Query Expansion using Phonetic Confusions for Chinese Spoken Document Retrieval

Yuk-Chi LI(1), Wai-Kit LO(2), Helen M. MENG(1) and P.C. CHING(2)

(1) Department of Systems Engineering and Engineering Management and
(2) Department of Electronic Engineering, The Chinese University of Hong Kong
Email: ycli,hmmeng@se.cuhk.edu.hk, wklo,pcching@ee.cuhk.edu.hk

Abstract

This paper presents a method of query expansion based on phonetic confusions for retrieving spoken documents using text queries. This method is applied to a Chinese spoken document retrieval task. A series of experiments have been carried out for Cantonese broadcast news documents using a multi-scale syllable-based retrieval approach. Our results show an improvement from AIR (average inverse rank) of 0.481 to 0.491 when we apply query expansion based on phonetic confusions to our retrieval task.

Keywords: query expansion; Chinese spoken document retrieval; multi-scale syllable-based retrieval approach.

1 INTRODUCTION

Information retrieval (IR) has become an important problem for searching and managing massive amounts of information available on the Internet [1]. Much effort in IR has been devoted to textual information, but since information comes in multimedia forms, retrieval of audio information, especially speech, has become a hot topic of interest [2]. Spoken document retrieval involves the indexing and retrieval of speech data according to the user's queries. The prevalent technique is to split the task into two sub-tasks, e.g. in the recent TREC systems [3] – to convert the transcription of the spoken document into text by large-vocabulary continuous speech recognition (LVCSR), followed by the application of text-based retrieval techniques on the document transcriptions. Certain types of spoken documents, e.g. broadcast news, have open vocabularies. The use of LVCSR for indexing these spoken documents may encounter the out-of-vocabulary (OOV) problem, i.e. should an excerpt of the audio contain words absent from the recognizer's vocabulary, that part will not be adequately indexed for subsequent retrieval.

Indexing spoken documents by means of sub-word units may circumvent the OOV problem and fully indexing the audio data. The approach has been adopted by Ng et al. [4] By virtue of the syllabic nature of the Chinese language, the syllable seems to be a suitable sub-word unit for indexing spoken documents. Wang et al. [5] and Meng et al. [6] have used the syllable-based approach for indexing spoken documents in two Chinese dialects – Mandarin and Cantonese. This work extends the study reported in [6], where textual queries are used to retrieve Cantonese spoken documents. Textual queries can be mapped into syllable sequences by pronunciation dictionary lookup. Such syllable sequences are "clean" in the sense that they do not contain recognition errors. However, the syllable-based document representations produced by our recognizer during audio indexing are susceptible to speech recognition errors. In order to bridge the gap between the error-free syllable-based queries and errorful syllable-based documents, we attempt to apply a query expansion technique based on recognition confusions to improve the performance of retrieval.

2 CANTONESE SPOKEN DOCUMENTS

We have acquired a Cantonese news archive over the Internet. The video news clips are in RealMedia format, and cover a six-month period between June 1997 to February 1998. This includes 1801 news stories, each an individual video clip, with duration ranging from 11 seconds to 24 minutes. The average duration is 1.5 minutes. Each video clip is also accompanied by a textual summary. Every summary is brief, and cannot serve as a verbatim transcription of the audio track of the corresponding video clip.

Proceedings of the 5th International Workshop Information Retrieval with Asian Languages

3 SPOKEN DOCUMENT RETRIEVAL

3.1 Audio Indexing

Audio indexing is carried out using a Cantonese speech recognizer. We extracted the audio tracks from the video clips, and hence we have to run our recognizer on audio data that is in RealAudio™ format. Since we do not have transcribed RealAudio™ data for training a speech recognizer, we adapted from existing resources.

CUSENT [7] is a Hong Kong Cantonese continuous speech corpus. It is a recording of read sentences over a close-talking microphone from a quiet studio, and the script is a list of short sentences extracted from newspaper text. This is, to our knowledge, the only publicly available Cantonese corpus that supports large-vocabulary speech recognition development. Compared to the RealAudio™ format, CUSENT has audio of higher quality.

We first encoded the CUSENT corpus using the 8.5kbps CELP-based speech codec of the RealAudio™ format. This degraded data is then used to train our sub-syllabic acoustic models. We have also transcribed a small amount of original RealAudio™ data, and used 1.75 hours for further training, and 0.5 hours for testing. The training and testing data are disjoint. Evaluation based on the test set gives a syllable recognition accuracy of 34%. This reflects the harsh acoustic environments present in the broadcast news data, which consist of anchors' speech (from the studio), reporters' / interviewees' speech (from the field), and other types of audio (music, announcements, etc.) There is a substantial portion of wrongly transcribed syllables in the recognition outputs. The transcribed base syllable sequences are concatenated together to form the document representations for spoken document retrieval.

3.2 Retrieval Model

For retrieval, we used the vector-space model in SMART [8]. The term-weighing strategies used are shown in Equations (1) and (2):[1]

For term i in document d:

$$d[i] = \ln(tf_d[i]) + 1.0 \quad \ldots\ldots(1)$$

where $tf_d[i]$ is the frequency of term i in document d.

For term i in query q:

$$q[i] = [\ln(tf_q[i]) + 1.0] \times \ln\left(\frac{N+1}{n_i}\right) \quad \ldots\ldots(2)$$

where $tf_q[i]$ is the frequency of term i in query q,

N is the total number of queries, and

n_i is the number of queries with term i.

