domain, implementation of (25) can be quite efficient, [47]. Besides, RM can combine with the baseline n-gram language model to obtain a better recognition result, since the baseline *n*-gram language model trained on a large general corpus can offer the generic constraint cue of lexical regularities;

$$\widetilde{P}(w \mid H) = \lambda \cdot P_{\mathrm{RM}}(w \mid H) + (1-\lambda) \cdot P_{n\text{-gram}}(w \mid H), \tag{26}$$

where λ is the interpolation parameter, which balances the degree of reliance between RM model and *n*-gram language model.

3

Effective Pseudo-Relevance Feedback & Proximity Information

In order to effectively extract a smaller set of helpfully representative feedback documents from a small amount of top-ranked documents obtained from initial retrieval, ULM retrieval model was employed with initial query to acquire a number of top-ranked documents $\mathbf{D}_{\text{Top}} = \{D_1, D_2, \ldots, D_M\}$. To make a good selection on feedback documents, the document D in the top-ranked list $\mathbf{D}_{\text{Top}}$ is graded based on four different point of view of the document D, namely relevance, non-relevance, diversity and density cues, and then selected one document at a time. Specifically, in the selection process, each candidate feedback document D is scored according to a linear combination of measures of these cues as following:

$$D^* = \arg\max_{D \in \mathbf{D}_{\text{Top}} - \mathbf{D}_{\text{P}}} \left[(1-\alpha-\beta-\gamma) \cdot M_{Rel}(Q,D) + \alpha \cdot M_{NR}(Q,D) + \beta \cdot M_{Diversity}(D) + \gamma \cdot M_{Density}(D) \right], \tag{27}$$

where D^* denotes the final feedback document set; $\mathbf{D}_{\text{Top}}$ is the top-ranked document set obtained from initial retrieval; $\mathbf{D}_{\text{P}}$ indicates the set of already selected feedback

documents; $M_{Rel}(Q,D)$, $M_{NR}(Q,D)$, $M_{Diversity}(D)$ and $M_{Density}(D)$ stand for the measures of relevance, non-relevance, diversity and density to each candidate document D in $\mathbf{D}_{\mathrm{Top}}$; α, β and γ are the weighting parameters to balance the degree of importance or reliance among these four cues. The final used set of feedback document is then iteratively selected the highest score (26) document one by one from a small amount of the top-ranked documents until $\mathbf{D}_{\mathrm{P}}$ achieves the pre-defined number of feedback documents. It is worth mentioning that to some extent the iterative selection of the algorithm shown in (26) resembles maximal marginal relevance (MMR) ranking algorithm [34,48] which was originally developed for extractive document summarization. To get a more clear view on the implementation of these four cues, the detail measure techniques will be introduced below. First of all, $M_{Rel}(Q,D)$ denotes relevance measure between query and document, which can be realized by ULM retrieval model depicted in (1), the initial retrieval we used here.

3.1 Diversity Measure

In recent years, diversification no matter in retrieval results or pseudo relevance feedback documents has attracted much attention in the text IR community, since the conventional document ranking criteria often considers merely relevance information in a document to a given query and will inevitably suffer from too many redundant documents shown in the top-ranked list, which may be quite annoying and even not

effective for query modeling. In terms of pseudo-relevance feedback, if top-ranked documents contain too much redundant information and are used to build the query model, the second round of retrieval, which is anticipated to return more relevant documents, may result in returning too many "redundant" documents to the user. Hence, it is good to consider diversity cues within feedback documents, especially those used to construct a query model. For better query reformulation, the diversity measure of a candidate feedback document with respect to the already selected feedback documents $\mathbf{D}_{\mathrm{P}}$ is defined as following:

$$M_{Diversity}(D) = \min_{D_j \in \mathbf{D}_{\mathrm{P}}} \frac{1}{2} \cdot [KL(D_j \parallel D) + KL(D \parallel D_j)]. \tag{28}$$

As shown above, each candidate document attempts to compute the model distance between the candidate and the already selected one by one and records the smallest grades as its diversity score, which can gain maximum diversity effect when the objective function (26) prefer a candidate document with higher score.

