MULTI-SCALE AND MULTI-MODEL INTEGRATION FOR IMPROVED PERFORMANCE IN CHINESE SPOKEN DOCUMENT RETRIEVAL

Wai-Kit LO[1], *Helen MENG*[2], *and P. C. CHING*[1]

[1]Department of Electronic Engineering,
The Chinese University of Hong Kong,
Hong Kong SAR, China
{email: wklo1@ee.cuhk.edu.hk, pcching@ee.cuhk.edu.hk}

[2]Department of Systems Engineering
and Engineering Management,
The Chinese University of Hong Kong,
Hong Kong SAR, China
{email: hmmeng@se.cuhk.edu.hk}

ABSTRACT

This paper describes our attempt to combine the relative merits of different indexing units (scales) and different retrieval models to improve performance in Chinese spoken document retrieval. Our study includes indexing units from three scales: words, character bigrams and syllable bigrams. We also include two different retrieval models: the HMM-based model and the vector space model (VSM). Our retrieval task is based on the TDT-2 Mandarin collection - news text is used to retrieve relevant Mandarin audio. We experimented with different scales and retrieval models. The HMM-based model retrieves better at the word scale (mAP=0.566). For the VSM, better performance is obtained at the character bigram scale (mAP=0.562). We proceeded with a series of integration experiments where the ranked retrieval lists from different runs are combined by rank-based re-scoring. The best retrieval performance (mAP=0.591) is achieved when we integrate the HMM-word and VSM-character configurations. These results suggest that retrieval based on different scales and different models capture different kinds of knowledge, which can be integrated to improve retrieval performance.

1. INTRODUCTION

Spoken document retrieval is an essential technology for accessing large archives of audio or audio-visual materials. Spoken document retrieval enables users to search for personally relevant information from a diversity of multimedia information sources. A popular approach towards spoken document retrieval is to apply automatic speech recognition to the audio and then to perform information retrieval on the recognizer's transcriptions. Much previous work has been conducted in efforts like TREC[1] and TDT[2].

Indexing units at different scales capture different kind of knowledge. Common scales of indexing units for Chinese spoken document retrieval include words and subwords, where subwords include both characters and syllables. While words are meaningful and contain lexical knowledge, characters and syllables can provide full textual and phonological coverage for Chinese textual and spoken documents. Therefore, retrieval at word scale gives better precision while retrieval using characters and syllables are robust to word segmentation and recognition errors.

Furthermore, different retrieval models have different strengths and characteristics. For instance, HMM-based [1, 2, 3, 4, 5] retrieval models are formulated as finding the probability of generating the given query by a document. This probability is estimated using the statistics from document as well as the language in general. The vector space model (VSM) [6] represent queries and documents as vectors, the retrieval problem is formulated as finding the similarity between the document and query vectors. In this work, the cosine measure is used as the similarity measure.

Different indexing scales and different retrieval models capture different knowledge sources for retrieval. In this work, we propose to use multi-scale and multi-model integration for Chinese spoken document retrieval in order to improve retrieval performance.

2. MULTI-SCALE INDEXING UNITS

Multi-scale retrieval refers to the use of words and subwords as indexing units [7]. As mentioned above, the various scales of indexing units can capture different kinds of linguistics information as well as robustness to errors.

2.1. Words

The word unit contains lexical information that can help improve precision during retrieval. The word unit is commonly used in information retrieval systems for most European languages. However, there is no explicit word delimiter in Chinese, automatic word segmentation is required to identify words from the character sequences. Much ambiguity exists in segmenting a sequence of Chinese character into words. This ambiguity can affect retrieval performance. In our retrieval task, the textual queries are segmented for words. While for the spoken documents, words are obtained from the output of LVCSR and errors may be introduced during recognition.

