An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval

Berlin Chen[1,2], Hsin-min Wang[1], and Lin-shan Lee[1,2]

[1] Institute of Information Science, Academia Sinica,
[2] Dept. of Computer Science & Information Engineering, National Taiwan University,
Taipei, Taiwan, Republic of China
E-mail: {berlin, whm, lsl}@iis.sinica.edu.tw

ABSTRACT

In this paper an HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval is presented. The underlying characteristics and different structures of this approach were extensively investigated. The retrieval capabilities were verified by tests with indexing features of word- and syllable(subword)-levels and comparison with 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. The information fusion of indexing features of word- and syllable-levels 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.

1. INTRODUCTION

The hidden Markov model (HMM) has been the most prevailing approach for speech recognition [1-2]. In this approach, a set of statistical phoneme- or word-level HMMs was trained beforehand with the labeled speech corpus, and the probability of the test speech utterance with respective to these HMMs was then evaluated on the HMM network to find the optimal phoneme or word sequence with the maximal likelihood probability. This statistical paradigm was first introduced to the information retrieval problem by the BBN Group [3] with very good potential indicated. A similar approach was also proposed by Song and Croft [4]. 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 was posed. Based on Bayes' theorem and some assumptions, this relevance measure can be approximated by $P(Q|D \text{ is } R)$, which stands for the probability of the query Q being posed, under the hypothesis that document D is relevant. The documents can therefore be ranked based on this relevance measure.

In this paper, we presented an HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval based on the above concepts. We modeled the query Q as a sequence of input observations (indexing terms, e.g. words and syllables) and each document D as a discrete HMM composed of distributions of N-gram parameters of such observations (indexing terms) of different scales. We applied the expectation-maximization (EM) training algorithm [1] to optimize the weights for the N-gram parameters in the document HMMs instead of using empirically selected weights [4]. We also incorporated the N-gram parameters estimated from a general text corpus to the HMMs of the documents. Furthermore, we investigated the retrieval capabilities by tests with indexing features of word- and syllable(subword)-levels and comparison with the conventional vector space model approach. The minimum classification error (MCE) training procedure was also introduced in training, and very encouraging improvements were achieved. Finally, the information fusion of indexing features of different levels was studied as well.

In the following, 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 broadcast news stories (audio) in the document collection. Such a retrieval context is termed *query-by-example*.

2. EXPERIMENTAL CORPORA

We used two Topic Detection and Tracking (TDT) collections for this work. TDT-2 is taken as the development test set while TDT-3 is taken as the evaluation test set. The Chinese news stories (text) from Xinhua News Agency were used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. Table 1 describes the details for the corpora used in this paper.

The Dragon large-vocabulary continuous speech recognizer [5] provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3), such that the results here may be compared with works done by other groups. We have spot-checked a fraction of the TDT-2 development set (of 39.90 hours) by comparing the Dragon recognition hypotheses with the 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 language models and vocabulary definition while the manual transcriptions are running texts without word boundaries. Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with the 24k words extracted from Dragon's word recognition output, and used the augmented LDC lexicon in tokenizing the manual transcriptions for computing error rates. We also used this augmented LDC lexicon in tokenizing the text query

	TDT2 (Development) 1998, 02~06			TDT3 (Evaluation) 1998, 10~12		
# Spoken documents	2,265 stories, 46.03hrs of audio			3,371stories, 98.43hrs 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

Table 1: Statistics of TDT-2 and TDT-3 collections used in this paper.

exemplars in the retrieval experiments.

3. RETRIEVAL MODELS

3.1 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 D is relevant, conditioned on the fact that the query Q is observed; i.e., $P(D \text{ is } R|Q)$, which can be transformed to the following equation by applying Bayes' theorem [1]:

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

where $P(Q|D \text{ is } R)$ is the probability of the query Q being posed under the condition 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 of query Q being posed. $P(Q)$ in (1) can be eliminated because it is identical for all documents. Furthermore, because there is no general way to estimate the probability $P(D \text{ is } R)$, we can simply set it to unity for simplicity and approximate the probability $P(D \text{ is } R|Q)$ by the probability $P(Q|D \text{ is } R)$ for the problem studied here.

