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

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 newswire 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,

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

where $c(t)$ is the occurrence count of the term t within the document D or query Q, and $\ln(N / N_t)$ 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:

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

Because $P(Q)$ is identical for all documents, and it is difficult to estimate the probability of $P(D \text{ is } R)$, the remaining term $P(Q|D \text{ is } R)$ is used to rank the documents [3]. The query Q is treated as a sequence of input observations (or indexing terms), $Q=q_1q_2..q_n..q_N$, 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 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].

Non-Interpolated Average Precision		Word	Syllable	Fusion
TDT2	TD	0.555	0.538	0.577
	SD	0.512	0.518	0.531
TDT3	TD	0.651	0.668	0.676
	SD	0.621	0.652	0.666

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

Non-Interpolated Average Precision		Word	Syllable	Fusion
TDT2	TD	0.633	0.572	0.635
	SD	0.566	0.531	0.573
TDT3	TD	0.657	0.656	0.681
	SD	0.631	0.643	0.673

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

Non-Interpolated Average Precision		Word	Syllable	Fusion
TDT2	TD	0.551	0.512	0.566
	SD	0.531	0.496	0.543
TDT3	TD	0.644	0.628	0.665
	SD	0.639	0.544	0.645

Table 3: Baseline retrieval results for the latent semantic indexing (LSI) approach

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

semantic indexing (LSI) approach in Table 3, the syllable-level information used is based on syllable pairs and the word-level information used is based on single words. The number of the former used in the experiments is on the order of 168,100 (ignoring the tones), and that of the latter is on the order of 50,000. Therefore, the size of the original term/document matrix is huge. The number of factors (number of singular triplets reserved when constructing the *latent space*), *k*, is thus critical [6]. In the experiments, the value of *k* ranged from 10 to 1000. Those reported in Table 3 are the best results for best values of *k*. From Table 3 we can see LSI achieves either comparable or better performance than VSM. Also, LSI is specially robust in the presence of recognition error, so the results for spoken documents case (SD) is usually not too far from those for text documents case (TD). This is because LSI does not rely on literal term matching, therefore especially useful when the document is noisy, as in the speech recognition case here. Furthermore, fusion of word- and syllable-level information always provides better performance. This is consistent for all the three models. In addition to the non-interpolated average precision in Tables 1, 2 and 3, recall-precision characteristics provide some more information, and are therefore plotted as well for the three baseline models, as shown in Figure 1 at 11 standard recall levels [15] obtained for TDT-3 SD collection. It is clear that LSI performs significantly better than the other two models at high levels of recall, while HMM retrieves more relevant documents at low levels of recall. Such results strongly suggest that adequately interpolating these models may further improve the retrieval performance, as discussed below.

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

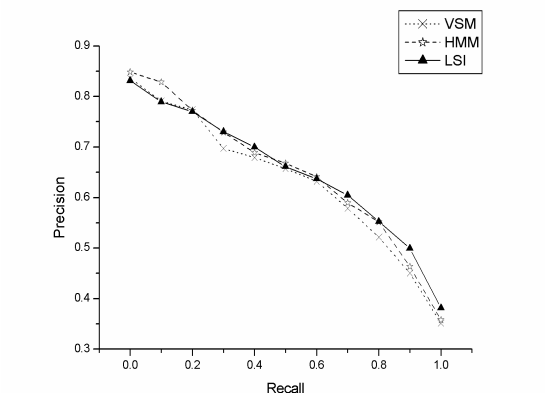

Figure 1: Recall-Precision curve at 11 standard recall levels evaluated on three basic retrieval models

Non-Interpolated Average Precision		Word	Syllable	Fusion
TDT2	TD	0.582	0.545	0.582
	SD	0.533	0.536	0.552
TDT3	TD	0.670	0.682	0.690
	SD	0.656	0.663	0.679

Table 4: Retrieval results of the *Hybrid model 1*(VSM plus LSI)

Non-Interpolated Average Precision		Word	Syllable	Fusion
TDT2	TD	0.634	0.589	0.640
	SD	0.571	0.553	0.581
TDT3	TD	0.676	0.695	0.707
	SD	0.670	0.673	0.698

Table 5: Retrieval results of the *Hybrid model 2* (HMM plus LSI)