這一晚會如常舉行	**Character sequence**
Different segmentations	***Different meanings***
這一 晚會 如常 舉行	**This banquet will be held as usual**
這一晚 會 如常 舉行	**Tonight an event will be held as usual**
這一 晚會 如 常舉行	**If this banquet is held very often**

2.2. Subwords

Subword indexing units include the character and the syllable. Usually, overlapping n-grams of these units are used. They can improve the robustness of retrieval (to be explained later). Subword bigrams are commonly used in Chinese information retrieval. They can also capture certain degree of lexical information too. It

[1] Text Retrieval Conference, http://trec.nist.gov

[2] Topic Detection and Tracking, http://www.itl.nist.gov/iad/894.01/tests/tdt

is because in the Chinese language, most of the words are two characters in length.

Subword bigram can circumvent the errors due to ambiguity in automatic word segmentation. For a given Chinese character sequence with multiple segmentations (as illustrated above), retrieval using word units may result in reduction of performance. If subword bigram is used, the retrieval will not be affected by this ambiguity in segmentation.

In spoken document retrieval, subword bigram are also useful because it is robust to errors due to recognition. For a given 4-character word $\{C_1\ C_2\ C_3\ C_4\}$, supposed that a single error is made and turns C_4 into E_4, the word will become $\{C_1\ C_2\ C_3\ E_4\}$. Therefore, matching at word level will fail. However, if overlapping subword bigram is used, there are two correct bigrams preserved (C_1C_2, C_2C_3) to offer match for retrieval.

Moreover, every Chinese character is pronounced as a single syllable. The mapping between Chinese characters and syllables is many-to-many. There are also a large number of Chinese homophones. These homophones can cause character confusions during automatic transcription of the spoken documents by LVCSR. Incorrect homophonic characters may be returned. By indexing based on syllable scales, this problem can be circumvented.

3. RETRIEVAL MODELS

In this work, the HMM-based model and the VSM are used in the retrieval experiments. Their configurations are given in the following sections.

3.1. HMM-based model

For the HMM-based retrieval experiments, the HMM are augmented by a general language model [1, 2, 3, 4, 5]. The equation for this model is shown in Eq. (1).

$$p(Q|D_i) = \prod_{q_j \in Q} \left[w_{doc} \cdot p(q_j|D_i) + w_{glm} \cdot p_{glm}(q_j) \right] \qquad (1)$$

where w_{doc} and w_{glm} are the weights for the document model and the general language model respectively; and $p_{glm}(q_j)$ is the probability for the term q_j generated by the general language model.

In this model, the probability estimates are obtained using maximum likelihood estimation. The formula for the document model $p(q_j|D_i)$ and general language model $p_{glm}(q_j)$ are shown in Eq. (2) and (3) respectively.

$$\begin{aligned} p(q_j|D_i) &= \frac{Count(q_j\ in\ D_i)}{Count(all\ terms\ in\ D_i)} && (2) \\ p_{glm}(q_j) &= \frac{Count(q_j\ in\ collection)}{Count(all\ terms\ in\ collection)} && (3) \end{aligned}$$

3.2. Vector space model

For the VSM experiments, the queries and documents are transformed into vector representations by indexing.

$$\begin{aligned} \overrightarrow{q} &= [q_1\ q_2\ \cdots\ q_n] && (4) \\ \overrightarrow{d} &= [d_1\ d_2\ \cdots\ d_n] && (5) \end{aligned}$$

where q_i and d_i are the weights for the i^{th} term and n is the dimension of the vector representations.

In this work, the term weighting equations for query and document are defined as shown in Eq. (6) and (7) respectively.

$$\begin{aligned} q_i &= (\log(TF_{q_i}) + 1) \cdot \log\left(\frac{N+1}{n_i}\right) && (6) \\ d_i &= \log(TF_{d_i}) + 1 && (7) \end{aligned}$$

where TF is the term frequency, N is the total number of documents in the collection and n_i is the number of documents containing the i^{th} term.

When retrieval is performed, the document vectors are compared to the query vector using the cosine similarity measure as defined in Eq. (8).

$$Score\left(\overrightarrow{q}, \overrightarrow{d}\right) = \frac{\overrightarrow{q} \cdot \overrightarrow{d}}{\|\overrightarrow{q}\| \cdot \|\overrightarrow{d}\|} \qquad (8)$$

4. RANK-BASED INTEGRATION

Integration of multiple scales of units has been demonstrated to be beneficial to retrieval performance [8, 9, 10, 11]. In [8], the retrieval scores from different subword n-grams are integrated for a monolingual English task. In [9], this integration is applied to a cross-language English-Mandarin spoken document retrieval task. In [11], similar approach was applied to a monolingual Mandarin retrieval task.

