ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

A DATA-DRIVEN INDEXING APPROACH FOR CHINESE SPOKEN DOCUMENT RETRIEVAL

Chun-Jen WANG, Berlin CHEN, Lin-shan LEE

Graduate Institute of Communication Engineering,
National Taiwan University,Taipei
stanley@speech.ee.ntu.edu.tw,berlin@iis.sinica.edu.tw,lsl@iis.sinica.edu.tw

ABSTRACT

The choice of indexing features is critical to the performance of a retrieval system. Predefined, overlapping, fixed-length term sequences are widely used in many retrieval systems. However, predefined feature sets are often riddled with meaningless and non-informative terms, which unavoidably degrades retrieval performance, and explodes the feature set. In this paper we present a statistical approach to derive data-driven term segments as features. We let the data to tell which features are important and which are not. The results show that very satisfactory performance can be achieved with these data-driven indexing features while retaining very compact feature set size. This approach also has the potential to identify domain-specific terminologies or newly-generated phrases.

1. INTRODUCTION

With the exponential accumulation of multimedia information over the Internet, large repositories of documents include some information in speech form. Intelligent and efficient retrieval techniques have been developed to provide the users with easy access to all kinds of spoken documents [1]. Despite these developments, spoken document retrieval remains a very challenging research topic.

Retrieval systems construct representations of documents and queries based on indexing features. Word- and subword-based indexing features have been explored extensively for spoken document retrieval. Experiments have shown that word-based features are more important in English [2], while subword-based information is highly discriminative for Chinese due to the monosyllabic structure of the language[3, 4]. In Chinese language, words carry semantically complete meaning, therefore, is suitable for indexing. However, important keywords might not be included in the system lexicon, which is known as “Out-of-Vocabulary” problem. Syllable-level (subword-based) information is more robust to recognition error. The use of syllable-level indexing features limits the size of vocabulary needed for retrieval, but may bring more ambiguity. Character-level (subword-based) features can cope with out-of-vocabulary problem well ,but may be more susceptible to imperfections of lexicons and language models during recognition stage.

Currently most indexing features adopted are predefined fixed-length term segments. They performed well at moderate length, 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 term segments as features. We verified this indexing scheme on both syllable- and character-level. It was found that the features derived in this way are very often semantically meaningful and therefore can capture more intrinsic concepts during retrieval. Very encouraging results were obtained and reported.

2. EXPERIMENTAL CORPORA

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 [5] 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 [6].

3. RETRIEVAL MODELS

In the following, the two 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,

$$g(t) = (1 + \ln(c(t))) \cdot \ln(\frac{N}{N_t}) \qquad (1)$$

where $c(t)$ is the occurrence count of the term t within the document D or query Q , and $\ln(\frac{N}{N_t})$ 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 document 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' theoremG

$$P(D\ is\ R|Q) = \frac{P(Q|D\ is\ R)P(D\ is\ R)}{P(Q)} \quad (2)$$

Because $P(Q)$ is identical for all documents, and it is difficult to estimate the probability of $P(D\ is\ R)$, the remaining term $P(Q|D\ is\ R)$ is used to rank the documents [7]. The query Q is treated as a sequence of input observations (or indexing terms), $Q = q_1 q_2 \ldots q_n \ldots q_N$, while each document D is modeled as a single-state discrete HMM. We adopted the same HMM structure as in the previous work [6] 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. [6].

4. DRAWBACKS OF PREDEFINED FIXED-LENGTH INDEXING FEATURES

Predefined fixed-length indexing features are widely used in many retrieval systems. However, they have some defects. First, the dimension of feature space grows prohibitively large even at small N, which often increase computation cost and/or total number of parameters to be estimated. Second, many of such features are meaningless and ambiguous. Third, the number of N is limited below 3 considering computation and storage cost, which makes it difficult to capture longer context information. Fourth, such features are often hard to follow rapid emerging vocabulary. New words are easily generated everyday, and there exists many domain-specific terminologies. Based on these observations, we presented an precious indexing scheme that can derive data-driven term sequences from corpus statistics.

5. DATA-DRIVEN INDEXING APPROACH

Since many predefined indexing features are meaningless and ambiguous, why not let data "tell" which indexing features are useful and which are not? This is the basic idea behind data-driven indexing approach. The measure we used for feature selection is the geometrical average of the forward and backward bigram [8] of adjacent terms (ω_i, ω_j)

$$FB(\omega_i, \omega_j) = \sqrt{P_f(\omega_j|\omega_i)P_b(\omega_i|\omega_j)} \quad (3)$$

where

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

$$P_b(\omega_i|\omega_j) = \frac{P(W_{t+1} = \omega_j\ , W_t = \omega_i)}{P(W_{t+1} = \omega_j)} \quad (5)$$