3.2 Density Measure

Another interest but effective approach is to take into account the structural information or distribution information among the top-ranked documents [49]. To realize this idea, the average symmetric probability distance between a candidate document D to all the other documents D_M in $\mathbf{D}_{\mathrm{Top}}$ is computed as following:

$$M_{Density}(D) = \frac{-1}{|\mathbf{D}_{\text{Top}}| - 1} \cdot \sum_{\substack{D_h \in \mathbf{D}_{\text{Top}} \\ D_h \neq D}} [KL(D_h \parallel D) + KL(D \parallel D_h)], \tag{29}$$

where $|\mathbf{D}_{\text{Top}}|$ indicates the total number of documents in $\mathbf{D}_{\text{Top}}$. Based on this similarity (or distance) measure, a document D is deemed to be more similar (or closer) to all the other documents, which can be more preventative in this group $\mathbf{D}_{\text{Top}}$. Thus, density measure $M_{Density}(D)$ can be used to measure the representative of a document. In this case, a higher value of $M_{Density}(D)$ means more representative the document D is in $\mathbf{D}_{\text{Top}}$. That is, this information can measure the representative of a candidate document and generality of a candidate document among the others as well. More specifically, this measure can be visualized as following:

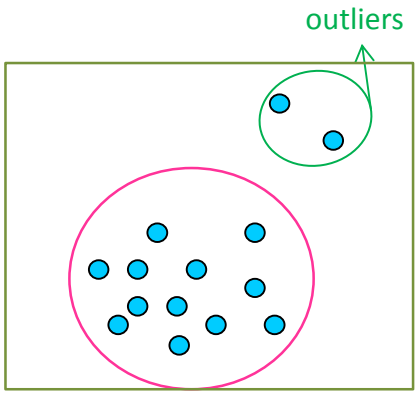

Figure 6 : A diagram of density measure

In Figure 6, each blue point stands for a candidate document of the top-ranked documents; and the similarity distance or relationship between arbitrary two blue points (namely, two assumed-relevant documents) is the density measure equation

attempts to capture. That is, if the average distance of an assumed-relevant document to all the others is small, that means this assumed-relevant document is close to all other top-ranked documents, which essentially illustrates this document is located in a larger similar group like the blue points circled by pink circle in Figure 7, and form a dominant group in these pseudo-relevance feedback documents. A candidate document like this may be most representative document in the candidate set. Conversely, if the density measure score of a top-ranked document is large, that means the document is far away from most of the top-ranked documents, which is probably as an outlier in the green circle of Figure 6. Therefore, if these assumed-relevant documents are actually relevant to some degree, this measure can be utilized to keep away from the information or documents not in popular demand, which might have higher probability to satisfy the user information need. The above three measures are the primary features Active-RDD is concerned.

3.3 Non-Relevance Measure

To measure non-relevance, there are some investigations have been placed on modeling non-relevance information and exploiting to support information retrieval [18,50]. Ideally, non-relevance modeling can be estimated according to each given query Q respectively and uniquely. To this end, a straightforward approach is to exploit low-ranked documents acquired from the initial round of retrieval to construct

the non-relevance model $P(w \mid NR_Q)$ specially for each single query. Then, the non-relevance measure of a candidate feedback document D can be expressed as

$$M_{NR}(D) = KL(NR_Q \| D). \tag{30}$$

where NR_Q denotes non-relevance model with respect to a query Q.

It is worth noting that to further incorporate $M_{NR}(D)$ for feedback document selection will constrain the selection process to keep selected feedback documents not only have a probability distance close to the original query model but with a probability distance far away from the non-relevance model, which is quite useful when an uncertain document is consider with only knowledge of its relevance information measure but have no idea to its potential non-relevance information within.

However, non-relevance information is not easy to correctly estimate even it has strong relationship with the query. It might imply that when a query in this moment can not be correctly estimated so as the non-relevance model. Therefore, a more easy to obtain and estimate model, which is available to approximate non-relevance information, is needed. Thus, owing to the fact that in reality the number of relevant documents in response to a given query is usually very small when compared to that of non-relevant ones from the view of the whole retrieval corpus, non-relevance model may be assumed and approximated by the entire spoken document collection. More

specifically, the background language model $P(w|BG)$ can be an alternative estimate for the non-relevance model. Another advantage to use background language model as an alternative is that the clarity measure of background model is shown to have ability to tell vague information of a query in text IR[51], which in turn can assist feedback document selection in identifying vague expression of a candidate document. As a result, in addition to three measures introduced above, namely measures of relevance, diversity and density, background language model $P(w|BG)$ is further explored to support feedback document selection.

3.4 Proximity Information for RM

On the other hand, model-based pseudo relevance feedback can also be an effective method for speech recognition. Relevance model (RM) based on "*bag-of-words*" assumption, which can facilitate the derivation and estimation, may be oversimplified for the task of language modeling in speech recognition, which pays more attention on the regularity of a language, such as the success of n-gram language model. Therefore, in order to better adapt RM (*cf.* Section 2.2.5) for speech recognition, one possible approach is to take advantage of word order and adjacency relationships among history words and the upcoming word from pseudo relevance feedback documents. To this end, the joint probability of observing $H = h_1, h_2, \ldots, h_L$ together with w in a pseudo-relevant

