IMPROVED CHINESE SPOKEN DOCUMENT RETRIEVAL WITH HYBRID MODELING AND DATA-DRIVEN INDEXING FEATURES

Chun-Jen Wang, Berlin Chen, and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

stanley@speech.ee.ntu.edu.tw, {berlin, lsl}@iis.sinica.edu.tw

ABSTRACT

Different models retrieve the documents based on different approaches of extracting the underlying content. Different levels of indexing features also offer different functionalities and discriminabilities when retrieving the documents. In this paper, we present results for Chinese spoken document retrieval with hybrid models to integrate the knowledge obtainable from three basic retrieval models, namely, the standard vector space model (VSM), the hidden Markov model (HMM), and the latent semantic indexing (LSI) model. The characteristics of retrieval performance using both word-level and syllable-level indexing features were extensively explored. In addition, a data-driven approach to derive variable-length indexing features is also presented. Very satisfactory performance can be achieved with these data-driven features while retaining very compact feature set size. Experiments showed that this approach has the potential to identify domain-specific terminologies or newly-generated phrases. It is therefore very useful not only in Chinese document retrieval, but also in detecting out of vocabulary (OOV) words in Chinese. Very encouraging results were obtained when the hybrid models were used with the data-driven indexing features as well.

1. INTRODUCTION

With the explosive increase of various types of data over the Internet, large repositories of information have become available to the public. Intelligent and efficient retrieval techniques have been developed to provide the users with easy access to all kinds of documents. As the speech recognition techniques evolved in the past decades, extensive efforts have been made to use speech recognition technologies to retrieve spoken documents. The TREC evaluation [1] is one example for such works. Despite all these developments, efficient retrieval of spoken documents remains a very challenging and attractive research topic.

Retrieval approaches nowadays in general adopt one out of the two matching strategies to determine the degree of relevance for a document with respect to a query, namely, literal term matching and concept matching. The classical vector space model (VSM) and probability-based model are primarily based on literal term matching. VSM is simple and fast while achieving satisfactory performance. This is why it has been popularly used. Its limited capabilities in utilizing the local context ordering information is well known though. The probability-based approach, on the other hand, attempts to handle the retrieval problem within a statistical framework. The language modeling approach [2] and the hidden Markov model (HMM) [3] are good examples of this category. HMM was shown to outperform the standard VSM on both the TDT-2 and TDT-3 Chinese collections [4], but some popular techniques in information retrieval, such as the relevance feedback and automatic query expansion, seem less convenient to be integrated into the framework [5].

Most approaches based on literal term matching are kind of limited due to the problem of word usage diversity [6], or the so-called “vocabulary mismatch” problem [7], i.e., many relevant documents can’t be retrieved even though they are really “about” the given query but are using a different set of words. Concept matching approaches, on the other hand, is based on the conceptual topics of the documents. Latent Semantic Indexing (LSI) is one example. It transforms the (high dimensional) representation of documents and terms to the so-called latent semantic space. The similarities among the documents can then be estimated in the reduced space. This approach was shown to be very promising [8,9], especially at higher levels of recall. Because literal term matching and concept matching approaches seem to be complementing each other, in this paper we’ll present hybrid models to integrate the nice features of these two matching approaches.

Furthermore, word- and subword-based indexing features have been exploited extensively for spoken document retrieval. Experiments indicated that word-based features are more important in English [10], while syllable-level (subword-based) information is highly discriminative for Chinese due to the monosyllabic structure of the language [11,12]. Currently most syllable-level features adopted for Chinese spoken document retrieval are predefined fixed-length syllable segments. They performed very well at moderate length (for example, N<5) [11], but the total number of possible segments is huge, making it difficult for real-world applications. This is why in this paper we developed a data-driven indexing feature selection approach to derive variable-length syllable segments as features. It was found that the features derived in this way are very often semantically meaningful and therefore can capture more intrinsic concepts during retrieval.
2. EXPERIMENTAL SETUP