		Predefined (167281)		Data-Driven (3800-5600)	
		VSM	HMM	VSM	HMM
TDT-2	TD	0.538	0.572	0.577	0.588
	SD	0.518	0.531	0.531	0.567
TDT-3	TD	0.668	0.656	0.662	0.677
	SD	0.652	0.643	0.651	0.651

Table 1. Retrieval results using syllable-level, both predefined and data-driven, indexing features

		Predefined (4.9×10^7)		Data-Driven (14000-20000)	
		VSM	HMM	VSM	HMM
TDT-2	TD	0.549	0.584	0.582	0.616
	SD	0.511	0.543	0.510	0.558
TDT-3	TD	0.668	0.653	0.667	0.672
	SD	0.642	0.609	0.644	0.641

Table 2. Retrieval results using character-level, both predefined and data-driven, indexing features

For clarity, we take syllable-level as an example. 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 syllable pairs scored above a threshold became a new feature and all instances of these pairs in the corpus were replaced by the new features. The procedures above applies to character-level as well. We empirically chose an optimal threshold based on TDT-2 collection, and evaluated on TDT-3 collection using this optimal threshold. Both indexing schemes were explored with both syllable- and character-level.

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). Two retrieval models, the vector space model (VSM) and hidden Markov model (HMM), were explored with both predefined and data-driven indexing approaches.

Experimental results are listed in Tables 1 and 2 respectively. All retrieval results were presented in terms of *non-interpolated average precision* [1]. For syllable-level indexing features, as shown in Table 1, the numbers in parentheses indicate the size of lexicon (i.e. total number of unique indexing features) under each indexing scheme. For the VSM model, predefined column use base syllable pairs (N=2) as features. It is clear that using the very compact data-driven indexing features produced results at least comparable to those produced by predefined indexing features, but with much smaller size of lexicon (the size of lexicon of data-driven indexing scheme is only 3% of that of predefined indexing scheme). For the HMM model, data-driven indexing approach also offers similar advantages. In this paper, we adopted the Uni+Bi* as in previous work [6]. In fact, in almost all cases the data-driven features outperformed the predefined features. For character-level indexing features, we use base character pairs (N=2) as features for the VSM model. It can be found in Table 2 that data-driven indexing features can achieved equal performance as predefined indexing features, but with much smaller lexicon size. For the HMM model, data-driven features are still superior to predefined features. Take the first row of Table 2 as example, for the VSM model, the size of the data-driven feature set is only 20,705 (including character sequences with length ranging from 1 to 4), which is far less than that of predefined character pairs, on the order of 10^7 , but giving better performance. This is quite amazing.

It is interesting to mention that we found many of the data-driven indexing features can be directly translated to 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 data-driven indexing features include terms composed of three to five single syllables/ characters. It is well known that for Chinese words with three or more syllables/characters, there is almost a one-to-one correspondence between the words and the syllable/character segments [9]. Predefined 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 lexicon. Therefore, the data-driven approach here can derive information-abundant indexing features very efficiently while discarding non-informative features.

6. CONCLUSIONS

In this paper, we present a statistical approach to derive data-driven indexing features and compare this mechanism to the widely-used predefined indexing scheme. Experimental results strongly supported the concept of this indexing approach. Altogether, this data-driven indexing mechanism featured some advantages: First, feature sets derived in this way are very compact, thus reducing the load of computation and storage. Second, documents/queries indexed with data-driven feature sets can provide more precise representations, thus achieving better retrieval performance. Third, they can capture longer distance of context information while retaining compact feature set. Fourth, their potential use lies in detecting newly-generated keywords. Fifth, this statistical approach may filter out non-informative or meaningless indexing features. Further investigations are under progress.

7. REFERENCES

[1] D. Harman, "Overview of the Ninth Text REtrieval Conference (TREC-9)," 2000.

[2] K. Ng, "Information fusion for spoken document retrieval," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000.

[3] B. Chen,H.M. Wang,L.S. Lee, "Retrieval of broadcast news speech in mandarin chinese collected in taiwan using syllable-level statistical characteristics," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000.

[4] Berlin Chen, *Speech Information Retrieval for Mandarin Chinese — Syllable-Based Indexing Features,Statistical Retrieval Models and Improved Approaches*. PhD thesis, National Taiwan University, 2001.

[5] P. Zhan,S. Wegmann,L. Gillick, "Dragon systems' 1998 Broadcast News Transcription System for Mandarin," *Proc. of the DARPA Broadcast News Workshop*, 1999.

[6] B. Chen,H.M. Wang,L.S. Lee, "An hmm/n-gram-based linguistic processing approach for mandarin spoken document retrieval," *Proc. European Conf. On Speech Communication and Technology (EUROSPEECH)*, 2001.

[7] David R.H. Miller, T. Leek, and R. Schwartz, "A hidden markov model information retrieval system," *Proc. ACM SIGIR*, 1999.

[8] G. Saon,M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, May 2001.

[9] L.S. Lee, "Voice Dictation of Mandarin Chinese," IEEE Signal Processing Magazine, 1997.