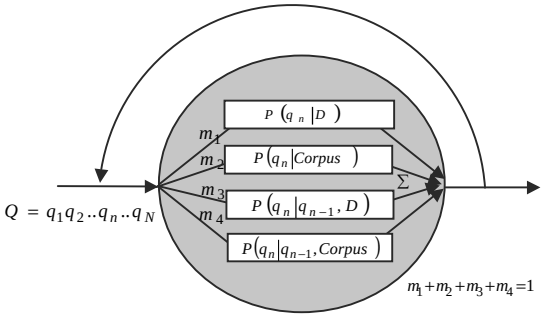
In this research, the query Q is treated as a sequence of input observations (or indexing terms), $Q = q_1 q_2 .. q_n .. q_N$, where each q_n can be a word or a syllable, while each document D is modeled by a single-state discrete HMM as shown in Figure 1. The observation probabilities for this HMM are modeled by the weighted sum of N-gram probabilities of words or syllables. Therefore, 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 the document D. In the present work, both unigram and bigram parameters were incorporated into the HMM representation and three types of HMM structures were studied:

Type I: Unigram-based (Uni)

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

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

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



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

Figure 1: The HMM structure for a specific document D.

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

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

where $P(q_n|D)$ is the unigram probability of a specific indexing term q_n within the document D and $P(q_n|q_{n-1}, D)$ is the bigram probability of a specific indexing term sequence $q_{n-1}q_n$ within the document D. In order to model the general distribution of the indexing terms, both the unigram and bigram parameters trained by a large text corpus; i.e., $P(q_n|Corpus)$ or/and $P(q_n|q_{n-1}, Corpus)$, were included as well in Equations (2)-(4). In addition, for each of the above three equations, the weights m_i were summed to 1 (e.g. $\sum_{i=1}^{4} m_i = 1$ in Equation (4)) and the weights were tied among all documents. These weights can be optimized using the expectation-maximization (EM) algorithm [1] given a training set of query exemplars and their corresponding query-document relevance information. For example, the weight m_1 of Equation (2) can be estimated using the following equation:

$$m_1 = \frac{\sum_{Q \in [TrainSet]_Q} \sum_{D \in [Doc]_{R \text{ to } Q}} \sum_{q_n \in Q} \left[\frac{m_1 P(q_n|D)}{m_1 P(q_n|D) + m_2 P(q_n|Corpus)} \right]}{\sum_{Q \in [TrainSet]_Q} |Q| \cdot |[Doc]_{R \text{ to } Q}|}, \quad (5)$$

where $[TrainSet]_Q$ is the set of training query exemplars, $[Doc]_{R \text{ to } Q}$ is the set of documents that are 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 the Type III (Uni+Bi*) HMM structure for a specific document D.

3.2 Vector Space Model

In the vector space model approach, a document D can be represented by a set of feature vectors $\vec{d}_s$, each consisting of information for one type of indexing terms [6-7], such as a single word or a word-pair, and so on. Each component $g(t)$ of a feature vector $\vec{d}$ for a document D is associated with the statistics of a specific indexing term t:

$$g(t) = (1 + \ln(c(t))) \cdot \ln(N/N_t), \quad (6)$$

where $c(t)$ is the occurrence count of the indexing term t within the document D, and the value of $1 + \ln(c(t))$ denotes

the term frequency for indexing term t, where the logarithmic operation is 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 include the 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 $\vec{q}_s$ constructed in the same way. The Cosine measure is used to estimate the query-document relevance for each type of indexing terms:

$$R_s(\vec{q}_s, \vec{d}_s) = \left(\vec{q}_s \bullet \vec{d}_s\right) \Big/ \left(\|\vec{q}_s\| \cdot \|\vec{d}_s\|\right). \quad (7)$$

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

$$R(Q, D) = \sum_s w_s \cdot R_s(\vec{q}_s, \vec{d}_s), \quad (8)$$

where w_s are empirically tunable weights.

4. INITIAL EXPERIMENTAL RESULTS

4.1 Experiment Setup

All the three types of HMM structures specified by Equations (2)-(4) were tested. The probabilities of $P(q_n|Corpus)$ and $P(q_n|q_{n-1}, Corpus)$ in these equations were estimated using a general text corpus consisting of 40 million Chinese characters. 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 and their corresponding query-document relevance information with respect to the development set of TDT-2 document collection. These weights were applied to the evaluation set of the TDT-3 document collection. In addition, because every Chinese word is composed of one to several syllables and syllable-level indexing features have been shown to have high discriminating capabilities in retrieving Mandarin spoken documents [6-8], both the word-level and syllable(subowrd)-level indexing features were studied. The test results with manual transcriptions of the spoken documents (denoted as TD in the result tables below) were also shown for reference, compared to the results with the erroneous transcriptions obtained from speech recognition (denoted as SD below). The retrieval results were expressed in terms of *non-interpolated average precision* [3,9].