document D_m can be alternatively decomposed as follows, which simultaneously models the pairwise word order and (immediate or intermediate) adjacency relationships as well:

$$\begin{aligned}\widetilde{P}_{\mathrm{RM}}(H,w\,|\,D_m) \\ = P(h_1\,|\,D_m)\Big[\prod_{l=2}^{L}P(h_l\,|\,h_{l-1},D_m)\Big]P(w\,|\,h_L,D_m),\end{aligned} \tag{31}$$

where $P(h_l\,|\,h_{l-1},D_m)$ and $P(w\,|\,h_L,D_m)$ resemble traditional bigram language model in capturing the pairwise proximity (namely, word order and adjacency) relationships between history words and the decoded word in a pseudo-relevant document. To be more specifically, the conditional probability $P(w\,|\,h_L,D_m)$ (which is the building block of RM) is further realized by the following formulation:

$$P(w\,|\,h_L,D_m)=\frac{C_\tau(h_{L,},w,D_m)}{\sum_{w'}C_\tau(h_{L,},w',D_m)}. \tag{32}$$

where $C_\tau(h_L,w,D_m)$ stand for the frequency co-occurrence count of h_L and w observed within a fixed-length sliding window in a pseudo-relevant document D_m, where the sliding window places immediately after each position of h_L and has a window size of τ words. It is worth noting that when the window size τ is set to one, then the conditional probability $P(w\,|\,h_L,D_m)$ in (31) is actually equivalent to a conventional bigram language model estimated by a document D_m. Therefore, it is quite interesting to modulate the window size and explore the impact of word proximity with different degree of closeness on relevance modeling. The resulting language model is hereafter named as the proximity-based RM model (PRM).

3.5 Topic-based Proximity Information for RM

As an inspiration similar to PLSA and LDA, latent topic information is explored to work with RM modeling [47]. To this end, a set of latent topic variables $\{T_1, T_2, \ldots, T_K\}$ is employed to portray the "*word-document*" co-occurrence relationship from the pseudo-relevant documents of a search history $H = h_1, h_2, \ldots, h_L$. Accordingly, the conditional probability that the search history H together with a decoded word w are observed from a pseudo-relevant document D_m is not computed directly based on the number of times of H and w co- occurring in the document D_m, but instead based on the frequency count of H and w co-occurring in the latent topics as well as the likelihood that D_m generates the respective topics:

$$\begin{aligned} &\widehat{P}_{\mathrm{RM}}(H, w \mid D_m) \\ &\quad = \sum_{k=1}^{K} \left[\prod_{l=1}^{L} P(h_l \mid T_k) \right] P(w \mid T_k) P(T_k \mid D_m). \end{aligned} \tag{33}$$

where the component probabilities can be estimated exploiting the expectation-maximization (EM) inference algorithm [38]. Substituting (33) into (25), to some extent, offers a mechanism to depict the proximity information between the search history H and the upcoming word w in the latent topic space related to a pseudo-relevant document. The relevance model (33) is referred to as the topic-based relevance model (TRM).

4

Experiments on Spoken Document Retrieval

4.1 Spoken Document Collections & Evaluation Metrics

For this study, we used the Topic Detection and Tracking collections (TDT-2) [52]. Spoken documents are gathered from the Mandarin news stories by Voice of America news broadcasts. The test queries were collected via compiling the title fields of the Chinese text news stories from Xinhua News Agency. Therefore, the task of news monitoring and tracking is especially suitable on this corpus. Performance evaluation of all news stories were judged based on the exhaustively labeling with event-based topic. Table 1 demonstrates some basic statistics about the TDT-2 collections used in this study. The TDT-2 collection is leveraged to tune the optimal parameters and to see the best performance for various retrieval models. In addition, the number of latent topics used to build TRM, PLSA and LDA is set to 32. The number of pseudo-relevant documents acquired from the initial round of retrieval for the various query models is set to 25. It is worth mentioning that all the parameters used in this study can be further fine-tuned to achieve better performance for different spoken document collections via

appropriate experimentation.

Table 2 : Statistics for the TDT-2 Collections

	TDT-2 1998, 02~06			
# Spoken documents	2,265 stories, 46.03 hours of audio			
# Distinct test queries	16 Xinhua text stories (Topics 20001~20096)			
	Min.	Max.	Med.	Mean
Doc. Length (in characters)	23	4,841	153	287.1
Length of test query (in characters)	8	27	13	14
# Relevant doc. per test query	2	95	13	29.3

The Chinese word transcripts of the Mandarin audio collections (TDT-2) were recognized by Dragon large-vocabulary continuous speech recognizer. To evaluate the performance of Dragon's recognizer, a fraction of the TDT-2 (approximately 39.90 hours) is spot-checked. The error rates of word, character and syllable are 35.38%, 17.69% and 13.00%, respectively. Due to the fact that it is not available to obtain Dragon's lexicon, the LDC Mandarin Chinese Lexicon is augmented with 24k words extracted from Dragon's word recognition output, and for computing error rates, the manual transcripts are tokenized by the augmented LDC lexicon (about 51,000 words). The query sets are also tokenized by this augmented LDC lexicon in the retrieval experiments.

