A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents

BERLIN CHEN, National Taiwan Normal University
HSIN-MIN WANG, Academia Sinica
and
LIN-SHAN LEE, National Taiwan University

In recent years, statistical modeling approaches have steadily gained in popularity in the field of information retrieval. This article presents an HMM/N-gram-based retrieval approach for Mandarin spoken documents. The underlying characteristics and the various structures of this approach were extensively investigated and analyzed. The retrieval capabilities were verified by tests with word- and syllable-level indexing features and comparisons to the conventional vector-space model approach. To further improve the discrimination capabilities of the HMMs, both the expectation-maximization (EM) and minimum classification error (MCE) training algorithms were introduced in training. Fusion of information via indexing word- and syllable-level features was also investigated. The spoken document retrieval experiments were performed on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). Very encouraging retrieval performance was obtained.
Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*
General Terms: Algorithms, Performance, Theory
Additional Key Words and Phrases: Hidden Markov models, Mandarin spoken documents, syllable-level Indexing features

1. INTRODUCTION

Over the past three decades, statistical modeling approaches for speech and language processing have been studied extensively. Among the approaches, hidden Markov modeling (HMM) for speech recognition is undoubtedly the most prevalent and effective [Jelinek 1997]. In this approach, a set of statistical phoneme- or word-level HMMs was trained beforehand with a labeled speech corpus; the probability of the test speech utterance with respect to the HMMs was then evaluated on the HMM network to find the optimal phoneme or word sequence with the maximum likelihood. This statistical paradigm was first introduced for the information retrieval problem by BBN Technologies [Miller et al., 1999] and by Ponte and Croft [1998] and Song and Croft [1999], indicating very good potential, and was then extended in a number of publications: [Berger and Lafferty 1999; Hoffmann 1999; Lafferty and Zhai 2001; Lavrenko 2002]. Excellent survey articles on the use of statistical modeling approaches

Authors' addresses: Berlin Chen, Graduate Institute of Computer Science & Information Engineering, National Taiwan Normal University, Taipei, Taiwan; Hsin-Min Wang, Institute of Information Science, Academia Sinica, Taipei, Taiwan; Lin-shan Lee, Graduate Institute of Computer Science & Information Engineering, National Taiwan University, Taipei, Taiwan.

for information retrieval can also be found in the work of Croft and Lafferty [2003]; Liu and Croft [2003]; and Allan et al. [2003]. In these approaches, the relevance measure between the query Q and the document D is expressed as $P(D \text{ is } R|Q)$; i.e., the probability that D is relevant given that the query Q is posed. Based on the Bayes theorem and other assumptions, this relevance measure can be approximated by $P(Q|D \text{ is } R)$, or the probability of the query Q being posed, under the hypothesis that document D is relevant. The documents can therefore be ranked on the basis of this relevance measure; but most of the approaches above only address the results by using words as the indexing units, and also treat the documents to be retrieved as bags of words such that the contextual information in the documents is inevitably ignored.

Based on the observations above, we propose an HMM/N-gram-based retrieval approach for Mandarin spoken documents [Chen et al. 2001; 2002]. We model the query Q as a sequence of input observations (indexing terms, e.g., words and subword units such as syllables) and each document D as a discrete HMM composed of distributions of N-gram parameters of such observations (indexing terms) at different scales. To further improve retrieval performance, several techniques were integrated into the proposed approach. First, instead of using empirically selected weights [Ponte and Croft 1998; Song and Croft 1999], we apply the expectation-maximization (EM) training algorithm [Dempster et al. 1977] to optimize the weights for the N-gram parameters in the document HMMs; both supervised and unsupervised modes have been explored. In order to tackle the inevitable data-sparseness problem while training the N-gram probabilities from a specific document, and to model the general distribution of the indexing terms in the target language, we incorporate the N-gram parameters estimated from a general text corpus into the HMMs of the documents. The general text corpus can be a collection of texts related to the spoken-document collection. For example, we can use a newswire text corpus for a broadcast news retrieval task. In addition, we investigate the retrieval capabilities by using tests with word- and syllable(subword)-level indexing features and by comparing them to the conventional vector space model approach. We also integrate the minimum classification error (MCE) training procedure [Juang et al. 1997] into the model training process. Finally, we study the fusion of indexing features of different levels.

All the experiments mentioned in this article were performed on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). The TDT corpora have been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project [Meng et al. 2004], which is an NSF-sponsored project conducted at the Johns Hopkins University Summer Workshop 2000. Project MEI investigated the use of an entire English newswire story (text) as a query to retrieve relevant Mandarin Chinese radio news stories (audio) from the document collection. In this article we study the monolingual spoken document retrieval task instead. All the experiments were tested on the task involving the use of an entire Chinese newswire story (text) as a query to retrieve relevant Mandarin Chinese radio news stories (audio) from the document collection. Such a retrieval context is termed *query-by-example*, and can help users find the corresponding video or audio news reports, which may seem more attractive and informative than newswire text reports. Most prior work on Chinese spoken document retrieval is focused on retrieving spoken documents with short queries [Wang 2000; Meng et al. 2000; Chen et al. 2002; Chang et al. 2002].

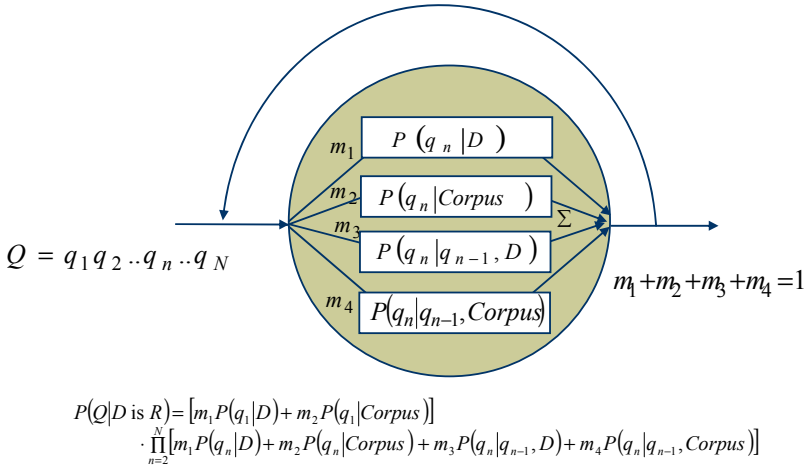
The rest of this article is organized as follows. Considerations for using word- and syllable(subword)-level indexing features for Mandarin Chinese spoken document retrieval are discussed in Section 2. The proposed retrieval model is introduced in Section 3, and some initial experimental results are discussed in Section 4. The online weight estimation, the minimum classification error (MCE) training approach, and the information fusion approach, with their corresponding experimental results, are presented in Sections 5, 6, and 7, respectively. We conclude in Section 8.

