Comparison of Word and Subword Indexing Techniques for Mandarin Chinese Spoken Document Retrieval

Hsin-min Wang and Berlin Chen

Institute of Information Science, Academia Sinica
128 Academia Road, Section 2, Taipei 115, Taiwan
{whm, berlin}@iis.sinica.edu.tw

Abstract. In this paper, we investigate the use of words and subwords (including both characters and syllables) in audio indexing for Mandarin Chinese spoken document retrieval. Two retrieval approaches, including the well-known vector space model approach and the newly proposed HMM/N-gram-based approach, are used in the present work. We focus on the use of an entire Chinese textual story (from a newspaper) as a query to retrieve Mandarin Chinese spoken documents (from news broadcasts). Experiments are based on the Topic Detection and Tracking Corpora.

1 Introduction

Massive quantities of audio and multimedia content, such as broadcast radio and television programs, are becoming increasingly available in the global information infrastructure. Since users need to be able to search for desired information efficiently, there is increasing demand for multimedia information retrieval technologies. As a result, the spoken document retrieval (SDR) task has been extensively studied in recent years [1-2]. In the area of Mandarin Chinese spoken document retrieval, some research works have been conducted at Academia Sinica, Taipei [3], and at The Chinese University of Hong-Kong [4]. In addition, Mandarin-English Information (MEI), a research project conducted in the Johns Hopkins University Summer Workshop 2000, investigated the use of an entire English newswire story (text) as a query to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) in the document collection [5].

In Mandarin Chinese, there exists an unlimited number of words, though only tens of thousands of them are commonly used. Each word is composed of from one to several characters. Each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are easily generated every day by combining a few characters or syllables. For example, the combination of 電 (electricity) and 腦(brain) gives a new word, 電腦(computer). Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio. On the other hand, an inventory of about 6,800 characters provides full textual coverage of written Chinese (in GB code), and one of about 13,000 characters does so for conventional Chinese (in Big5 code).

There is a many-to-many mapping between characters and syllables. For example, the character 乾 may be pronounced as /gan1/ or /qian2/ while all of the characters 甘干柑肝竿尷疳 are also pronounced as /gan1/ and all of 前錢潛黔虔掮 are pronounced as /qian2/. Consequently, a foreign word can very often be translated into different Chinese words. For example, “Kosovo” in “As the Kosovo peace talks in France…” may be translated into 科索沃/ke1-suo3-wo4/, 科索佛/ke1-suo3-fo2/, 科索夫/ke1-suo3-fu1/, 科索伏/ke1-suo3-fu2/, or 柯索佛/ke1-suo3-fo2/.

Word-level indexing features possess more semantic information than subword-level features; thus word-based retrieval enhances precision. On the other hand, subword-level indexing features are more robust against Chinese word tokenization ambiguity, Chinese homophone ambiguity, the open vocabulary problem, and speech recognition errors; thus, subword-based retrieval enhances recall. Consequently, there is good reason to study information fusion of indexing features of different levels. In this paper, we first investigate the use of words and subwords (including both characters and syllables) in audio indexing for Mandarin Chinese spoken retrieval and then explore information fusion.

In the following, all the experiments were conducted to study the use of an entire Chinese newswire story (text) as a query to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) from the document collection. Such a retrieval context is termed *query-by-example*. The experiments were based on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). Two retrieval approaches were adopted in this work: the well-known vector space model approach and the HMM/N-gram-based approach that we recently proposed [6].

2 Experimental Corpora

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

The Dragon large-vocabulary continuous speech recognizer provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3). We 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 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 (about 51k words) to tokenize the manual transcriptions for computing error rates. We also used this augmented LDC lexicon to tokenize the text query exemplars in the retrieval experiments.