In terms of non-interpolated mean average precision (mAP), the retrieval results are defined following the TREC evaluation [5,6]:

$$\mathrm{mAP} = \frac{1}{E}\sum_{i=1}^{E}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}} \tag{34}$$

where E indicates the number of test queries, N_i means the total number of relevant documents pertaining to query Q_i, and $r_{i,j}$ denotes the position (rank) of the j–th relevant document pertaining to query Q_i, counting down from the top of the ranked list.

4.2 Subword-level Index Features

In Mandarin Chinese, although only some (e.g., 80 thousands, depending on the domain) are commonly used, there is an unknown number of words. Each word includes one or more characters, each of which is pronounced as a monosyllable and is a morpheme with its own meaning. Moreover, full textual coverage of written Chinese is almost covered by an inventory of about 6,000 characters. There is a many-to-many mapping between characters and syllables.

The characteristics of the Chinese language result in some special considerations when performing Mandarin Chinese speech recognition. Mandarin Chinese speech recognition evaluation is usually based on syllable and character accuracy, rather than word accuracy. Due to the exclusive characteristics of the Chinese language, SDR has some special considerations. Fist, word-level index features embrace more semantic

information than subword-level index features; consequently to the fact of that, word-based retrieval enhances precision. Second, subword-level index features are more robust in contrast to the Chinese word tokenization ambiguity, homophone ambiguity, open vocabulary problem, and speech recognition errors; as a result, subword-based retrieval enhances recall. Accordingly, it is good to combine the information acquired from indexing the features of different levels [9].

In this study, different levels of index features are utilized for construct the query and document models involved in the KL-divergence measure, including words, syllable-level units, and their combination. To this end, in addition to words, syllable pairs are taken as the basic units for indexing. Both the recognition transcript and the manual transcript of each spoken document, which were originally tokenized as words, were automatically transferred to overlapping syllable pairs. Then, all the distinct syllable pairs occurring in the spoken document collection were identified and collected to construct a vocabulary or lexicon of syllable pairs for indexing. Syllable pairs can be used to replace words, to represent the query and spoken documents, and thereby to construct the associated query and document models.

4.3 Baseline Experiments

In the first set of experiments, 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 leveraged for constructing the refined query models. The corresponding results are shown in Table 3, where the results of ULM and LDA (latent Dirichlet allocation) [12] are also listed for reference. LDA is a state-of-the-art (more sophisticated) LM-based retrieval model that employs a set of latent topics for representing (spoken) documents. It is worth mentioning that both ULM and LDA perform retrieval only with the initial query. Take a look on Table 3 reveals two noteworthy points. First, in terms of mAP, the performance gap between the retrieval using manual transcripts (denoted by TD) and the recognition transcripts (denoted by SD) is about 0.05, such degradation is apparently less significant as compared to the WER of spoken documents [18]. Second, RM and SMM tend to perform on a par with each other, and they gain substantial improvements over ULM (and are comparable to LDA). TRM demonstrates superior performance over RM and SMM, confirming the merits of leveraging topical information for query modeling.

Table 3 : Retrieval results (in mAP) achieved by various retrieval models.

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

4.4 Using Effective Pseudo-Relevance Feedback

In the second set of experiments, the performances of the various feedback document selection methods investigated in this study are evaluated, including “Gapped Top K” (denoted by “Gapped” for short), “Cluster Centroid” (denoted by “Cluster” for short), “Active-RDD” and our purposed method (cf. Chapter 3), incorporating with some of the above retrieval (query) models (the target 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 5 for comparison. There are three noteworthy points to these results. First, no matter using “Active-RDD” or our proposed method to further select feedback documents seems to outperform the retrieval models simply using the top 5 feedback documents (cf. Table 5) or the top 25 documents (cf. Table 3 and 5) obtained from the initial round of retrieval as the feedback documents by a big margin, implying that appropriate feedback document selection is important to the success of query reformulation. Second, our proposed method outperformances “Active-RDD” for all cases, which comfirms the advantage of using the non-relevance cue for feedback document selection. Third, the performance result of both of “Gapped Top K” and “Cluster Centroid” appears to be much inferior to that of “Active-RDD” and our proposed

method.