4.2 Word- vs. Syllable-level Indexing Features

Table 2 shows the retrieval results of the HMM/Ngram-based approach on both the TDT-2 and TDT-3 collections. It can be found from the first three columns of Table 2 that, for the word-level indexing features, using unigram information alone achieved reasonable performance while including bigram information offered only limited improvements, if not degrading the retrieval performance. On the other hand, for the syllable-level indexing features (the last three columns), using unigram information alone seemed inadequate (Column 4), while including bigram information always helps (Columns 5 and 6). By comparing the best performance of the word- and syllable-level indexing features, the word-level indexing features outperformed the syllable-level indexing features in more cases, but the syllable-level indexing features performed the best for the real desired case, the erroneous speech transcriptions (SD) of the TDT-3 evaluation set. On the other hand, though the word error rates for both the TDT-2 and TDT-3 spoken document collections were higher than 35%, the performance for the SD cases were only slightly lower than those for the TD cases.

		Word-level			Syllable-level		
		Uni	Uni+Bi	Uni+Bi*	Uni	Uni+Bi	Uni+Bi*
TDT2	TD	**0.6327**	0.6069	0.5427	0.4698	0.5220	0.5697
(Dev.)	SD	0.5658	**0.5702**	0.4803	0.4411	0.5011	0.5305
TDT3	TD	**0.6569**	0.6542	0.6141	0.5343	0.5970	0.6544
(Eval.)	SD	0.6308	0.6361	0.5808	0.5177	0.5678	**0.6413**

Table 2: Retrieval results of the HMM/N-gram-based retrieval approach.

		Word-level		Syllable-level	
		S(N), *N*=1	*S(N)*, *N*=1~2	*S(N)*, *N*=1	*S(N)*, *N*=1~2
TDT2	TD	0.5548	**0.5623**	0.3412	0.5254
(Dev.)	SD	0.5122	**0.5225**	0.3306	0.5077
TDT3	TD	0.6505	**0.6531**	0.3963	0.6502
(Eval.)	SD	0.6216	0.6233	0.3708	**0.6353**

Table 3: Retrieval results of the vector space model retrieval approach.

4.3 Comparisons with Vector Space Model

The retrieval results of the vector space model approach are shown in Table 3, in which "*S(N)*, *N*=1" means using the single word or single syllable as the indexing terms and "*S(N)*, *N*=1~2" means using both the single word and the word-pair, or both the single syllable and the syllables-pair as the indexing terms. Several observations could be drawn from Table 3. First, similar to the HMM-based approach, the word-level features outperformed the syllable-level features in several cases, but the syllable-level features performed the best for the real desired case of SD for TDT-3. Second, unlike the HMM-based approach, using both the single word and the word-pair for indexing always outperformed that using the single word only. Third, using the single syllable only for indexing in the vector space model approach always gave significantly poorer performance than using the syllable unigram information only in the HMM/N-gram-based approach. Fourth, the HMM/N-gram-based approach was consistently better than the vector space model approach, and the difference is significantly larger for the TDT-2 development set from which the linear combination weights were trained.

5. MINIMUM CLASSIFICATION ERROR (MCE) TRAINING FOR THE HMM/N-GRAM-BASED RETRIEVAL APPROACH

The minimum classification error (MCE) training algorithm [10] widely used 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 desired relevant document D^*, we can define the classification error function as follows:

$$E(Q, D^*) = \frac{1}{|Q|}\left[-\log P\left(Q|D^* \text{ is } R\right) + \max_{D'} \log P\left(Q|D' \text{ is not } R\right)\right], \quad (9)$$

where D' is the irrelevant document that has the highest relevance score. This classification error function can be

transformed to a loss function ranging from 0 to 1 with the Sigmoid operator:

$$L(Q, D^*) = \frac{1}{1 + \exp(-\alpha E(Q, D^*) + \beta)}, \quad (10)$$

where α is used to control the slope of the function, and β is an offset factor (set to 0 here for simplicity). Equation (10) was then applied iteratively to update the linear combination weights m_i of the HMMs for the document D^*. For example, for the Type I HMM structure in Equation (2), the weight m_1 of the document D^* can be iteratively adjusted by using the following two equations:

$$m_1(i+1) = \frac{m_1(i) \cdot e^{-\nabla_{D^*, m_1}(i)}}{m_1(i) \cdot e^{-\nabla_{D^*, m_1}(i)} + m_2(i) \cdot e^{-\nabla_{D^*, m_2}(i)}}, \quad (11)$$

