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

Berlin Chen\(^1\), Hsin-min Wang\(^1\), and Lin-shan Lee\(^1\)^2

\(^1\) Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China
\(^2\) 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 \( p(D \mid Q) \); 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 \( p(Q \mid D \cup I) \), 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.
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., $r(D \mid Q)$, which can be transformed to the following equation by applying Bayes’ theorem [1]:

$$r(D \mid Q) = \frac{p(Q \mid D) r(D)}{p(Q)}$$  \hspace{1cm} (1)

where $p(Q \mid D)$ is the probability of the query $Q$ being posed under the condition that document $D$ is relevant, $r(D)$ is the prior probability that document $D$ is relevant, and $p(Q)$ is the prior probability of query $Q$ being posed. $r(D)$ in (1) can be eliminated because it is identical for all documents. Furthermore, because there is no general way to estimate the prior probability $r(D)$, we can simply set it to unity for simplicity and approximate the probability $r(D \mid Q)$ by the probability $r(Q \mid D)$ for the problem studied here.

In this research, the query $Q$ is treated as a sequence of input observations (or indexing terms), $Q = q_1, q_2, \ldots, q_n$, where each $q_i$ 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 a single-state discrete HMM as shown in Figure 1. The observation probabilities for this HMM are modeled by a single-state discrete HMM as shown in Figure 1.

$$p(Q \mid D) = \frac{1}{N} \sum_{n=1}^{N} r(q_n \mid D)$$  \hspace{1cm} (2)

where $r(q_n \mid D)$ is the unigram probability of a specific indexing term $q_n$ within the document $D$ and $p(q_n \mid Corpus)$ is the unigram probability of a specific indexing term sequence $q_1, \ldots, q_n$ 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., $p(q_n \mid Corpus)$ or/and $p(q_{n-1}q_n \mid Corpus)$, 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. $\sum m_i = 1$ 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 $m_4$ of Equation (2) can be estimated using the following equation:

$$m_i = \frac{\sum_{q_i \in \text{TrainSet}_{Q_{\ell}}} \frac{1}{|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|} \sum_{q_{i-1} \in \text{TrainSet}_{Q_{\ell}}} \frac{1}{|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|} p(q_i \mid q_{i-1}) p(q_{i-1} \mid \text{Corpus})}{\sum_{q_i \in \text{TrainSet}_{Q_{\ell}}} \frac{1}{|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|} \sum_{q_{i-1} \in \text{TrainSet}_{Q_{\ell}}} \frac{1}{|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|} p(q_i \mid q_{i-1})}$$  \hspace{1cm} (5)

where $\text{TrainSet}_{Q_{\ell}}$ is the set of training query exemplars, $|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|$ is the set of documents that are relevant to a specific training query exemplar $Q_{\ell}$, $|\text{Doc}|$ is the length of the query $Q$, and $|\text{Doc}_{\ell \in \text{TrainSet}_{Q_{\ell}}}|$ 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 $d$, 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 $f(t)$ of a feature vector $d$ for a document $D$ is associated with the statistics of a specific indexing term $t$:

$$f(t) = (1 + \ln(c(t))) \ln(|\text{Doc}|),$$  \hspace{1cm} (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
4.2 Word- vs. Syllable-level Indexing Features

Table 2 shows the retrieval results of the HMM/N-gram-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.

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.

4.4 Experiment Setup

All the three types of HMM structures specified by Equations (2)-(4) were tested. The probabilities of $p_{y_{n}x_{n}}$ and $p_{y_{n}x_{n},Corpus}$ in these equations were estimated using a general text corpus consisting of 40 million Chinese characters. The weights $m_{n}$ 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(subword)-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].

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:

$$E(q,D^*) = \frac{1}{|q|} \left[ -\log P[D^* | q] + \max_{D \neq D^*} \log P[D | q] \right].$$

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:
\[ L(Q, D^*) = \frac{1}{1 + \exp(-\alpha E(Q, D^*) + \beta)}, \]  
(10)
where \( \alpha \) is used to control the slope of the function, and \( \beta \) 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:
\[
\begin{aligned}
    m_1(t + 1) &= m_1(t) \cdot \frac{e^{-\epsilon(i) \cdot \alpha}}{\sum_{l=1}^{n} \log p(l|Q,D^*) + \beta}, \\
    \epsilon(i) &= \frac{1}{\sum_{l=1}^{n} \log p(l|Q,D^*) + \beta} \cdot \sum_{l=1}^{n} m_l(t) \cdot \log p(l|Q,D^*) 
\end{aligned}
\]
(11)
where \( \epsilon(i) \) 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 word 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:
\[ R(Q,D) = w_w R_w(Q,D) + w_s R_s(Q,D), \]
(13)
which is simply the weighted sum of the relevance scores obtained with the word- and 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.

### 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


<table>
<thead>
<tr>
<th></th>
<th>Word-level</th>
<th>Syllable-level</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT2</td>
<td>0.6459</td>
<td>0.6743</td>
<td>0.6983</td>
</tr>
<tr>
<td>(Dev.)</td>
<td>0.5810</td>
<td>0.6224</td>
<td>0.6326</td>
</tr>
</tbody>
</table>

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