Table 4 : 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 5 : 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

4.5 IDF-Based Term Weighting

In the third set of experiments, how to emphasize the words occurring in the feedback documents that have higher descriptive capabilities in the estimation of the refined query models is explored in this section. 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, demonstrating how predictive a word is, usually is denoted as a function of the inverse logarithm of the number of documents that contain the word [5]. It is apparent from Table 4 that leveraging such an IDF-based weighting scheme (denoted by TW for short) can further boost the retrieval performance, in combination with the various retrieval modeling techniques (*cf.* Rows "Our Method" vs. "Our Method + TW" in Table 4). More, in contrast to the baseline results of ULM and LDA shown in Table 3, it shows that more elaborate query modeling is very importance to an LM-based SDR system.

4.6 Fusion of Different Levels of Indexing Features

In the final set of experiments, how the word- and syllable-level index features complement each other to represent both the test queries and spoken documents are explored. The results of the SD with different query modeling techniques (i.e., RM, TRM and SMM) are shown in Table 5. As the experiments mentioned above, query modeling techniques are further combined with our feedback document selection

method and IDF-based term weighting method (denoted by “Our Method + TW” in Table 5). From Table 5, in general, the results for the various query modeling techniques, embrace consistent trends with that of the previous experiments. Specifically, there are two worth noting points from these results. First, the subword-level (syllable-level) index features exhibits competitive or even better performance than the word-level index features when the retrieval system conducts on top of imperfect recognition transcripts (i.e., for the SD case). Second, not surprisingly, referred 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 to achieve better performance. It implies that fusion of different granularities of index features incorporating with the presented feedback document selection method and IDF-based term weighting method has good performance for SDR, .

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

Experiments on Speech Recognition

5.1 Speech Recognition Corpus & Evaluation Metrics

The speech corpus consists of about 196 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News) [53]. A subset of 25-hour speech data compiled during November 2001 to December 2002 was used to bootstrap the acoustic training with the minimum phone error rate (MPE) criterion and the training data selection scheme. Another subset of 3-hour speech data collected within 2003 is preserved for the development set (1.5 hours) and the test set (1.5 hours). The statistical information on the exactly number of sentences and hours of the development set and the test set is shown in Table 6.

Table 6 : Statistics for the Speech Corpus

	# sentences	# hours
adaptation text corpus	3643	20
development corpus	292	1.5
test corpus	307	1.5

The vocabulary size is about 72 thousand words. With the SRI Language

Modeling Toolkit (SRILM) [54] the trigram language model used in this study was estimated from a background text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The adaptation (contemporaneous) text corpus used for training the proposed various RM models and the other adaptation methods was collected from MATBN 2001, 2002 and 2003 (excluding the test set), which consists of one million Chinese characters (3,643 documents) of the orthographic broadcast news transcripts (*cf*. Table 6).

In tradition, the performance evaluation metric used in Mandarin speech recognition is usually the character error rate (CER) rather than the word error rate (WER), which is defined as the sum of the insertion (Ins), deletion (Del), and substitution (Sub) errors between the recognized and reference Chinese character strings, divided by the total number of Chinese characters in the reference string (ref):

$$\text{CER} = \frac{Ins + Sub + Del}{ref} \tag{35}$$

In this study, all the language model adaptation experiments were performed in word graph rescoring. The word graphs of the speech data were built beforehand with a typical large vocabulary continuous recognition (LVCSR) system [55,56]. The baseline rescoring procedure with the background trigram language model results in a character error rate (CER) of 20.22% on the test set. Notice that the constants or

weighting (interpolation) coefficients of all the language models investigated in this study were all tuned at optimum values by the development corpus and tested in the test corpus with the best parameters obtained from the development corpus.

5.2 Baseline Experiments

In this section, the performance of RM is compared with several well-practiced language models, including PLSA, LDA, TBLM, RNNLM and DLM; their corresponding CER results are shown in Table 7. It should be noted that in this study, RNNLM was implemented with the toolkit released by [57]. In addition, DLM leverages features composed of word unigram, bigram and trigram counts was trained with different algorithms, including minimum error rate training (denoted by "MERT" for short), global conditional log-linear model (denoted by "GCLM" for short) and weighted global conditional log-linear model (denoted by "WGCLM" for short). Interested readers may refer to [46,58] for an idea and updated introduction to various training algorithms designed and developed for DLM. Inspection of Table 7 reveals several noteworthy points. First, TBLM obtains an almost negligible improvement as compared to the baseline trigram model. Second, RNNLM achieves the best performance, which leads to a relative CER improvement of 5.5% over the baseline trigram model. Third, RM, PLSA and LDA seem to perform slightly worse than RNNLM but apparently better than the three variants of DLM. Fourth, although RM is

not the best performing one among these language models, the relevance information it attempts to explore is orthogonal (complementary) to those discovered by the other language models. It is worth mentioning that the CER(%) of various language models shown in Table 7 are estimated based on the development set.