We used two Topic Detection and Tracking (TDT) collections for this work, TDT-2 as the development set while TDT-3 as the evaluation set. In both cases the Chinese news stories (in text form) from Xinhua News Agency were used as queries, and the Mandarin news stories (in audio form) from Voice of America news broadcast as the spoken documents. All the experiments reported in this paper involve the use of an entire Chinese newspaper story (text) as a query, to retrieve relevant Chinese broadcast news stories (audio) in the document collection, or the so-called query-by-example. The Chinese word transcriptions were given by the Dragon large-vocabulary continuous speech recognizer [13] for Mandarin audio collections (TDT-2 and TDT-3). We spot-checked a fraction of the TDT-2 (46 hours) and the TDT-3 (76 hours), and obtained word error rates of 35.38% and 36.97% respectively [4].

3. A BRIEF REVIEW OF THE BASIC INFORMATION RETRIEVAL MODELS

In the following, the three basic retrieval models used in this paper are briefly reviewed.

3.1. Vector Space Model (VSM)

In this approach, every document \( D \) and query \( Q \) is represented as a feature vector. Each component in the vector, \( g(t) \), is associated with the statistics of a specific indexing term \( t \),

\[
g(t) = \frac{1 + \ln(c(t))}{\ln(N/T)}
\]

where \( c(t) \) is the occurrence count of the term \( t \) within the document \( D \) or query \( Q \), and \( \ln(N/T) \) is the inverse document frequency (IDF). The popular cosine measure is used to estimate the query-document relevance.

3.2. Hidden Markov Model (HMM)

In this model, a documents \( D \) is ranked according to the probability that \( D \) is relevant, conditioned on the fact that the query \( Q \) is produced, which can be further transformed as in the following by Bayes’ theorem:

\[
p(D | Q) = \frac{p(Q | D) p(D)}{p(Q)}
\]

Because \( P(Q) \) is identical for all documents, and it is difficult to estimate the probability of \( P(D | R) \), the remaining term \( P(Q | D) \) is used to rank the documents [3]. The query \( Q \) is treated as a sequence of input observations (or indexing terms), \( Q = q_1, q_2, \ldots, q_N \), while each document \( D \) is modeled as a single-state discrete HMM. We adopted the same HMM as in the previous work [4] in this paper, where the document-specific and general distributions of unigram and/or bigram probabilities were used. The weights for combining these probabilities were estimated using the expectation-maximization (EM) algorithm [4].

3.3. Latent Semantic Indexing (LSI)

This model starts with a term/document matrix, and singular value decomposition (SVD) is applied to reduce the dimension and construct the latent semantic space, in which the original documents and terms is properly represented, and queries or documents which are not part of the original matrix can be

Non-Interpolated Average Precision

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Word</th>
<th>Syllable</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT2</td>
<td>TD</td>
<td>0.555</td>
<td>0.538</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.512</td>
<td>0.518</td>
<td>0.531</td>
</tr>
<tr>
<td>TDT3</td>
<td>TD</td>
<td>0.651</td>
<td>0.668</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.621</td>
<td>0.652</td>
<td>0.666</td>
</tr>
</tbody>
</table>

Table 1: Baseline retrieval results for the vector space model (VSM)

Non-Interpolated Average Precision

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Word</th>
<th>Syllable</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT2</td>
<td>TD</td>
<td>0.633</td>
<td>0.572</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.566</td>
<td>0.531</td>
<td>0.573</td>
</tr>
<tr>
<td>TDT3</td>
<td>TD</td>
<td>0.657</td>
<td>0.656</td>
<td>0.681</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.631</td>
<td>0.643</td>
<td>0.673</td>
</tr>
</tbody>
</table>

Table 2: Baseline retrieval results for the hidden Markov model (HMM)

folded-in by matrix multiplication. The rationale is that terms which occur in similar context will be near each other in the latent semantic space even if they never co-occur in the same document. The degree of relevance between documents and queries are then estimated by computing the cosine measure in the latent semantic space [8,9,14].