$$\nabla_{D^*, m_1}(i) = -\varepsilon(i) \times \alpha \times L(Q, D^*) \times [1 - L(Q, D^*)] \times \left[-m_1(i) + \frac{1}{|Q|} \sum_{q_n \in Q} \frac{m_1(i) P(q_n | D^*)}{m_1(i) P(q_n | D^*) + m_2(i) P(q_n | Corpus)} \right], \quad (12)$$

where $\varepsilon(i)$ is a constant factor used in each iteration i. The training query exemplars used in Section 4 for the EM training were again used here for the MCE training. Only very preliminary tests were performed. For the word-level indexing features only the Type I model (Uni) was tested, while for the syllable-level indexing features only the Type III model (Uni+Bi*) was tested, both with TDT-2 only. Both cases gave reasonable performance in the previous experiments in Table 2. The retrieval results with the MCE training are shown in the first two columns of Table 4. It can be found that, with the syllable-level indexing features the result was significantly improved from 0.5697 (Table 2) to 0.6743 in the TD case and from 0.5305 (Table 2) to 0.6224 in the SD case. Similar improvements were achieved with the word-level indexing features though not as significant. It is also very interesting to note that here the syllable-level indexing features outperformed the word-level indexing features for both TD and SD cases. Since the weights of the HMMs of documents were no longer tied together, those obtained for TDT-2 could not be used for TDT-3.

6. INFORMATION FUSION FOR THE HMM/N-GRAM-BASED RETRIEVAL APPROACH

The word-level indexing features possess more semantic information than the syllable-level features. On the other hand, the syllable-level indexing features provide a more robust relevance measure between queries and documents when dealing with such problems as those arising from the flexible wording structure in Mandarin Chinese and the speech recognition errors in spoken documents [6-8]. It was believed that a proper fusion of the syllable- and word-level information would be useful for the retrieval task studied here. As a result, the fusion of the best approaches using the word- and syllable-level indexing features achieved in Section 5 using the following equation was tested:

$$R(Q, D) = w_w R_w(Q, D) + w_s R_s(Q, D), \quad (13)$$

which is simply the weighted sum of the relevance scores obtained with the word- and the syllable-level indexing features. The results on the TDT-2 collection are shown in the right column of Table 4. As compared to the results of using either the word-level or syllable-level information alone (in the first two columns of Table 4), the fusion is indeed helpful for both the TD and SD cases.

		Word-level	**Syllable-level**	**Fusion**
		Uni	Uni+Bi*	
TDT2	TD	0.6459	**0.6743**	**0.6983**
(Dev.)	SD	0.5810	**0.6224**	**0.6326**

Table 4: Retrieval results of the HMM/N-gram-based retrieval approach after the MCE-training was applied.

7. CONCLUDING REMARKS

In this paper, we presented an HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval. We have intensively investigated its underlying characteristics and structures, verified its retrieval capabilities by using different levels of indexing features, and compared it with the vector space model approach. The minimum classification error training has been introduced in the training phase to improve the discrimination among the HMMs for the documents. We found that given a set of training query exemplars, the retrieval performance could be significantly improved, which means that an HMM/N-gram-based retrieval system can be incrementally improved through use. In addition, the information fusion of different levels of indexing features has been shown useful. Many other advanced retrieval techniques such as the blind relevance feedback and the query expansion by term associations have been successfully applied to this HMM/N-gram-based approach to further improve the retrieval performance as well. Nevertheless, they were not included in this paper due to the space limit.

8. REFERENCES

[1] Jelinek Frederick, Statistical Methods for Speech Recognition. The MIT Press 1999.

[2] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, "Speech and Language Techniques for Audio Indexing and Retrieval," *Proc. IEEE*, Vol. 88, No. 8, Aug. 2000.

[3] David R. H. Miller, T. Leek, and R. Schwartz, "A Hidden Markov Model Information Retrieval System," in *Proc.* ACM SIGIR 1999.

[4] F. Song and W. B. Croft, "A General Language Model for Information Retrieval," in *Proc.* CIKM 1999.

[5] P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 Broadcast News Transcription System for Mandarin," in *Proc. of the DARPA Broadcast News Workshop*, 1999.

[6] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan Using Syllable-Level Statistical Characteristics," in *Proc.* ICASSP 2000.

[7] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of Mandarin Broadcast News Using Spoken Queries," in *Proc.* ICSLP 2000.

[8] H. M. Wang, H. Meng, P. Schone, B. Chen and W. K. Lo, "Multi-Scale Audio Indexing for Translingual Spoken Document Retrieval, " in *Proc.* ICASSP 2001.

[9] D. Harman, Overview of the Fourth Text Retrieval Conference (TREC-4). 1995.

[10] B. H. Juang, W. Chou, and C. H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 5, No. 3, May 1997.