2. CONSIDERATIONS FOR USING WORD- AND SYLLABLE-LEVEL INDEXING FEATURES

There is an unknown number of words in Mandarin Chinese, although only some (e.g., 80 thousand, depending on the domain) are commonly used. Each word is composed of one or more characters, and each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are easily generated every day by combining a few characters. For example, the combination of the characters "電(electricity)" and "腦(brain)" yields the word "電腦(computer)" while the combination of "火(fire)" and "山(mountain)" yields the word "火山(volcano)". Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio, if the differences in tones are disregarded. On the other hand, an inventory of about 6000 characters provides full textual coverage of written Chinese.[1] There is a many-to-many mapping between characters and syllables. For example, the character "乾" may be pronounced as /gan1/ or /qian2/, while all of the characters "甘", "干", "柑", "肝", "竿", "尷", and "疳" are also pronounced as /gan1/, and all of the characters "前", "錢", "潛", "黔", "虔", and "掮" are pronounced as /qian2/. Consequently, a foreign word can be translated into different Chinese words based on its pronunciation. For example, Kosovo may be translated as "科索沃/ke1-suo3-wo4/", "科索佛/ke1-suo3-fo2/", "科索夫/ke1-suo3-fu1/", "科索伏/ke1-suo3-fu2/", "柯索佛/ke1-suo3-fo2/", etc.; while Al Qaeda may be translated as "蓋達/gai4-da2/", "凱達/kai3-da2/", "卡達/ka3-da2/", "卡伊達/ka3-i1-da2/", "阿爾蓋達/a1-er3-gai4- da2/". Different translations usually have some syllables in common, or may have exactly the same syllables.

The characteristics of the Chinese language lead to some special considerations when performing Mandarin Chinese speech recognition; e.g., Lee [1997] indicated that syllable recognition is an important problem. Mandarin Chinese speech recognition evaluation is usually based on syllable and character accuracy, rather than word accuracy. The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task [Chen et al. 2002; Meng et al. 2004]. Word-level indexing features possess more semantic information than subword-level features; hence, word-based retrieval enhances precision. On the other hand, subword-level indexing features behave more robustly against the Chinese word tokenization ambiguity, homophone ambiguity, open vocabulary problem, and speech recognition errors; hence, subword-based retrieval enhances recall. Accordingly, there is good reason to fuse the information obtained from indexing the features of different levels. It has been shown [Chen et al. 2002] that syllable level indexing features are very effective for Mandarin

[1] According to the GB-2312 character set, there are about 13000 traditional Chinese characters in BIG5 code.



$$P(Q|D \text{ is } R) = [m_1 P(q_1|D) + m_2 P(q_1|Corpus)] \cdot \prod_{n=2}^{N} [m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus)]$$

Fig. 1. The HMM structure for a specific document D.

Chinese spoken document retrieval; retrieval performance can be improved further by integrating information from character-level and word-level indexing features.

3. RETRIEVAL MODELS

3.1 The HMM/N-Gram-Based Model

Given a query Q and a set of documents, the retrieval system ranks the documents according to the probability that a document D is relevant, on condition that the query Q is observed; i.e., $P(D \text{ is } R|Q)$, which can be transformed to the following equation by the Bayes theorem [Jelinek 1997; Chou and Juang 2003]:

$$P(D \text{ is } R|Q) = \frac{P(Q|D \text{ is } R) P(D \text{ is } R)}{P(Q)}, \quad (1)$$

where $P(Q|D \text{ is } R)$ is the probability that query Q is posed, provided that document D is relevant, $P(D \text{ is } R)$ is the prior probability that document D is relevant, and $P(Q)$ is the prior probability that query Q is posed. The $P(Q)$ in Eq. (1) can be eliminated because it is identical for all documents and will not affect their ranking. Furthermore, since the way to estimate the probability $P(D \text{ is } R)$ is still unknown, we may simply assume that $P(D \text{ is } R)$ is uniformly distributed, or identical, for all documents. In this way we can approximate the probability $P(D \text{ is } R|Q)$ by means of the probability $P(Q|D \text{ is } R)$ for the problem here. That is, in practice, the documents are ranked according to probability $P(Q|D \text{ is } R)$.

In this research, query Q is treated as a sequence of input observations (or indexing terms), $Q = q_1 q_2 .. q_n .. q_N$, where each q_n can be a word or a syllable, while each document D is modeled by a single-state discrete HMM, as shown in Figure 1. The observation probabilities for this HMM are modeled by the weighted sum of the N-gram probabilities of words or syllables. So the relevance measure, $P(Q|D \text{ is } R)$, can be estimated by the N-gram probabilities of the indexing term sequence for the query, $Q = q_1 q_2 .. q_n .. q_N$, predicted by document D. Since the discrimination capabilities of

syllable-based bigram and unigram indexing features were shown by the vector space model approach [Chen et al. 2002], here both unigram and bigram parameters are incorporated into the HMM representation. Three types of HMM structures are studied:

Type I: Unigram-based (Uni)

$$P(Q|D \text{ is } R) = \prod_{n=1}^{N} [m_1 P(q_n|D) + m_2 P(q_n|Corpus)], \tag{2}$$

Type II: Unigram/Bigram-based (Uni+Bi)

$$P(Q|D \text{ is } R) = [m_1 P(q_1|D) + m_2 P(q_1|Corpus)] \cdot \prod_{n=2}^{N} [m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D)], \tag{3}$$

Type III: Unigram/Bigram/Corpus-based (Uni+Bi*)

$$P(Q|D \text{ is } R) = [m_1 P(q_1|D) + m_2 P(q_1|Corpus)] \cdot \prod_{n=2}^{N} [m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus)], \tag{4}$$

where m_i are weights for N-gram probabilities; $P(q_n|D)$ is the unigram probability of a specific indexing term q_n within document D; and $P(q_n|q_{n-1}, D)$ is the bigram probability for a specific indexing term sequence $q_{n-1}q_n$ within document D. In order to tackle the inevitable data sparseness problem when training the N-gram probabilities from a specific document, and to model the general distribution of the N-gram probabilities for the indexing terms, both the unigram and bigram parameters trained by a large text corpus; i.e., $P(q_n|Corpus)$ or/and $P(q_n|q_{n-1}, Corpus)$, are included in Eqs. (2) to (4) as well. The large text corpus can be a collection of texts related to the spoken document collection. For example, we can use a newswire text corpus for a broadcast news retrieval task. Notice that we did not test on the pure bigram case, since some smoothing techniques are necessary to avoid the zero-count bigrams due to the sparse data problem. In fact, the underlying idea for using Eqs. (3) and (4) is similar to that for interpolation-based language model smoothing or adaptation for speech recognition [Chen and Goodman 1999; Zhai and Lafferty 2001; Bellegarda 2004]. This can also be viewed as a combination of information from a local source (the document) and a global source (the large text corpus) [Liu and Croft 2003]. For implementing each of the three equations above, the N-gram (unigram and bigram) probabilities for generating the query observations in a specific document and in the large text corpus are estimated on the basis of the maximum likelihood principle [Chen and Goodman, 1999]. On the other hand, the weights m_i, which are summed to 1 (e.g., $\sum_{i=1}^{4} m_i = 1$ in Eq. (4)) and tied among all documents, are optimized using the expectation-maximization (EM) algorithm [Dempster et al. 1977], given a training set of query exemplars and their corresponding query-document relevance information. For example, the weight m_1 of Eq. (2) can be estimated using the following equation [Jelinek 1997; Chen et al. 2001]:

$$m_1 = \frac{\sum_{Q \in [TrainSet]_Q} \sum_{D \in [Doc]_{R \text{ to } Q}} \sum_{q_n \in Q} \left[\frac{\hat{m}_1 P(q_n|D)}{\hat{m}_1 P(q_n|D) + \hat{m}_2 P(q_n|Corpus)} \right]}{\sum_{Q \in [TrainSet]_Q} |Q| \cdot |[Doc]_{R \text{ to } Q}|}, \tag{5}$$

where $\hat{m}_1$ and $\hat{m}_2$ are the weights estimated in the previous iteration; $[TrainSet]_Q$ is the set of training query exemplars; $[Doc]_{R \text{ to } Q}$ is the set of documents that is relevant to a

specific training query exemplar Q ; $|Q|$ is the length of the query Q ; and $|[Doc]_{R\text{ to }Q}|$ is the total number of documents relevant to the query Q . Figure 1 depicts a Type III (Uni+Bi*) HMM structure for a specific document D .

3.2 Vector Space Model

The vector space model approach [Salton and McGill 1983; Baeza-Yates and Ribeiro-Neto 1999], widely used in many text information retrieval systems, was used for comparison. In this model, a document D can be represented by a set of feature vectors $\vec{d}_j$, each consisting of information for one type of indexing term, such as single words (referred to as word segments with length 1 [Chen et al. 2002]; or word unigrams [Meng et al. 2004]); or word-pairs (referred to as overlapping word segments with length 2 [Chen et al. 2002]; or overlapping word bigrams [Meng et al. 2004]). Each component z_{jt} of the feature vector $\vec{d}_j$ for a document D is associated with the weighted statistics of a specific indexing term t :

$$z_{jt} = (1 + \ln(c(t))) \cdot \ln(N/N_t), \tag{6}$$

where $c(t)$ is the occurrence count for the indexing term t within document D ; the value $1+\ln(c(t))$ denotes the term frequency for indexing term t , where the logarithmic operation is used to condense the distribution of the term frequency; $\ln(N/N_t)$ is the Inverse Document Frequency (IDF), where N_t is the number of documents that contain term t ; and N is the total number of documents in the collection. A query Q is also represented by a set of feature vectors $\vec{q}_j$ constructed in the same way. The Cosine measure is used to estimate the query-document relevance for each type of indexing terms:

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \bullet \vec{d}_j) / (\|\vec{q}_j\| \cdot \|\vec{d}_j\|). \tag{7}$$

The overall relevance is the weighted sum of the relevance scores of all types of indexing terms:

$$R(Q,D) = \sum_j w_j \cdot R_j(\vec{q}_j, \vec{d}_j), \tag{8}$$

where w_j are empirically tunable weights. We primarily use single words and word-pairs (or single syllables and syllable-pairs) in this article, since previous work indicates that they are the most effective [Chen et al. 2002; Meng et al. 2004].

4. INITIAL EXPERIMENTAL RESULTS

4.1 Corpora In the Experiments

We used two Topic Detection and Tracking (TDT) collections [LDC 2000] in this work. TDT is a DARPA-sponsored program, where participating sites tackle tasks such as identifying the first time a news story is reported on a given topic, or grouping news stories with similar topics from the audio and textual streams of newswire data. Both the English and Mandarin Chinese corpora have been studied in the recent past. The TDT corpora have also been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project [Meng et al. 2004]. In this article we use the Mandarin Chinese collection of the TDT corpora for the retrospective retrieval task, such that the statistics for the entire document collection is obtainable. The Chinese news

Table I. Statistics for TDT-2 and TDT-3 Collections In this Article

	TDT-2 (Development set) 1998, 02~06			TDT-3 (Evaluation set) 1998, 10~12		
# Spoken documents	2,265 stories, 46.03 hours of audio			3,371stories, 98.43 hours of audio		
# Distinct text queries	16 Xinhua text stories (Topics 20001~20096)			47 Xinhua text stories (Topics 30001~30060)		
	Min.	Max.	Mean	Min.	Max.	Mean
Doc. length (characters)	23	4841	287.1	19	3667	415.1
Query length (characters)	183	2623	532.9	98	1477	443.6
# Relevant documents per query	2	95	29.3	3	89	20.1

stories (text) from Xinhua News Agency are used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts are used as the spoken documents. All news stories are exhaustively tagged with event-based topic labels, which serve as the relevance judgments for performance evaluation. Table I describes the details for the corpora used in this article. The TDT-2 collection is taken as the development set, which forms the basis for tuning parameters, e.g., the weights for N-gram probabilities in Eqs. (2) to (4) and the relevance scores in Eq. (8), the number of top-ranked documents for online weight estimation (see Section 5). The TDT-3 collection is taken as the evaluation set; i.e., all the experiments performed on it were conducted following the parameter setting that was optimized based on the TDT-2 development set. Therefore, the experimental results can validate the effectiveness of the proposed approaches on comparable real-world data.

The Dragon large-vocabulary continuous speech recognizer [Zhan et al. 1999] provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3), such that the results reported here may be compared to work done by other groups. To assess the performance level of the recognizer, we spot-checked a fraction of the TDT-2 development set (about 39.90 hours) by comparing the Dragon recognition hypotheses with manual transcriptions, and obtained error rates of 35.38% (word), 17.69% (character), and 13.00% (syllable). Spot-checking approximately 76 hours of the TDT-3 test set gave error rates of 36.97% (word), 19.78% (character), and 15.06% (syllable). Notice that Dragon's recognition output contains word boundaries (tokenizations) resulting from its own language models and vocabulary definition, while the manual transcriptions ran texts without word boundaries. Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with 24k words extracted from Dragon's word recognition output, and for computing error rates used the augmented LDC lexicon (about 51k words) to tokenize the manual transcriptions. We also used this augmented LDC lexicon to tokenize the text query exemplars in the retrieval experiments.

4.2 Experimental Setup

All three types of HMM structures specified by Eqs. (2) to (4) were tested. The probabilities $P(q_n|Corpus)$ and $P(q_n|q_{n-1},Corpus)$ in these equations were estimated using a general text corpus consisting of 40 million Chinese characters, which are mainly

newswire texts collected from the Internet during January to June 2000. The weights m_i were derived by the EM training formula, as described in equation (5), using an outside training query set consisting of 819 query exemplars[2] and their corresponding query-document relevance information, with respect to the development set of TDT-2 document collection. These weights were applied to the retrieval experiments conducted on the development set (TDT-2) and the evaluation set (TDT-3). In addition (as mentioned earlier), since every Chinese word is composed of one to several syllables and syllable-level indexing features have high discriminating capabilities in retrieving Mandarin spoken documents [Chen et al., 2002; Meng et al. 2004], both the word-level and syllable(subword)-level indexing features are studied. The test results, assuming manual transcriptions for the spoken documents to be retrieved (denoted TD, text documents, in the tables below) are known, are also shown for reference, compared to the results when only the erroneous transcriptions by speech recognition are available (denoted SD, spoken documents, below). The retrieval results are expressed in terms of non-interpolated *mean* average precision (*m*AP) following the TREC evaluation [Harman 1995; Baeza-Yates and Ribeiro-Neto, 1999], which is computed by the following equation:

$$m\,\mathrm{AP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{j}{r_{i,j}}, \tag{9}$$

where L is the number of testing queries, M_i is the total number of documents that are relevant to query Q_i, and $r_{i,j}$ is the position (rank) of the j-th document that is relevant to query Q_i, counting down from the top of the ranked list.