Table 7 : The speech recognition results (in CER (%)) of various language models compared in this study.

RM	PLSA	LDA	TBLM	RNNLM	DLM (MERT)	DLM (GCLM)	DLM (WGCLM)
19.21	19.28	19.22	20.09	19.10	19.74	19.89	19.62

5.3 Using Proximity Information

In the second set of experiments for speech recognition, the utility of additionally incorporating the proximity information (i.e., the word order and adjacency cues) is evaluated in order to better describe the word-word co-occurrence relationships in a pseudo-relevant document for relevance modeling; the resulting model is designated as PRM (*cf.* Section 3.4). Table 8 shows the corresponding CER results with respect to different window lengths being used to capture the word order and adjacency cues. Consulting Table 8 we notice two particularities. One is that the performance of PRM is improved when the length of the window becomes larger; the improvements, however, seem to soon reach a plateau when the length of the sliding window is set to 2 words. The other is that PRM performs better than RNNLM and RM, and yields a

relative CER improvement of 6.6% over the baseline trigram model (when τ is optimal set to 2 for PRM). Significance tests based on the standard NIST MAPSSWE [59] indicate the statistical significance of such a CER reduction (note here that the statistical significance was determined at the 95% confidence level).

Table 8 : The speech recognition results (in CER (%)) of PRM.

PRM (τ=1)	PRM (τ=2)	PRM (τ=3)	PRM (τ=4)	PRM (τ=5)
18.91	18.89	18.97	18.98	19.07

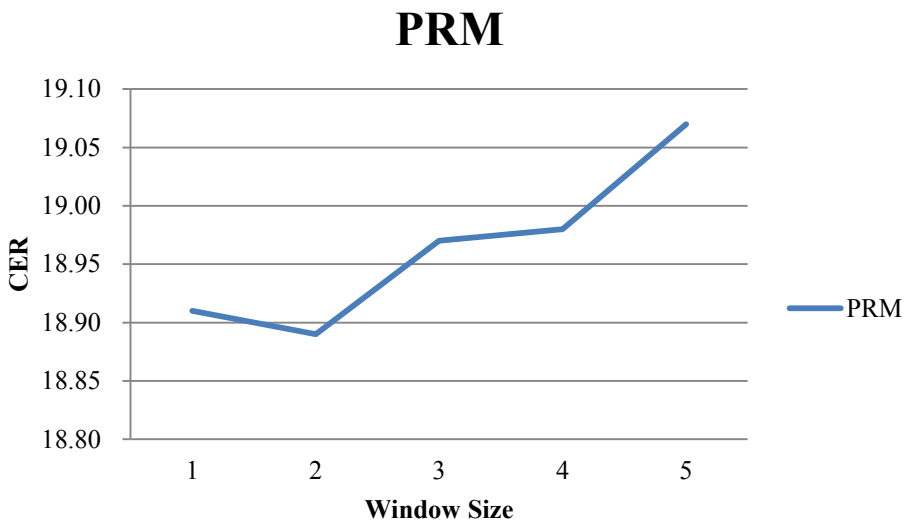

Figure 7 : The speech recognition results (in CER (%)) of PRM

5.4. Using Latent Topic Proximity Information

In the third set of experiments in speech recognition, the possibility of leveraging latent topic information is investigated, namely topic-based relevance model (TRM) which provides a mechanism to describe the proximity of the search history H and the upcoming word w in the latent topic space within a pseudo-relevant document,

for the RM modeling. Additionally, we can further combine PRM with TRM (through a simple linear interpolation) so as to maximize on two different sources of proximity information simultaneously for RM modeling. As can be seen from Table 9, the improvement brought by TRM is less pronounced as compared to that of PRM, which to some extent confirms the intuition that proper modeling of word order and adjacency information is quite useful to the success of speech recognition. The combination of PRM and TRM, however, offers a moderate improvement over PRM in isolation. Additionally, if PRM is further linearly combined with PLSA, the CER can be ultimately reduced as well. Finally, Table 10 exhibits the results of significance tests between RM and PRM, RM and PRM+TRM, which demonstrates the improvement of our approach is statistically significant.

Table 9 : The speech recognition results (in CER (%)) of TRM and PLSA, and their combination with PRM respectively.