4. BASELINE EXPERIMENTAL RESULTS

In this paper, the completely correct manual transcriptions of the spoken documents (denoted as TD, text documents) were also used in the retrieval experiments for reference, as compared to the erroneous transcriptions obtained from speech recognition (denoted as SD, spoken documents). All the three basic retrieval models, the vector space model (VSM), hidden Markov model (HMM), and latent semantic indexing (LSI) model were explored with both word- and syllable-level indexing features as well as the fusion of them. The retrieval results were presented in terms of non-interpolated average precision [1]. The baseline results for the three basic models, VSM, HMM and LSI are listed in Tables 1, 2, 3 respectively. For the VSM model, as shown in Table 1, syllable information is very useful especially for spoken documents case (SD). For TDT-3, the syllable information even outperforms the word information for the text documents case (TD). Also, the fusion of the two is always better. For the HMM model as shown in Table 2, word-level indexing features perform better in TDT-2 collection, while syllable-level indexing features outperform in TDT-3 collection in the spoken documents case (SD). In all the cases the fusion of the two gives better results. It is also clear that the HMM approach in Table 2 is consistently better than the VSM approach in Table 1. For the retrieval results of latent
5. HYBRID MODELS

Since LSI is based on concept matching while VSM and HMM on literal term matching, integrating LSI with either HMM or VSM seems to be natural and advantageous. Experimental results verified the above inference. Tables 4 and 5 show the performance of two such hybrid models, hybrid model 1 (VSM plus LSI) and 2 (HMM plus LSI) respectively. Several observations can be drawn from these two tables. First, the retrieval effectiveness of the component models in these hybrid models is apparently additive, i.e., in all cases the integrated model is better than any individual models. Second, HMM and LSI (for the hybrid model 2 in Table 5) are especially complementary. In fact, linearly interpolating these two models improves the precision at all levels of recall. There can be several reasons for this. LSI estimates the global associations, but unavoidably loses the word ordering information. On the contrary, HMM captures the local context constraints by its n-gram paradigm, and it also provides a framework for incorporating several knowledge sources. Finally, fusion of word- and syllable-level information again improves the performance for these hybrid models.

6. DATA-DRIVEN INDEXING FEATURES

All the previous results used predefined fixed-length indexing features, in terms of either words or syllables. However, many of such fixed-length features even never occur in any documents. There is no way to tell how much indexing information each of these features actually carry either. But all these features very often make the dimension of feature space prohibitively large for real-world applications. This is why the data-driven indexing feature concept proposed here makes sense. We let the data tell which indexing features are useful, while deleting all those which are not helpful. The measure we used for feature selection is the geometrical average of the forward and backward bigram [16] of adjacent terms \((\omega_1, \omega_2)\).

\[
FB(\omega_1, \omega_2) = \frac{\sqrt{P_j(\omega_1 | \omega_2)P_k(\omega_2 | \omega_1)}}{P(W_{ij} = \omega_1, W_{ji} = \omega_2)}
\]

where

\[
P_j(\omega_1 | \omega_2) = \frac{P(W_{ij} = \omega_1, W_{ji} = \omega_2)}{P(W_{ij} = \omega_1)}
\]

This measure was used to select syllable segments as indexing features in the experiments below. We started with a feature set consisting of all single syllables only, and applied the above measure iteratively to find all useful syllable segments. In each iteration, the term pairs scored above a threshold became a new term and all instances of these pairs in the corpus were replaced by the new terms. We empirically chose an optimal threshold based on TDT-2 collection, and evaluated on TDT-3 collection using this optimal threshold. Preliminary experimental results
In this paper, we present hybrid retrieval approaches to incorporate different advantages of the three basic models with word- and syllable-level indexing features on Chinese spoken document retrieval. Experimental results supported the integration concept of the hybrid models. A statistical approach to derive data-driven indexing features is also developed. Almost equal retrieval performance can be achieved with a very compact syllable-level indexing feature set, which is actually significantly better than using single words. Further investigations are under progress.

### 8. REFERENCES