4.3 Word-Level vs. Syllable-Level Indexing Features

Table II shows the retrieval results of the HMM/Ngram-based approach on both the TDT-2 and TDT-3 collections. From the first three columns of Table II, we can see that, for word-level indexing features, using unigram information alone achieves reasonable performance, while including bigram information generally degrades retrieval performance. By contrast, for the syllable-level indexing features (the last three columns), using unigram information alone seems inadequate (column 4), while including bigram information always improves performance significantly (columns 5 and 6). Since the number of distinct words (51k) is relatively large compared to the number of distinct syllables (0.4k), estimates of bigram probabilities for the word-level indexing features inherently suffer from the sparse data problem. The smoothing terms in Eq. (4) obtained from the 40M general text corpus obviously work pretty well for syllable-level indexing features (columns 6 versus. 5), but not as well when using word-level features (columns 3 versus 2), probably because the 40M genera text corpus is still not large enough for word bigram training; the word bigrams obtained in this way even disturbed (slighty) the uni/bigrams obtained for each document. However, by comparing the performance of the word-level against the syllable-level indexing features, we see that word-level indexing features prove superior in most cases in Table II. Syllable-level indexing features with the model in Eq. (4) perform the best for the real, desired case, i.e., the erroneous speech transcriptions (SD) of the TDT-3 evaluation set (the last row and column 6 in Table II). A larger text corpus might help improve performance in the word-based cases, but the idea

[2] Notice that the 819 query exemplars are different from the 16 test query exemplars in the development set.

Table II. Retrieval Results for HMM/N-Gram-Based Retrieval

Average Precision		Word-level			Syllable-level		
		Uni	Uni+Bi	Uni+Bi*	Uni	Uni+Bi	Uni+Bi*
TDT-2 (Dev.)	TD	0.6327	0.6069	0.5427	0.4698	0.5220	0.5718
	SD	0.5658	0.5702	0.4803	0.4411	0.5011	0.5307
TDT-3 (Eval.)	TD	0.6569	0.6542	0.6141	0.5343	0.5970	0.6560
	SD	0.6308	0.6361	0.5808	0.5177	0.5678	0.6433

Table III. Retrieval Results for the Vector Space Model

Average Precision		Word-level		Syllable-level	
		S(N), *N*=1	*S(N)*, *N*=1~2	*S(N)*, *N*=1	*S(N)*, *N*=1~2
TDT-2 (Dev.)	TD	0.5548	0.5623	0.3412	0.5254
	SD	0.5122	0.5225	0.3306	0.5077
TDT-3 (Eval.)	TD	0.6505	0.6531	0.3963	0.6502
	SD	0.6216	0.6233	0.3708	0.6353

proposed here gave the best performance in the syllable-based cases (also, due to time constraints, we were not able to enlarge the text corpus to further test the word-based cases). It is also interesting that although the word error rates for both TDT-2 and TDT-3 spoken document collections are higher than 35%, the retrieval performance for the SD cases is only slightly lower than that for TD cases. These results parallel those reported by others [Srinivasan and Petkovic 2000; Federico 2000; Renals et al. 2000].

4.4 Comparisons to the Vector Space Model

The retrieval results for the vector space model approach are shown in Table III, in which "*S(N)*, *N*=1" means using single words or single syllables as the indexing terms; "*S(N)*, *N*=1~2" means using both single words and word-pairs, or both single syllables and syllable-pairs as indexing terms. Several observations may be made from the information in Table III. First, similar to the HMM/N-gram-based approach, the word-level features outperform syllable-level features in most cases; but syllable-level features with *S(N)*, *N*=1~2 (last column) perform best for the real, desired case SD for TDT-3. Second, unlike the HMM/N-gram-based approach, using both single words and word-pairs for indexing always outperforms single words alone, although the difference is not as significant as using syllable-level indexing features. Third, using only single syllables for indexing in the vector space model approach always gives significantly poorer results than using only syllable unigram information in the HMM/N-gram-based approach. Fourth, the HMM/N-gram-based approach is consistently better than the vector space model approach (although the difference between the two is significant for the TDT-2 development set, from which the linear combination weights were trained; it is relatively small for the TDT-3 evaluation set).

5. ONLINE WEIGHT ESTIMATION

As mentioned in Section 3, the weights m_i of the document HMMs for the HMM/N-gram retrieval approach can be optimized by using the expectation-maximization (EM)

algorithm, given a training set of query exemplars and their corresponding query-document relevance information. The retrieval results of using an outside training query set to train the weights of the document HMMs are shown in Section 4. In this section we investigate the possibility of estimating the weights m_i of the document HMMs directly in the retrieval process, instead of using an outside training query set. First, for each input query, an initial retrieval is performed with the weights of the N-gram probabilities set to equal. For example, the weights m_1 and m_2 for the documents using Type I (Uni) HMM structures in Eq. (2) are set equally to 0.5 in the initial retrieval. After the initial retrieval, a list of documents, ranked according to the relevance between the query and documents, can be obtained. The top L-ranked documents are then assumed to be relevant to the query, and hence are selected for training. The EM training procedure is done in an unsupervised mode using the input query and the selected top L-ranked documents (the weights are tied among all documents, as before). Finally, there is a second retrieval which is based on the newly estimated weights m_i of the document HMMs; of course, the initial weights do not have to be equal. The weights obtained with the EM algorithm by an outside training query set, as mentioned previously, can also be used as the initial weights.

Here we investigate the retrieval performance of such an online weight estimation method by varying the initial setting of weights and the number of selected documents, L, for training. The retrieval results for the TDT-2 development set are shown in Table IV. The Type I HMM structure (Uni) in equation (2) is tested for word-level indexing features; while the Type III HMM structure (Uni+Bi*) in equation (4) is tested for syllable-level indexing features. In Table IV, the "Initial" column lists the results for initial retrieval (without using the online weight estimation method); the rest of the columns list the results for different choices of the parameter L; the row "Equal" means that the initial weights of the document HMMs are set to be equal; the row "EM" means that the initial weights of the document HMMs are trained beforehand using the outside training query set mentioned in Section 4. By setting the weights of the document HMMs as equal in the initial retrieval (in the "Equal" rows) and using an adequate number of top L-ranked documents for EM training, the online weight-estimation method can be as effective as the one that uses an outside training query set for EM training, and sometimes even performs better. For example, for the SD case with the syllable-level indexing features, the average precision increases from 0.5061 (Equal-Initial) to 0.5384, with the top 10 ranked documents applied in online weight estimation, while the average precision is 0.5307 for the EM-Initial condition. Furthermore, even if the initial weights were already trained beforehand using an outside training query set (in the "EM" rows), the online weight estimation method may sometimes improve retrieval performance further. For example, for the SD case with the syllable-level indexing features, the average precision increases from 0.5307 to 0.5399, with the top 10 ranked documents applied in online weight estimation. Similar trends are observable for both the word-level indexing features and the syllable-level indexing features in Table IV. Considering performance and efficiency, selecting the top 10 ranked documents for online weight estimation turns out to be adequate.