TRM	PRM (τ =2) + TRM	PRM (τ =2) + PLSA
19.18	18.41	18.71

Table 10 : The p-value obtained from the pair t-test on CER(%)of PRM with respect to that of RM and CER(%) of PRM + TRM with respect to that of RM respectively.

	p-value (PRM (τ =2))	p-value (PRM (τ =2) + TRM)
RM	4.99E-02	4.77E-05

6

Conclusion and Future Work

In this study, to further enhance query formulation especially for SDR, a language modeling (LM) framework is proposed to combine several kinds of information cues, namely, relevance, diversity, density and non-relevance into the process of feedback document selection. The utility of the retrieval methods also been validated by extensively comparisons with several existing methods. The experimental results seem to show the superiority of our LM framework for SDR. As to future work for SDR, we would like to adopt this LM framework for speech recognition and summarization [47,60]

On the other hand, a novel extension of the RM framework for language modeling in speech recognition has been presented as well. Our contribution to speech recognition is two-fold. First, the so-called "*bag-of-word*s" assumption of RM is relaxed by incorporating word proximity evidence into the RM formulation. Second, topic-based proximity information is additionally explored in an effort to enhance the proximity-based RM framework. Experimental results reveals that the various

language models deduced from our framework are very comparable to existing language models for LVCSR. In this aspect, we would like to adopt this LM framework for speech retrieval and summarization applications for future work[60,61].

Bibliography

[1] L. Lin-shan and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine,* vol. 22(5), pp. 42-60, 2005.

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

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

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

[5] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval: The Concepts and Technology behind Search": Addison-Wesley Professional, 2011.

[6] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceeding 8th Text REtrieval Conference (TREC-8)*, 2000, pp. 107-129.

[7] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, 1998, pp. 275-281.

[8] D. R. H. Miller, T. Leek, and R. M. Schwartz, "A hidden Markov model information retrieval system," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, 1999, pp. 214-221.

[9] B. Chen, H.-M. Wang, and L.-S. Lee, "A discriminative HMM/N-gram-based retrieval approach for mandarin spoken documents," vol. 3(2), pp. 128-145, 2004.

[10] 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(1), pp. 1-30, 2010.

[11] C. X. Zhai, "Statistical language models for information retrieval: A critical review", *Foundations and Trends in Informational Retrieval*, vol. 2,no. 3, pp. 137-213, 2008.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3(1), pp. 993-1022, 2003.

[13] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* vol. 42(1), pp. 177-196, 2001.

[14] D. Blei and J. Lafferty, "Topic models," in *Text Mining: Theory and Applications*, A. Srivastava and M. Sahami, Eds., ed New York: Taylor and Francis, 2009.

[15] X. Yi and J. Allan, "A Comparative Study of Utilizing Topic Models for Information Retrieval," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, Toulouse, France, 2009, pp. 29-41.

[16] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, 2007, pp. 631-638.

[17] S. Parlak and M. Saraclar, "Performance Analysis and Improvement of Turkish Broadcast News Retrieval," *IEEE Transactions on Audio, Speech, and Language Processing, ,* vol. 20(3), pp. 731-741, 2012.

[18] 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(9), pp. 2602-2612, 2012.

[19] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, 2001, pp. 120-127.

[20] F. Jelinek, "Statistical methods for speech recognition", Cambridge, MA: MIT Press, 1999.

[21] C. D. Manning and H. Schutze, "Foundations of statistical natural language processing", Cambridge, MA: MIT Press, 1999.

[22] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA, 2006, pp. 178-185.

[23] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics,* vol. 22(1), pp. 79-86, 1951.

[24] L. Shih-Hsiang, Y. Yao-Ming, and C. Berlin, "Leveraging Kullback-Leibler Divergence Measures and Information-Rich Cues for Speech Summarization " *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19(4), pp. 871-882, 2011.

[25] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to Ad Hoc information retrieval," in *Proceedings of the 24th annual*

international ACM SIGIR conference on Research and development in information retrieval, ed. New Orleans, Louisiana, USA: ACM, 2001, pp. 334-342.

[26] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management*, Atlanta, Georgia, USA, 2001, pp. 403-410.

[27] X. Shen and C. Zhai, "Active feedback in ad hoc information retrieval," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005, pp. 59-66.

[28] J. Xu and W. B. Croft, "Query Expansion Using Local and Global Document Analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996, pp. 4-11.

[29] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, Philadelphia, Pennsylvania, USA, 1997, pp. 84-91.

[30] T. Sakai, M. Kajiura, and K. Sumita, "A First Step towards Flexible Local Feedback for Ad hoc Retrieval," in *Proceedings of the fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 95-102.

[31] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Transactions on Information Systems,* vol. 18(1), pp. 79-112, 2000.

[32] S. E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," in *The 8th Text REtrieval Conference (TREC 8)*, 2000, p. 151.

[33] T. Sakai, T. Manabe, and M. Koyama, "Flexible pseudo-relevance feedback via selective sampling," *ACM Transactions on Asian Language Information Processing,* vol. 4(2), pp. 111-135, 2005.