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

	TDT-2 (Dev.) 1998, 02-06			**TDT-3 (Eval.) 1998, 10-12**		
# Spoken documents	2,265 stories, ~46hrs of audio			3,371stories, ~98hrs 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 doc. per query	2	95	29.3	3	89	20.1

3 Retrieval Models

3.1 The 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 term [3], such as word unigrams or overlapping word bigrams (or called word pairs). Each component $g(t)$ of a feature vector $\vec{d}_s$ for a document D is associated with the statistics of a specific indexing term t:

$$g(t) = (1 + \ln(c(t))) \cdot \ln(N / N_t), \tag{1}$$

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

$$\mathrm{R}_s(\vec{q}_s, \vec{d}_s) = (\vec{q}_s \bullet \vec{d}_s) \Big/ (\|\vec{q}_s\| \cdot \|\vec{d}_s\|). \tag{2}$$

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

$$\mathrm{R}(Q, D) = \sum_s w_s \cdot \mathrm{R}_s(\vec{q}_s, \vec{d}_s), \tag{3}$$

where w_s represents empirically tunable weights. We mainly use unigrams and overlapping bigrams (also called overlapping pairs) since previous works [3-5] indicated that they are most effective.

3.2 The HMM/N-gram-based Model

In the probability model approach, 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 query Q is observed; i.e., $P(D \text{ is } R|Q)$, which can be transformed into the following equation by applying Bayes' theorem:

$$P(D \text{ is } R|Q) = \frac{P(Q|D \text{ is } R)P(D \text{ is } R)}{P(Q)}, \tag{4}$$

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 Equation (4) 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 means of the probability $P(Q|D \text{ is } R)$ for the problem studied here.

In the HMM/N-gram-based approach, a 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 subword, while each document D is modeled by a single-state discrete HMM as shown in Fig. 1. The observation probabilities for this HMM are modeled by the weighted sum of N-gram probabilities of words or subwords. Therefore, the relevance measure, $P(Q|D \text{ is } R)$, can be estimated by means of the N-gram probabilities of the indexing term sequence for the query, $Q = q_1 q_2 .. q_n .. q_N$, predicted by document D . As mentioned earlier, in the present work, we mainly use unigrams and bigrams. Equations (5) and (6) illustrate, respectively, the estimation of $P(Q|D \text{ is } R)$ based on unigrams alone and based on both unigrams and bigrams:

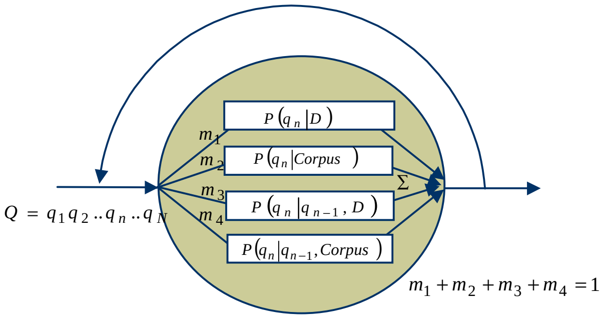
Type I: Unigram-based (Uni)

$$P(Q|D \text{ is } R) = \prod_{n=1}^{N} \left[m_1 P(q_n|D) + m_2 P(q_n|Corpus) \right]; \tag{5}$$

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

$$\begin{aligned} P(Q|D \text{ is } R) = & \left[m_1 P(q_1|D) + m_2 P(q_1|Corpus) \right] \times \\ & \prod_{n=2}^{N} \left[m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus) \right]; \end{aligned} \tag{6}$$

here, $P(q_n|D)$ is the unigram probability of a specific indexing term q_n within 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 document D . In order to model the general distribution of the indexing terms, both unigram and bigram parameters trained by a large text corpus, i.e., $P(q_n|Corpus)$ and $P(q_n|q_{n-1}, Corpus)$, were also included in Equations (5) and (6). In addition, for Equations (5) and (6), the weights m_i were summed to 1 (e.g., $\sum_{i=1}^{4} m_i = 1$ in Equation (6)), and the weights were tied among all the documents. These weights can be optimized using the expectation-maximization (EM) algorithm given a training set of query exemplars and their corresponding query-document relevance



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

Fig. 1. The HMM structure for a specific document D .

information. For example, the weight m_1 of Equation (5) can be estimated using the following equation:

$$m_1 = \frac{\sum\limits_{Q \in [TrainSet]_Q} \sum\limits_{D \in [Doc]_{R \text{ to } Q}} \sum\limits_{q_n \in Q} \left[\frac{m_1 P(q_n|D)}{m_1 P(q_n|D) + m_2 P(q_n|Corpus)} \right]}{\sum\limits_{Q \in [TrainSet]_Q} |Q| \cdot |[Doc]_{R \text{ to } Q}|}, \tag{7}$$

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 query Q , and $|[Doc]_{R \text{ to } Q}|$ is the total number of documents relevant to query Q . Fig. 1 depicts the Type II (Uni+Bi) HMM structure for a specific document D .