We further evaluate the performance of the online weight estimation method with the parameter L=10, as chosen above, on the TDT-3 evaluation set. The retrieval results are shown in Table V; detailed results with $L\neq10$ are also provided for reference. Similar trends to those in Table IV can be seen in Table V. For the SD case with syllable-level

Table IV. Retrieval Results for the TDT-2 Development Set Using Online Weight Estimation for HMM/N-Gram-Based Retrieval

Average Precision			Initial	*L* for online weight estimation						
				1	5	10	20	30	40	50
Word -level (Uni)	TD	Equal	0.5744	0.6326	0.6326	0.6333	0.6359	0.6380	0.6413	0.6416
		EM	0.6327	0.6325	0.6346	0.6351	0.6338	0.6339	0.6377	0.6378
	SD	Equal	0.5300	0.5784	0.5779	0.5746	0.5748	0.5721	0.5721	0.5735
		EM	0.5658	0.5743	0.5747	0.5781	0.5746	0.5725	0.5801	0.5786
Syllable -level (Uni+Bi*)	TD	Equal	0.5409	0.5852	0.5823	0.5849	0.5823	0.5736	0.5699	0.5709
		EM	0.5718	0.5760	0.5818	0.5825	0.5791	0.5700	0.5700	0.5714
	SD	Equal	0.5061	0.5239	0.5341	0.5384	0.5410	0.5415	0.5390	0.5378
		EM	0.5307	0.5274	0.5385	0.5399	0.5377	0.5385	0.5384	0.5372

Table V. Retrieval Results for the TDT-3 Evaluation Set Using Online Weight Estimation for HMM/N-Gram-Based Retrieval

Average Precision			Initial	*L* for online weight estimation						
				1	5	10	20	30	40	50
Word -level (Uni)	TD	Equal	0.6100	0.6617	0.6594	0.6596	0.6602	0.6599	0.6603	0.6608
		EM	0.6569	0.6616	0.6605	0.6597	0.6600	0.6606	0.6607	0.6603
	SD	Equal	0.5733	0.6253	0.6253	0.6255	0.6268	0.6283	0.6279	0.6277
		EM	0.6308	0.6255	0.6265	0.6269	0.6286	0.6292	0.6281	0.6288
Syllable -level (Uni+Bi*)	TD	Equal	0.6015	0.6309	0.6379	0.6374	0.6418	0.6436	0.6448	0.6451
		EM	0.6560	0.6231	0.6340	0.6379	0.6473	0.6473	0.6466	0.6466
	SD	Equal	0.5821	0.6151	0.6296	0.6344	0.6412	0.6424	0.6418	0.6396
		EM	0.6433	0.6236	0.6328	0.6378	0.6434	0.6439	0.6427	0.6390

indexing features, the average precision improves from 0.5821 (Equal-Initial) to 0.6344 with the top 10 ranked documents applied in online weight estimation (although L=20~50 actually offers slightly better results in this case), and the precision rate of 0.6344 is in fact very close to the value 0.6433 for the EM-Initial condition, which requires an outside set of training queries.

Notice here that online weight estimation offers functions similar to the conventional blind relevance feedback [Baeza-Yates and Ribeiro-Neto 1999; Jourlin et al. 2000; Liu and Croft 2003], which automatically modifies or expands the original query representation by using the indexing terms of the top L-ranked documents and tries to improve the retrieval performance in the second retrieval. In the online weight estimation here, we re-estimate the weights of the document HMMs using the additional information from the top L-ranked documents. It is worth mentioning that to perform either the online weight estimation or the blind relevance feedback, the retrieval system needs to search the database twice. To avoid serious reductions in efficiency, it is recommended that such techniques are not applied iteratively more than once.

6. MINIMUM CLASSIFICATION ERROR (MCE) TRAINING

The minimum classification error (MCE) training algorithm [Juang et al. 1997; Chou and Juang 2003] used widely in HMMs for speech recognition can be applied here to improve the discrimination of the HMMs in the HMM/N-gram-based retrieval approach. Given a query Q and a relevant document D^*, we can define the classification error function as follows:

$$E(Q,D^*)=\frac{1}{|Q|}\left[-\log P\left(Q\middle|D^* \text{ is } R\right)+\max_{D'}\log P\left(Q\middle|D' \text{ is not } R\right)\right], \tag{10}$$

where D' is the irrelevant document with the highest relevance score (i.e., the highest ranked irrelevant document), and $|Q|$ is the length of the query Q. In order to find D', an initial retrieval is first conducted in each iteration of MCE training to obtain a ranked list of the documents for each training query Q. Then, with the query-document relevance information, it is easy to identify the irrelevant document with the highest relevance score by checking down from the top of the ranked list. Notice that the classification error function in equation (10) can be extended to take more than one irrelevant document into consideration, e.g., the first *K* irrelevant documents with the highest relevance scores, by modifying the second term in the right-hand side of the equation. However, for simplicity, only the irrelevant document with the highest relevance score, or *K*=1, is selected for training in this study. There were previous detailed discussions of this issue [Rahim et al. 1997; Juang et al. 1997]. The classification error function in equation (10) can be transformed into a loss function ranging from 0 to 1 with the Sigmoid operator:

$$L(Q,D^*)=\frac{1}{1+\exp(-\alpha E(Q,D^*)+\beta)}, \tag{11}$$

where α is a positive constant that controls the slope of the function and β is an offset factor (set to zero here, for simplicity). Obviously, when $E(Q,D^*)$ is much smaller than 0, which implies correct classification or retrieval, virtually no loss is incurred. When $E(Q,D^*)$ is positive, this leads to a penalty that becomes essentially a classification or retrieval error count. The loss function in Eq. (11) can be further minimized according to an iterative procedure, such that the linear combination weights m_i of the HMM for the document D^* may consequently be iteratively updated.