[34] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, 1998, pp. 335-336.

[35] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, 2003, pp. 10-17.

[36] V. Lavrenko, "A generative theory of relevance," University of Massachusetts Amherst, 2004.

[37] Y. Lv and C. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proceedings of the 18th ACM conference on Information and knowledge management*, Hong Kong, China, 2009, pp. 1895-1898.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of the Royal Statistical Society B,* vol. 39(1), pp. 1-38, 1977.

[39] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, 2004, pp. 5228-5235.

[40] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," *Proceedings of the IEEE,* vol. 88(8), pp. 1270-1278, 2000.

[41] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication,* vol. 42(1), pp. 93-108, 2004.

[42] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proceedings of European Conference on Speech Communication and Technology*, 1999, pp. 2167-2170.

[43] Y.-C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2005, pp. 5-8.

[44] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: a maximum entropy approach," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, pp. 45-48.

[45] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of Annual Conference of the International Speech Communication Association*, 2010, pp. 1045-1048.

[46] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Computer Speech and Language,* vol. 21(2), pp. 373-392, 2007.

[47] B. Chen and K.-Y. Chen, "Leveraging relevance cues for language modeling in speech recognition," *Information Processing & Management,* vol. 49(4), pp. 807-816, 2013.

[48] B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20(1), pp. 211-222, 2012.

[49] Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *Proceedings of the 29th European conference on IR research*, Rome, Italy, 2007, pp. 246-257.

[50] W. W. Edgar Meij, Jiyin He, Maarten de Rijke "Incorporating Non-Relevance Information in the Estimation of Query Models," in *TREC*, 2008.

[51] S. Cronen-Townsend and W. B. Croft, "Quantifying query ambiguity," presented at the Proceedings of the second international conference on Human Language Technology Research, San Diego, California, 2002.

[52] LDC, "Project Topic Detection and Tracking," *Linguistic Data Consortium,* 2000.

[53] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing,* vol. 10(1), pp. 219-235, 2005.

[54] A. Stolcke, *SRI Language Modeling Toolkit* (http://www.speech.sri.com/projects/srilm/), 2000.

[55] B. Chen, J.-W. Kuo, and W.-H. Tsai, "Lightly supervised and data-driven approaches to Mandarin broadcast news transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 777-780.

[56] H.-S. Lee and B. Chen, "Generalized likelihood ratio discriminant analysis," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 158-163.

[57] T. Mikolov, S. Kombrink, A. Deoras, L. a. s. Burget, and J. H. C. ´y, "RNNLM-Recurrent neural network language modeling toolkit," in *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, 2011.

[58] T. Oba, T. Hori, and A. Nakamura, "A comparative study on methods of weighted language model training for reranking LVCSR N-best hypotheses," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing* 2010, pp. 5126-5129.

[59] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 532-535.

[60] C. Berlin, H.-C. Chang, and K.-Y. Chen, "Sentence modeling for extractive speech summarization," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, 2013.

[61] Y.-W. Chen, K.-Y. Chen, H.-M. Wang, and B. Chen, "Effective pseudo-relevance feedback for spoken document retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

Publication List

[1] Yi-Wen Chen, Bo-Han Hao, Kuan-Yu Chen, Berlin Chen, "Incorporating proximity information for relevance language modeling in speech recognition," *the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, August 25-29, 2013.

[2] Yi-Wen Chen, Kuan-Yu Chen, Hsin-Min Wang, Berlin Chen, "Effective Pseudo-Relevance Feedback for Spoken Document Retrieval," *the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 26-31, 2013.

[3] Berlin Chen, Kuan-Yu Chen, Pei-Ning Chen, Yi-Wen Chen, "Spoken Document Retrieval with Unsupervised Query Modeling Techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.9, pp.2602-2612, November, 2012.

[4] Yi-Wen Chen, Jun-Yu Chen, Kuan-Yu Chen, Berlin Chen, "Empirical Comparisons of Various Pseudo-relevant Document Selection Methods for Improved Spoken Document Retrieval," *the 17th Conference on Technologies and Applications of Artificial Intelligence (TAAI 2012)*, November 16-18, 2012. (in Chinese)

[5] Ching-Huang Wang, Yi-Wen Chen, Tian-You Wu, " Self-Guided Bibliotherapy: A Case Study of a Taiwanese Doctoral Student," *the 8th International Conference on New Directions in the Humanities*, Los Angeles, USA, June 29 - July 2, 2010.

[6] Ching-Huang Wang, Yi-Wen Chen, Tian-You Wu, "Self- Guided Bibliotherapy: A Case Study of a Taiwanese Doctoral Student, " *the International Journal of the Humanities*, vol.8, no.1, pp.413-422, April, 2010.