Rank-based re-scoring is a simple and efficient integration approach. Ranked retrieval lists are integrated together as defined in Eq. (9).

$$Score(Q_i, D_j) = \frac{1}{\sum_{k \in K} Rank_k(Q_i, D_j)} \qquad (9)$$

where $Score(Q_i, D_j)$ is the integrated similarity score between query Q_i and document D_j; K are the set of retrieval runs to be integrated; $Rank_k(Q_i, D_j)$ is the rank of document D_j from the k^{th} runs when retrieving with query Q_i.

Rank-based integration is useful in many situations. While score-based approach is popular [8, 9, 10, 11], rank-based integration has its own merit. In case the integration is among different retrieval engines, the dynamic ranges of the retrieval scores from the engines may vary greatly. It would be difficult to choose appropriate weighting for the compositional scores. Even worse, retrieval scores are not always available from the retrieval engines. Rank-based retrieval on the other hand offers an alternative to circumvent these problems. In this work, we use rank-based integration to integrate ranked retrieval lists produced by different retrieval models using different indexing scales.

5. EXPERIMENTAL CORPORA

In the Mandarin spoken document retrieval experiments, we used the transcribed Mandarin news data in the TDT-2 collection obtained from LDC.[3] This includes Mandarin radio broadcast from Voice of America. There are a total of 48 hours of audio from 2,265 stories. The data spans the period from March 1998 to June 1998.

The textual queries are Chinese articles from the TDT-2 collection of the XinHua news articles. In the TDT-2 collection, both the Chinese textual queries and the Mandarin broadcast news are

[3] Linguistic Data Consortium, http://www.ldc.upenn.edu.

annotated with relevance judgements based on 17 predefined topics. For our experiments, we focus on the 15 topics that have more than two relevant Mandarin audio documents per topic. For each of these 15 topics, 3 textual news articles are randomly sampled for use as our textual queries.

The Mandarin spoken document retrieval task involves a query-by-example setup. News articles are used as queries for retrieving relevant spoken documents. The task is identical to that in [9]. The textual news articles are first segmented into words. The segmented words are also converted to syllable bigrams and character bigrams. Retrieval experiments are then performed on these scales separately.

For evaluation, the non-interpolated mean average precision (mAP) is used as defined in Eq. (10).

$$mAP_{non-int} = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{M}\sum_{j=1}^{M_i}\left\{\frac{1}{N_j}\sum_{k=1}^{N_j} precN_{Q_j}(k)\right\}\right\} \tag{10}$$

where $mAP_{non-int}$ is the non-interpolated mean average precision, N_j is the total number of relevant documents for topic j, M_i is the total number of queries in batch i, L is the total number of batches of queries, $precN_{Q_j}(k)$ is the precision for Q_j when k retrieved documents are relevant.

6. RESULTS

6.1. Retrieval performance without integration

Retrieval with the HMM-based model at the word scale gave the best performance (mAP=0.566). While the HMM-based model fared better with words than subwords, the VSM preferred subwords to words in term of retrieval performance. The retrieval results for these two models are shown in Table 1.

	Word	Char2	Syl2
HMM	0.566	0.559	0.556
VSM	0.539	0.562	0.562

Table 1. Retrieval performance for the HMM-based model and the VSM using different scales of indexing units. Syl2 refer to syllable bigrams and Char2 means character bigrams.

6.2. Multi-scale integration for individual retrieval models

For the multi-scale retrieval experiments, we have fixed the retrieval models and then multi-scale integration is applied to the different indexing scales. By applying Eqn. (9) to these ranked lists, the different indexing scales are integrated.