Here we take the Type I HMM structure defined in Eq. (2) as an example for describing the details. Since the weights m_1 and m_2 are nonnegative and must be subject to the constraint $m_1+m_2=1$ during the iterative updating, we can express them as follows to facilitate the derivation of formulas:

$$m_1=\frac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}} \text{ and } m_2=\frac{e^{\tilde{m}_2}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}. \tag{12}$$

where $\tilde{m}_1$ and $\tilde{m}_2$ are the transformed versions of weights m_1 and m_2, respectively. Once $\tilde{m}_1$ and $\tilde{m}_2$ are obtained, m_1 and m_2 can be derived easily. Therefore, we can minimize the loss function defined in Eq. (11) by iteratively adjusting the weight $\tilde{m}_1$

(similarly for $\tilde{m}_2$) of the document D^* according to the following gradient descent procedure:

$$\tilde{m}_1(i+1) = \tilde{m}_1(i) - \varepsilon(i) \cdot \frac{\partial L(Q, D^*)}{\partial \tilde{m}_1}\Big|_{D^* = D^*(i)}, \tag{13}$$

where $\varepsilon(i)$ is a monotonically decreasing positive number used in each iteration i to control the training rate, while $\tilde{m}_1(i)$ and $D^*(i)$ are, respectively, the values of $\tilde{m}_1$ and the corresponding HMM for D^* at the *i*-th iteration. Gradient descent is also often referred to as generalized probabilistic descent (GPD) in the literature [Amari 1972; Chou and Juang 1992]. If we define $\nabla_{D^*,\tilde{m}_1} \equiv \varepsilon(i) \cdot \frac{\partial L(Q, D^*)}{\partial \tilde{m}_1}$ as a weighted gradient function of the loss function $L(Q, D^*)$, then $\nabla_{D^*,\tilde{m}_1}$ can be written as

$$\nabla_{D^*,\tilde{m}_1} = \varepsilon(i) \cdot \frac{\partial L(Q, D^*)}{\partial E(Q, D^*)} \cdot \frac{\partial E(Q, D^*)}{\partial \tilde{m}_1}, \tag{14}$$

where $\frac{\partial L(Q, D^*)}{\partial E(Q, D^*)}$ and $\frac{\partial E(Q, D^*)}{\partial \tilde{m}_1}$ can be derived as, respectively,

$$\frac{\partial L(Q, D^*)}{\partial E(Q, D^*)} = \alpha \cdot L(Q, D^*) \cdot \left[1 - L(Q, D^*)\right], \tag{15}$$

and

$$\frac{\partial E(Q, D^*)}{\partial \tilde{m}_1} = -\left[-m_1 + \frac{1}{|Q|} \sum_{q_n \in Q} \frac{m_1 P(q_n | D^*)}{m_1 P(q_n | D^*) + m_2 P(q_n | Corpus)} \right]. \tag{16}$$

Based on equations (14) to (16), equation (13) can be further expressed as

$$\tilde{m}_1(i+1) = \tilde{m}_1(i) + \varepsilon(i) \cdot \alpha \cdot L(Q, D^*)\left[1 - L(Q, D^*)\right] \cdot \left[-m_1 + \frac{1}{|Q|} \sum_{q_n \in Q} \frac{m_1 P(q_n | D^*)}{m_1 P(q_n | D^*) + m_2 P(q_n | Corpus)} \right]. \tag{17}$$

As a result, the weight m_1 of the HMM of document D^* can be adjusted iteratively by using the following equation:

$$m_1(i+1) = \frac{m_1(i) \cdot e^{-\nabla_{D^*,\tilde{m}_1}(i)}}{m_1(i) \cdot e^{-\nabla_{D^*,\tilde{m}_1}(i)} + m_2(i) \cdot e^{-\nabla_{D^*,\tilde{m}_2}(i)}}. \tag{18}$$

The details for the derivation of equations (15), (16), and (18) are given in the Appendix. The weight m_2 of the document D^* can be updated iteratively in a similar way. When the document is relevant to more than one training query Q, this procedure can be performed consecutively with all these queries in a single iteration, and then repeated recursively.

The training query exemplars in Section 4 for EM training are again used here for MCE training. It is worth mentioning that the goal of MCE training is to correctly discriminate among query observations for the best retrieval results, rather than to fit the distributions of query observations, as in the EM training. In this research, the Type I HMM structure (Uni) in equation (2) is tested for word-level indexing features, while the

Table VI. Retrieval Results for HMM/N-Gram-Based Retrieval With and Without MCE Training

Average Precision		Word-level (Uni)		Syllable-level (Uni+Bi*)	
		EM	MCE	EM	MCE
TDT-2 (Dev.)	TD	0.6327	0.6459	0.5718	0.6858
	SD	0.5658	0.5810	0.5307	0.6300
TDT-3 (Eval.)	TD	0.6569	0.6503	0.6560	0.7026
	SD	0.6308	0.6331	0.6433	0.6814

Type III HMM structure (Uni+Bi*) in equation (4) is tested for syllable-level indexing features. The retrieval results of the TDT-2 development set with MCE training (after 100 iterations) are first shown in the upper part of Table VI, where we can see that with syllable-level indexing features (right-half of the table), the average precisions improve significantly from 0.5718 to 0.6858 in the TD case and from 0.5307 to 0.6300 in the SD case. There are similar improvements, although not as significant, with word-level indexing features (left-half of the table). It is also very interesting that, with MCE training, the syllable-level indexing features significantly outperform those of word-level indexing for both the TD and SD cases. The reverse is true when EM training is used alone. Figures 2 and 3 depict MCE training curves for, respectively, word- and syllable-based indexing approaches for this experiment with the TDT-2 development set. All the results reported with MCE training are obtained with 100 iterations.

Since the weights of the document HMMs are no longer tied together here, those for the TDT-2 development set cannot be used for the TDT-3 evaluation set. In order to validate the effectiveness of the MCE training on the TDT-3 evaluation set, another outside training query set consisting of 777 query exemplars,[3] with their corresponding query-document relevance information to the TDT-3 evaluation set, is used in MCE training. Because the experiment is now conducted on the evaluation set, the parameter settings (α and $\varepsilon(i)$), as well as the number of iterations in MCE training, are taken directly from those tuned in the TDT-2 development set. The retrieval results of the TDT-3 evaluation set are shown in the lower part of Table VI, where the effects of MCE training on retrieval performance for word-level indexing features is not apparent. Compared to the results of EM training, the average precision for TD degrades slightly from 0.6569 to 0.6503, while the average precision for SD improves slightly from 0.6308 to 0.6331. However, in contrast, MCE training provides a great boost to both the TD and SD cases when syllable-level indexing features are used. The average precisions improve considerably from 0.6560 and 0.6433 to 0.7026 and 0.6814, respectively, for the TD and SD cases.

7. INFORMATION FUSION FOR HMM/N-GRAM-BASED RETRIEVAL

The word-level indexing features have more semantic information than syllable-level features. On the other hand, syllable-level indexing features provide a more robust relevance measure between queries and documents when dealing with the problems arising from the flexible wording structure in Mandarin Chinese and the errors in speech recognition of spoken documents [Chen et al. 2002; Meng et al. 2004]. We believe that a proper fusion of syllable- and word-level information will be useful for the retrieval task studied here. As a result, fusion of the retrieval results with respect to the word- and

[3] Notice that the 777 query exemplars are different than the 47 test query exemplars in the evaluation set.