4 Experiments

4.1 Experiment Setup

In the HMM/N-gram-based approach, the probabilities of $P(q_n|Corpus)$ and $P(q_n|q_{n-1}, Corpus)$ in Equations (5) and (6) were estimated using a general text corpus consisting of 40 million Chinese characters. The weights m_i were derived by means of the EM training formula as described in Equation (7) 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 the TDT-2 document collection. These weights were applied to the evaluation set of the TDT-3 document collection. In the result tables below, the test results obtained for manual transcription of the spoken documents (denoted as TD) are also provided for comparison with the results obtained for erroneous transcription through speech recognition (denoted as SD). The test results are expressed in terms of the *mean non-interpolated average precision* (*mAP*) following the TREC evaluation [7], which is computed by the following equation:

$$mAP = \frac{1}{m}\sum_{i}^{m}\frac{1}{n_i}\sum_{j}^{n_i}\frac{j}{r_{i,j}}, \qquad (8)$$

where m is the number of queries, n_i is the total number of documents that are relevant to query i, and $r_{i,j}$ is the position (rank) of the j-th document that is relevant to query i, counting down from the top of the ranked list.

4.2 Word- vs. Subword-level Indexing Using The Vector Space Model

Table 2 shows the retrieval results obtained by applying the vector space model retrieval approach to both the TDT-2 and TDT-3 collections. It can be found from the first two columns of Table 2 that, for the word-level indexing features, using unigram information alone achieved reasonable performance while including overlapping bigram information offered only limited improvement. On the other hand, for the subword-level indexing features, including overlapping bigram information always gave significant improvement, especially for the syllable-level indexing features (the last two columns). In other words, for the subword-level indexing features, using unigram information alone seemed inadequate. Comparing the best performance of the word-, character- and syllable-level indexing features, the word-level indexing features outperformed the character- and syllable-level indexing features in most cases, but the syllable-level indexing features (the Uni+Bi case) performed best when applied to the real, desired case, the erroneous speech transcriptions (SD) of the TDT-3 evaluation set. Another interesting observation is that, 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 was only slightly lower than that for the TD cases.

4.3 Word- vs. Subword-level Indexing Using The HMM/N-gram-based Model

The retrieval results obtained when the HMM/N-gram-based retrieval approach was applied are shown in Table 3. Several observations could be made based on these results. First, similar to the case with the vector space model approach, the word-level indexing features in general outperformed the character- and syllable-level features, but the syllable-level features (the Uni+Bi case) performed best when applied to the real, desired case of SD for TDT-3. Second, unlike the vector space model approach, for the word- and character-level indexing features, including bigram information for indexing always degraded the retrieval performance instead of enhancing it. Since the numbers of distinct words and characters (51k and 6.8k) are relatively large compared to the number of syllables (0.4k), the estimation of bigram probabilities for the word- and character-level indexing features inherently suffered from the sparse data problem. Obviously, in Equation (6), the smoothing terms obtained from the general text corpus did not work well. This needs further study. Third, using syllable unigram information alone for indexing in the HMM/N-gram-based approach always gave significantly better performance than did using syllable unigram information alone in the vector space model approach. Fourth, the HMM/N-gram-based approach achieved consistently better performance than the vector space model approach, and the difference between the two was significantly larger for the TDT-2 development set from which the linear combination weights were trained.

Table 2. Retrieval results of the vector space model approach

		Word-level		Character-level		Syllable-level	
		Uni	Uni+Bi	Uni	Uni+Bi	Uni	Uni+Bi
TDT-2 (Dev.)	TD	0.5548	0.5623	0.5122	0.5441	0.3412	0.5254
	SD	0.5122	0.5225	0.4803	0.5176	0.3306	0.5077
TDT-3 (Eval.)	TD	0.6505	0.6531	0.6275	0.6373	0.3963	0.6502
	SD	0.6216	0.6233	0.5836	0.6106	0.3708	0.6353