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 (ω_i,ω_j)

$$FB(\omega_i,\omega_j) = \sqrt{P_f(\omega_j \mid \omega_i) P_b(\omega_i \mid \omega_j)} \quad (2)$$

where

$$P_f(\omega_j \mid \omega_i) = \frac{P(W_{t+1} = \omega_j, W_t = \omega_i)}{P(W_t = \omega_i)} \quad (3)$$

$$P_b(\omega_i \mid \omega_j) = \frac{P(W_{t+1} = \omega_j, W_t = \omega_i)}{P(W_{t+1} = \omega_j)} \quad (4)$$

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

Non-Interpolated Average Precision		VSM	HMM	LSI
TDT2	TD	0.577	0.588	0.582
	SD	0.531	0.567	0.538
TDT3	TD	0.662	0.677	0.664
	SD	0.651	0.651	0.660

Table 6: Retrieval results of the three basic models using data-driven indexing features

Non-Interpolated Average Precision		*Hybrid model 1* (VSM plus LSI)	*Hybrid model 2* (HMM plus LSI)
TDT2	TD	0.586	0.595
	SD	0.533	0.563
TDT3	TD	0.668	0.690
	SD	0.666	0.671

Table 7: Retrieval results of hybrid models using data-driven indexing features

are shown in Table 6 for the three basic models: VSM, HMM and LSI. As can be found in Table 6, using the very compact data-driven indexing features on the three basic models produced results at least comparable to those produced by fixed-length indexing features, but with much smaller size of feature set. In fact, in almost all cases the data-driven features outperformed the predefined single-word or syllable-pair indexing features. Take the fourth row of Table 6 for TDT-3 SD collection as an example, the size of the data-driven feature set is only 5,008 (including syllable segments with length ranging from 1 to 5), which is far less than that of the syllable-pair features, 168,100, and is also only a tenth of that of single-word features, 51,253, but giving comparable results. This is quite amazing.

It is interesting to note that we found many of the 5008 data-driven syllable segments mentioned above represent important keywords for retrieval including proper nouns, such as personal names and organization names, or a part of such keywords or proper nouns. This implies that this automatic feature selection method can also detect newly generated phrases and domain-specific terminologies. The new feature vocabulary also includes terms composed of three to five single syllables. These higher order terms almost directly translate to exact semantic meanings. It is well known that for Chinese words with three or more syllables, there is almost a one-to-one correspondence between the words and the syllable segments. Fixed-length indexing mechanism can capture this kind of higher order information only by increasing the length of segments, which unavoidably explodes the size of the indexing feature set. Therefore the data-driven approach here can derive information-abundant indexing terms very efficiently while discarding non-informative terms. Finally, these data-driven indexing features were applied to the very good hybrid models 1 and 2, as presented previously. The results are shown in Table 7. We can see that even better performance was achieved with very small features sizes.

7. CONCLUSIONS

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

[1] D. Harman, "Overview of the Ninth Text Retrieval Conference (TREC-9)", 2000. Available at http://trec.nist.gov/pubs/trec9/papers/overview_9.pdf.

[2] J.M. Ponte, W.B. Croft, "A Language Modeling Approach to Information Retrieval", in Proc. ACM SIGIR, 1998.

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

[4] B. Chen, H.M. Wang, L.S. Lee, "An HMM-based Language Modeling Approach for Mandarin Spoken Document Retrieval", in Proc. European Conf. On Speech Communication and Technology (EUROSPEECH), 2001.

[5] V. Lavrenko, W.B. Croft. "Relevance-Based Language Models", in Proc. ACM SIGIR, 2001.

[6] M.W. Berry, S.T. Dumais, G.W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval", SIAM Review, 1995.

[7] A. Singhal, F. Pereira, "Document Expansion for Speech Retrieval", in Proc. ACM SIGIR, 1999.

[8] S. Deerwester, S.T. Dumais, R. Harshman, "Indexing by Latent Semantic Analysis", in Journal of the Society of Information Science, 1990.

[9] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R. Harshman, L.A. Streeter, K.E. Lochbaum, "Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure", in Proc. ACM SIGIR, 1988.

[10] Kenney Ng, "Information Fusion for Spoken Document Retrieval", in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 2000.

[11] B. Chen, H.M. Wang, L.S. Lee, "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan Using Syllable-Level Statistical Characteristics", in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 2000.

[12] B. Chen, H.M. Wang, L.S. Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", to appear in IEEE Trans. On Speech and Audio Processing, 2002.

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

[14] J.R. Bellegarda, "A Multispan Language Modeling Framework for Large Vocabulary Speech Recognition", in IEEE Trans. on Speech and Audio Processing, 1998.

[15] R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley, 1999.

[16] G. Saon, M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition", in IEEE Trans. on Speech and Audio Processing, 2001.