Table VII. Retrieval Results for HMM/N-Gram-Based Retrieval After Fusion of Word- and Syllable-Level Information

Average Precision		λ=0.1	λ=0.2	λ=0.3	λ=0.4	λ=0.5	λ=0.6	λ=0.7	λ=0.8	λ=0.9
TDT2 (Dev.)	TD	0.6877	0.7113	0.7329	0.7392	0.7141	0.7048	0.6886	0.6814	0.6625
	SD	0.6353	0.6614	0.6914	0.6880	0.6602	0.6569	0.6483	0.6123	0.5983
TDT3 (Eval.)	TD	0.7194	0.7201	0.7174	0.7077	0.6964	0.6889	0.6797	0.6695	0.6586
	SD	0.7148	0.7156	0.7096	0.6989	0.6880	0.6778	0.6631	0.6547	0.6438

syllable-level indexing features in Section 6 is tested using the following formula:

$$R(Q,D) = \lambda \cdot \log R_w(Q,D) + (1-\lambda) \cdot \log R_s(Q,D), \tag{19}$$

which is simply the weighted sum of the log relevance scores, $R_w(Q,D)$ and $R_s(Q,D)$, respectively, obtained with word- and syllable-level indexing features alone. The value of the weight λ can be empirically adjusted between 0 and 1. We first attempt to investigate the results of fusing word- and syllable-level information on the TDT-2 development set (with MCE-trained models) by increasing the value of λ from 0.1 to 0.9 with step size 0.1. The retrieval results are shown in the upper part of Table VII. When compared to the results obtained by using either the word-level (i.e., λ=1, in the column denoted "Word-level (Uni)" and "MCE" in Table VI) or syllable-level information alone (i.e., λ=0, in the column denoted "Syllable-level (Uni+Bi*)" and "MCE" in Table VI), we can see that the average precision after information fusion is always better than obtained by world-level features alone. This is apparently due to the retrieval performance with the syllable-level indexing features is significantly better than that with the word-level indexing features. As the value of λ is equal to or lower than 0.7 (i.e., putting more emphasis on the syllable-level features), the average precisions after information fusion will start to become better than those obtained with the syllable-level features alone, and the best average precisions of 0.7392 (λ=0.4) and 0.6914 (λ=0.3) are obtained for the TD and SD cases, respectively, which indicates significant improvement. We then further apply these best settings of weight λ (i.e., λ=0.4 for TD and λ=0.3 for SD) to the TDT-3 evaluation set (with MCE-trained models). As shown in the lower part of Table VII, the average precisions are 0.7077 and 0.7096 for the TD and SD cases, respectively. Although these results are better than those obtained by using either word- or syllable-level features alone, the improvements are not as significant as those for the TDT-2 development set. If we also perform information fusion by varying the value of λ in the same way as we did before in the TDT-2 task, the best retrieval results of 0.7201 (λ=0.2) and 0.7156 (λ=0.2) are obtained for the TD and SD cases, respectively. Comparing these results to those following the best setting indicated by the TDT-2 development task, the setting tuned from the TDT-2 development set performs rather well in the TDT-3 evaluation set, although not optimally.

Based on the experimental results in this and previous sections, we conclude that the syllable-based approach is better than the word-based one for the Mandarin Chinese spoken document retrieval task, although many research results indicate that the word-based approach is very useful for similar tasks in Western languages such as English [Ng et al. 2000; Ng 2000]. All the experiments in this article have been carefully designed to avoid "testing on training"; i.e., all the parameters are tuned by using the TDT-2 development set and tested on both the TDT-2 development set and the TDT-3 evaluation

set. Generally speaking, the parameters tuned from the TDT-2 development set perform rather well in the TDT-3 evaluation set, although not optimally.

8. CONCLUSIONS

In this article we presented an HMM/N-gram-based retrieval approach for Mandarin Chinese spoken documents. We have extensively investigated its underlying characteristics and structures, verified its retrieval capabilities by using indexing features of different levels, and compared it with the vector space model approach. The minimum classification error (MCE) training was introduced during the training phase to improve discrimination among the document HMMs. We found that, given a set of training query exemplars, the retrieval performance can be significantly improved by MCE training, which means that an HMM/N-gram-based retrieval system can be incrementally improved through use. In addition, fusing the information from the indexing features of different levels is shown to be useful.

APPENDIX

The derivations of equations (15), (16), and (18) are detailed in the following three equations, respetively.

$$\frac{\partial L(Q,D^*)}{\partial E(Q,D^*)} = \frac{\partial\left\{\dfrac{1}{1+\exp\left(-\alpha E(Q,D^*)+\beta\right)}\right\}}{\partial\left\{\exp\left(-\alpha E(Q,D^*)+\beta\right)\right\}}$$

$$= \frac{\alpha\cdot\exp\left(-\alpha E(Q,D^*)+\beta\right)}{\left(1+\exp\left(-\alpha E(Q,D^*)+\beta\right)\right)^2} \tag{A1}$$

$$= \alpha\cdot\frac{1}{\left(1+\exp\left(-\alpha E(Q,D^*)+\beta\right)\right)}\left[1-\frac{1}{\left(1+\exp\left(-\alpha E(Q,D^*)+\beta\right)\right)}\right]$$

$$= \alpha\cdot L(Q,D^*)\cdot\left[1-L(Q,D^*)\right],$$

$$\frac{\partial E(Q,D^*)}{\partial \tilde{m}_1} = \frac{-1}{|Q|}\frac{\partial\left\{\sum_{q_n\in Q}\log\left[\dfrac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|D^*\right)+\dfrac{e^{\tilde{m}_2}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|Corpus\right)\right]\right\}}{\partial\tilde{m}_1}$$

$$= \frac{-1}{|Q|}\sum_{q_n\in Q}\frac{\dfrac{-e^{\tilde{m}_1}}{\left(e^{\tilde{m}_1}+e^{\tilde{m}_2}\right)^2}\left[e^{\tilde{m}_1}P\left(q_n\middle|D^*\right)+e^{\tilde{m}_2}P\left(q_n\middle|Corpus\right)\right]+\dfrac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|D^*\right)}{\dfrac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|D^*\right)+\dfrac{e^{\tilde{m}_2}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|Corpus\right)} \tag{A2}$$

$$= \frac{e^{\tilde{m}_1}}{\left(e^{\tilde{m}_1}+e^{\tilde{m}_2}\right)^2}-\frac{1}{|Q|}\sum_{q_n\in Q}\frac{\dfrac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|D^*\right)}{\dfrac{e^{\tilde{m}_1}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|D^*\right)+\dfrac{e^{\tilde{m}_2}}{e^{\tilde{m}_1}+e^{\tilde{m}_2}}P\left(q_n\middle|Corpus\right)}$$

$$= -\left[-m_1+\frac{1}{|Q|}\sum_{q_n\in Q}\frac{m_1P\left(q_n\middle|D^*\right)}{m_1P\left(q_n\middle|D^*\right)+m_2P\left(q_n\middle|Corpus\right)}\right],$$