Table 3. Retrieval results of the HMM/N-gram-based approach

		Word-level		Character-level		Syllable-level	
		Uni	Uni+Bi	Uni	Uni+Bi	Uni	Uni+Bi
TDT-2 (Dev.)	TD	0.6327	0.5427	0.5743	0.5204	0.4698	0.5697
	SD	0.5658	0.4803	0.5437	0.4804	0.4411	0.5305
TDT-3 (Eval.)	TD	0.6569	0.6141	0.6465	0.5843	0.5343	0.6544
	SD	0.6308	0.5808	0.6031	0.5309	0.5177	0.6413

4.4 Information Fusion

Word-level indexing features possess more semantic information than syllable-level features. On the other hand, 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 of Mandarin Chinese and speech recognition errors in spoken documents. This is shown by the above experimental results. It was believed that proper fusion of the word- and subword-level information would be useful in the retrieval task. As a result, fusion of the best approaches described in Sect. 4.3 and 4.4 using the following equation was tested:

$$\mathrm{R}(Q,D)=w_w\mathrm{R}_w(Q,D)+w_c\mathrm{R}_c(Q,D)+w_s\mathrm{R}_s(Q,D), \quad (9)$$

which is simply the weighted sum of the relevance scores obtained with the word-, character- and syllable-level indexing features.

The results are shown in Table 4, where for the vector space model (denoted as VSM) approach, all the indexing features include both unigram and bigram information, while for the HMM/N-gram-based (denoted as HMM) approach, the word-level and character-level features are based on unigrams only, while the syllable-level features use both unigrams and bigrams. Comparison with the results obtained using either word-, character- or syllable-level information alone shows that fusion was in general helpful for retrieval, though in some cases, it slightly degraded retrieval performance instead of enhancing it. We also combined the two approaches by using the weighted sum of their relevance scores, based in both cases on the S+C+W case. Fusion gave average precision results of 0.6218 and 0.5726 for TD and

SD for TDT-2, and of 0.6815 and 0.6650 for TD and SD for TDT-3. Based on the results shown in the last column of Table 4, fusion was indeed helpful with respect to the evaluation set (TDT-3).

Table 4. Retrieval results based on information fusion

			S+C	S+W	C+W	S+C+W
TDT-2 (Dev.)	TD	VSM	0.5620	0.5744	0.5619	0.5741
		HMM	0.5860	0.6264	0.6197	0.6254
	SD	VSM	0.5187	0.5293	0.5372	0.5397
		HMM	0.5302	0.5769	0.5664	0.5643
TDT-3 (Eval.)	TD	VSM	0.6605	0.6683	0.6540	0.6664
		HMM	0.6408	0.6545	0.6734	0.6697
	SD	VSM	0.6380	0.6447	0.6409	0.6456
		HMM	0.6210	0.6334	0.6471	0.6466

5 Concluding Remarks

In this paper, we have focused on the use of words, characters and syllables in audio indexing for Mandarin Chinese spoken retrieval. Though word-level indexing features outperformed character- and syllable-level features in most cases, syllable-level indexing features performed very well in the real, desired case of retrieval from the erroneous speech transcriptions (SD) of the evaluation set. We also found that information fusion of indexing features of different levels was, in general, useful for retrieval.

References

1. Jones, K. S., Jones, G. J. F., Foote, J. T., Young, S. J.: Experiments on Spoken Document Retrieval. Information Processing & Management 32(4) (1996) 399-417
2. Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., Srivastava, A.: Speech and Language Techniques for Audio Indexing and Retrieval. Proceedings of the IEEE 88(8) (2000) 1338-1353
3. Chen, B., Wang, H. M., Lee, L. S.: Retrieval of Mandarin Broadcast News Using Spoken Queries. Int. Conf. on Spoken Language Processing (2000)
4. Meng, H., Lo, W. K., Li, Y. C., Ching, P. C.: Multi-scale Audio Indexing for Chinese Spoken Document Retrieval. Int. Conf. on Spoken Language Processing (2000)
5. Meng, H., et al.: Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval. Human Language Technology Conf. (2001)
6. Chen, B., Wang, H. M., Lee, L. S.: An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval. European Conf. on Speech Communication and Technology (2001)
7. Harman, D.: Overview of the Fourth Text Retrieval Conference. The Fourth Text Retrieval Conf. (1995)