First, results from integration across different indexing scales for the HMM-based model are shown in Figure 1. We observed the trends based on our results:

1. The performance is improved by integrating the word scale with the subword scales (syllable bigrams or character bigrams). The best performance improvement (mAP=0.578) is obtained from integrating the word and character bigram scales.
2. Integration between subword scales shows no observable improvement in retrieval performance.

Figure 2 shows multi-scale integration results for the VSM. The following observations are made:

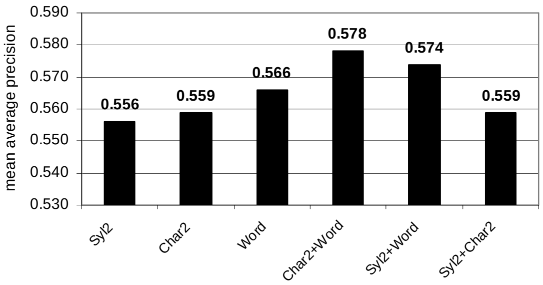

Fig. 1. Retrieval results for rank-based multi-scale integration using the HMM-based model.

1. Integrating the word and character bigram scales gave minor improvement in performance over the individual scales (mAP=0.563).
2. When integrating syllable bigrams with other scales, there is no improvement in performance.

Hence, multi-scale integration for the HMM-based model showed similar performance trends as the VSM. Integrating the word and character bigram gave the best performance for both models.

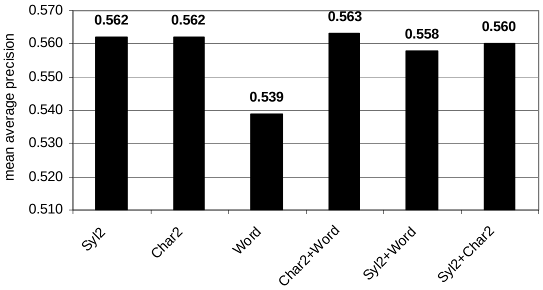

Fig. 2. Retrieval results for rank-based multi-scale integration using the VSM.

6.3. Multi-scale and multi-model integration

	VSM Word	 Char2	 Syl2	No integration
HMM-Word	0.563	**0.591**	**0.580**	0.566
HMM-Char2	**0.565**	**0.565**	0.560	0.559
HMM-Syl2	**0.561**	**0.569**	0.561	0.556
No integration	0.539	0.562	0.562	

Table 2. Rank-based multi-scale multi-model integration between HMM and VSM.

The multi-scale and multi-model retrieval results from integrating HMM-based model with VSM are shown in Table 2. It is observed from Table 2 that by integrating the ranked lists obtained using words in HMM-based model with the list obtained using character bigram in VSM, the best performance is achieved (mAP=0.591). It corresponds to an improvement of 4.4% over the best retrieval performance at the individual scales before integration.

From Table 2, it can also be seen that there are consistent performance improvements by integrating the word scale with any subword scale. By referring to Figure 1, the same observation is also found for integrating among word and subword scale results from HMM-based retrieval model. Moreover, it is also found that the improvement obtained by integrating across models is greater than that achieved from integration of different scales from the same retrieval model.

7. DISCUSSIONS

The multi-scale integration results show that by integrating word and subword scales (both within and across models), greater improvement can be achieved. This implies that these two indexing scales are complementary. However, between the subword scales of syllable bigrams and character bigrams, integration provides less performance improvement.

Word

Rank	1	2	3	4	5
Relevance	√	√	X	X	X

Char2

Rank	1	2	3	4	5
Relevance	X	√	√	X	X

Syl2

Rank	1	2	3	4	5
Relevance	X	√	X	√	X

Fig. 3. Extraction of ranked retrieval list from a selected query returned by retrieval at different scales using the VSM. Word and subword scales demonstrate different behaviour at low recall. While the two subwords scale (character bigram and syllable bigram) show similar retrieval performance.

Figure 3 shows the ranked retrieval list for the three indexing scales using the VSM: word, character bigram and syllable bigram. From the ranked lists of character bigram and syllable bigram, it can be seen that both scales have the relevant documents ranked second. In fact, both runs placed the same irrelevant document on the top of the lists. When we compare the ranked lists from word scale to that from subword scales (character bigram or syllable bigram), it is found that relevant documents are ranked higher in the list returned by word scale. These imply that both the character and the syllable scales demonstrate similar behaviour during retrieval. On the other hand, retrieving at the word scale and the subword scale demonstrate different performances.