$$
\begin{aligned}
m_1(i+1) &= \frac{e^{\tilde{m}_1(i+1)}}{e^{\tilde{m}_1(i+1)} + e^{\tilde{m}_2(i+1)}} \\
&= \frac{e^{\tilde{m}_1(i)} e^{-\nabla_{D^*,\tilde{m}_1}(i)}}{e^{\tilde{m}_1(i)} e^{-\nabla_{D^*,\tilde{m}_1}(i)} + e^{\tilde{m}_2(i)} e^{-\nabla_{D^*,\tilde{m}_2}(i)}} \\
&= \frac{e^{\tilde{m}_1(i)} e^{-\nabla_{D^*,\tilde{m}_1}(i)} \Big/ \left(e^{\tilde{m}_1(i)} + e^{\tilde{m}_2(i)}\right)}{\left[e^{\tilde{m}_1(i)} e^{-\nabla_{D^*,\tilde{m}_1}(i)} \Big/ \left(e^{\tilde{m}_1(i)} + e^{\tilde{m}_2(i)}\right)\right] + \left[e^{\tilde{m}_2(i)} e^{-\nabla_{D^*,\tilde{m}_2}(i)} \Big/ \left(e^{\tilde{m}_1(i)} + e^{\tilde{m}_2(i)}\right)\right]} \\
&= \frac{m_1(i) \cdot e^{-\nabla_{D^*,\tilde{m}_1}(i)}}{m_1(i) \cdot e^{-\nabla_{D^*,\tilde{m}_1}(i)} + m_2(i) \cdot e^{-\nabla_{D^*,\tilde{m}_2}(i)}}.
\end{aligned}
\tag{A3}
$$

REFERENCES

ALLAN, J. (ED.), ASLAM, J., BELKIN, N., BUCKLEY, C., CALLAN, J., CROFT, W. B. (ED.), DUMAIS, S., FUHR, N., HARMAN, D., HARPER, D., HIEMSTRA, D., HOFMANN, T., HOVY, E., KRAAIJ, W., LAFFERTY, J., LAVRENKO, V., LEWIS, D., LIDDY, L., MANMATHA, R., MCCALLUM, A., PONTE, J., PRAGER, J., RADEV, D., RESNIK, P., ROBERTSON, S., ROSENFELD, R., ROUKOS, S., SANDERSON, M., SCHWARTZ, R., SINGHAL, A., SMEATON, A., TURTLE, H., VOORHEES, E., WEISCHEDEL, R., XU, J., AND ZHAI, C. 2003. Challenges in information retrieval and language modeling. *SIGIR Forum 37,* 1 (2003), 31-47.

AMARI, S. I. 1972. Learning patterns andpattern sequences by self-organizing nets of threshold elements. *IEEE Trans. on Computer C-12*, 11 (1972), 1197-1206.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison Wesley, Reading, MA.

BELLEGARDA, J. R. 2004. Statistical language model adaptation: Review and perspectives. *Speech Communication 42* (2004), 93–108.

BERGER, A. AND LAFFERTY, J. 1999. Information retrieval as statistical translation. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

CHANG, E., SEIDE, F., MENG, H., CHEN, Z., SHI, Y., AND LI, Y. C. 2002. A system for spoken query information retrieval on mobile devices. *IEEE Trans. on Speech and Audio Processing 10,* 8 (2002), 531-541.

CHEN, B., WANG, H. M., AND LEE, L. S. 2001. An HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval. In *Proceedings of the European Conference on Speech Communication and Technology*.

CHEN, B., WANG, H. M., AND LEE, L. S. 2002. Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese. *IEEE Trans. on Speech and Audio Processing 10*, 5 (2002), 303-314.

CHEN, S. F. AND GOODMAN, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13* (1999), 359–394.

CHOU, W. AND JUANG, B. H. 1992. Adaptive discriminative learning in pattern recognition. Tech. Rep., AT&T Bell Laboratories, 1992.

CHOU, W. AND JUANG, B. H. (EDS.) 2003. *Pattern Recognition in Speech and Language Processing*. CRC Press.

CROFT, W. B. AND LAFFERTY, J. (EDS.) 2003. *Language Modeling for Information Retrieval*. Kluwer, Amsterdam.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B(39)*, (1977), 1-38.

FEDERICO, M. 2000. A system for the retrieval of Italian broadcast news. *Speech Communication 32* (2000), 37-47.

HARMAN, D. 1995. Overview of the Fourth Text Retrieval Conference (TREC-4). Available at http://trec.nist.gov/pubs/trec4/overview.ps.

HOFFMANN, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

JELINEK, F. 1997. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

JOURLIN, P., JOHNSON, S. E., SPARCK JONES, K., AND WOODLAND, P. C. 2000. Spoken document representations for probabilistic retrieval. *Speech Communication 32* (2000), 21-36.

JUANG, B. H., CHOU, W., AND LEE, C. H. 1997. Minimum classification error rate methods for speech recognition. *IEEE Trans. on Speech and Audio Processing 5, 3* (1997), 257-265.

LAFFERTY, L. AND ZHAI, C. 2001. Document language models and risk minimization for information retrieval. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

LAVRENKO, V. 2002. Optimal mixture models in IR. In *Proceedings of the 24th European Colloquium on IR Research* (*ECIR'02*).

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

LEE, L. S. 1997. Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine 14*, 4 (1997), 63-101.

LIU, X. AND CROFT, W. B. 2003. Statistical language modeling for information retrieval. To appear in *The Annual Review of Information Science and Technology 39* (2005).

MENG, H., LO W. K., LI, Y. C., AND CHING, P. C. 2000. Multi-scale audio indexing for spoken document retrieval. In *Proceedings of the International Conference on Spoken Language Processing*.

MENG, H., CHEN, B., KHUDANPUR, S., LEVOW, G. A., LO, W. K., OARD, D., SCHONE, P., TANG, K., WANG, H. M., AND WANG, J. 2004. Mandarin–English Information (MEI): Investigating translingual speech retrieval. *Computer Speech and Language 18,* 2 (2004), 163-179.

MILLER, D. R. H., LEEK, T., AND SCHWARTZ, R. 1999. A hidden Markov model information retrieval system. In *Proceedings of ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

NG, G., WILKINSON, R., AND ZOBEL, J. 2000. Experiments in spoken document retrieval using phoneme N-grams. *Speech Communication 32* (2000), 61-77.

NG, K. 2000. Information fusion for spoken document retrieval. In *Proceedings of IEEE International Conference on Acoustic, Speech, Signal Processing*.

PONTE, J. M. AND CROFT, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

RAHIM, M. G., LEE, C. H., AND JUANG, B. H. 1997. Discriminative utterance verification for connected digit recognition. *IEEE Trans. on Speech and Audio Processing 5,* 3 (1997), 266-277.

RENALS, S., ABBERLEY, D., KIRBY, D., AND ROBINSON, T. 2000. Indexing and retrieval of broadcast news. *Speech Communication 32* (2000), 5-20.

SALTON, G. AND MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Song, F. and Croft, W. B. 1999. A general language model for information retrieval. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management*. ACM, New York.

SRINIVASAN, S. AND PETKOVIC, D. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

WANG, H. M. 2000. Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese. *Speech Communication 32* (2000), 49-60.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the ACM SIGIR Conference on R&D in Information Retrieval*. ACM, New York.

Zhan, P., Wegmann, S., and Gillick, L. 1999. Dragon Systems' 1998 broadcast news transcription system for Mandarin. In *Proceedings of the DARPA Broadcast News Workshop*.

Received October 2003; revised February 2004; accepted June 2004