HMM Word

Rank	1	2	3	4	5
Relevance	√	X	X	X	√

VSM Char2

Rank	1	2	3	4	5
Relevance	X	√	X	√	X

HMM Word + VSM Char2

Rank	1	2	3	4	5
Relevance	√	X	√	X	√

Fig. 4. Extraction of ranked retrieval lists from a selected query for retrieval using word in HMM-based model, character bigram in VSM and the integrated results. The integrated results show improvement in retrieval performance.

When multi-scale multi-model integration is applied, further performance improvement is observed. The ranked lists for the multi-model retrieval runs are shown in Figure 4. It can be seen that by integrating the HMM-word and VSM-character configurations, the re-scored retrieval list improves the document rankings. This gain relies on the document rankings in the ranked lists of the compositional runs. As illustrated in the figure, when documents are ranked high by both models, they will preserve the high rankings after the integration. In this example, the document ranked first and second by the individual runs is boosted to be ranked first after integration. Similar improvement in ranking is also observed for the document ranked third in the integrated result. In general, if relevant documents are ranked higher at low recall level, the overall average precision will improve. On the other hand, for those documents determined irrelevant (low rankings) by both ranked lists, they are also further suppressed. As a result, the overall effect by integration is to amplify the mutual knowledge from different sources.

8. CONCLUSIONS

We proposed the use of multi-scale and multi-model integration for improving retrieval performance in Chinese spoken document retrieval. Different knowledge sources from the multiple indexing scales and retrieval models can be leveraged to improve the retrieval performance. The ranked retrieval lists from different experimental runs are integrated in a rank-based manner. Our experiments show that by integrating different scales using different retrieval models can improve the retrieval performance. The best performance is obtained by integrating word scale results from HMM-based model with character bigram scale results from VSM (mAP=0.591).

9. REFERENCES

[1] A. Berger and J. Lafferty, “The Weaver system for document retrieval,” in *Proc. of the TREC-8*, 1999, pp. 163–74.

[2] K. Ng, “A maximum likelihood ratio information retrieval model,” in *Proc. of the TREC-8*, 1999, pp. 483–92.

[3] F. Song and W. B. Croft, “A general language model for information retrieval,” in *Proc. of ACM CIKM*, 1999, pp. 316–21.

[4] J. Makhoul *et. al.*, “Speech and language technologies for audio indexing and retrieval,” *Proc. of IEEE*, vol. 88, pp. 1338–53, 2000.

[5] B. Chen *et. al.*, “An HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval,” in *Proc. of the 7th EUROSPEECH*, 2001, vol. 2, pp. 1045–8.

[6] G. Salton and M. McGill, *Introduction to modern information retrieval*, McGraw-Hill, New York NJ, 1983.

[7] H. Meng *et. al.*, “Multi-scale audio indexing for Chinese spoken document retrieval,” in *Proc. of the 6th ICSLP*, 2000, vol. IV, pp. 101–4.

[8] K. Ng, “Information fusion for spoken document retrieval,” in *Proc. of the ICASSP*, 2000, pp. 2405–8.

[9] H. Meng *et. al.*, “Mandarin-English Information (MEI): Investigating translingual speech retrieval,” Tech. Rep., Johns Hopkins University, Baltimore, USA, 2000, Final report : [online] $http//www.clsp.jhu.edu/ws2000/final\_reports/mei.$

[10] W. K. Lo, P. Schone, and H. Meng, “Multi-scale retrieval in MEI: an English-Chinese translingual speech retrieval system,” in *Proc. of the 7th EUROSPEECH*, 2001, vol. 2, pp. 1303–6.

[11] H. M. Wang and B. Chen, “Comparison of word and subword indexing techniques for Mandarin Chinese spoken document retrieval,” in *Proc. of the 2nd PCM*, 2001.