The similarity $S(q, d)$ between a query q and document d is measured by the normalized inner product, to form the basis of retrieval as show in Equation (3):

$$S(q, d) = \frac{q \bullet d}{\|q\| \bullet \|d\|} \quad \ldots\ldots(3)$$

3.3 Experiments

We investigate the use of textual queries for retrieving spoken documents. The titles from textual summaries of our news stories are extracted and used as textual queries. We have three types of documents derived from our Cantonese broadcast news archives: (i) textual documents represented in Chinese characters; (ii) documents represented as syllables converted from the text; and (iii) documents represented as syllables from speech recognition.

- **Textual Documents**

A benchmark is provided by the use of textual queries to retrieve textual documents. The textual summaries (excluding the titles) are used as our textual documents. Both queries and documents were first tokenized into Chinese words using a maximum-matching (greedy) algorithm together with CULEX [10] of size 47,000 words. The words become our indexing terms for retrieval. The SMART system has been augmented to process Chinese characters for this purpose.

- **Text-Converted Syllable Documents**

We performed retrieval using the syllable unit. Textual queries are mapped into their base syllable pronunciations by referencing CULEX. This puts the queries and documents in the same syllable space for retrieval. Indexing terms may thus be derived from the syllable unigrams, overlapping syllable bigrams/trigrams and also skipped syllable bigrams[2]. We feel that the overlap is important to avoid segmentation errors; and the skipped syllable bigrams can used to capture Chinese abbreviations e.g. in [11].

- **Recognized Syllable Documents**

The queries were mapped into their base syllable sequences. The recognized syllable sequences from the audio tracks were used as the spoken document representations. We used the same set of indexing terms mentioned previously (overlapping syllable n-grams and skipped syllable bigrams).

---

[1] We have used the "lnc/ltc" weighing, similar to [9].

[2] Skipped syllable bigrams are formed from three consecutive syllables by skipping the middle one.

Since our data lack relevance judgements between queries and documents, we have formulated the task of known-item retrieval. For each of the 1801 queries, there is only a single "relevant" document, which is the one corresponding to the same news story. Our evaluation metric is the average inverse rank (*AIR*). Retrieval for each query produces a ranked list of documents. We search for the rank (*R*) of the "relevant" document in the list, and score the performance for that query by means of the inverse rank (*1/R*). Averaging across all queries produces the AIR. Results are displayed in Figure 1.

- **Results**

The text retrieval benchmark (i.e. textual query, textual documents, indexed on words) gave AIR=0.834. From Figure 1, we see that overlapping bigrams gave superior performance when compared to unigrams or trigrams. The former may be too general and the latter may be too specific. Bigrams provide sequential constraints that are helpful for retrieval, and the use of overlapping character bigrams may ameliorate word tokenization errors. The overlapping character bigrams gave AIR=0.834, overlapping syllable bigrams gave AIR=0.830, and overlapping recognized syllables bigrams gave AIR=0.477. These results reflect the effect of the reduction of lexical knowledge as we migrate from character bigrams to syllable bigrams,[3] and speech recognition errors significantly degrade retrieval performances. We augmented the overlapping bigrams with skipped bigrams for capturing abbreviations. (see Figure 1). The use of overlapping and skipped bigrams gave AIR=0.481 for recognized syllables. The skipped bigrams seem to provide a slight improvement. These results prompted us to focus on the use of bigrams and skipped bigrams in our subsequent experiments.

4. QUERY EXPANSION USING PHONETIC CONFUSIONS

As mentioned previously, audio indexing by means of sub-word units can provide full coverage of the speech data, and circumvent the out-of-vocabulary problem. However, for a retrieval task involving textual queries and spoken documents, the textual queries need to be mapped into syllables by pronunciation dictionary lookup, and the spoken documents are transformed into a syllable-based representation via speech recognition. Hence the query syllables are error-free while the document syllables are errorful. In order to bridge this gap between queries and documents, we attempt to apply a query expansion technique based on syllable-syllable confusions produced by the recognizer. Our objective is to improve retrieval performance.

[3] Multiple Chinese characters may be pronounced as the same syllable. Migration from characters to syllables reduces the inherent lexical knowledge.

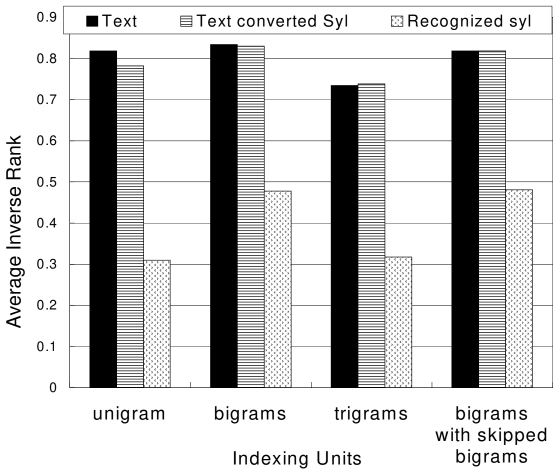

Figure 1. Retrieval performances based on average inverse rank, for a variety of indexing techniques – Chinese character or base syllable *n*-grams, where *n*=1,2 or 3 (i.e. unigrams, bigrams and trigrams), as well as skipped bigrams. Unigrams, bigrams and trigrams for "text" (the black bar graphs) refer to Chinese character unigrams, bigrams and trigrams respectively.

4.1 Syllable-Syllable Confusions from Recognition

A syllable confusion matrix is generated from our recognition outputs. In Section 3.1 we described a test set of 0.5 hours of transcribed RealAudio™ news audio, for which we obtained a syllable recognition accuracy of 34%. From the results of this experiment we extracted the syllable-syllable confusions. Furthermore, for each pair of reference syllable and hypothesized syllable, we tallied the number of reference syllable occurrences, as well as computed the rate of confusion. Examples are shown in Table 1.

4.2 Query Expansion and Robust Retrieval

We make use of two thresholds in expanding our queries based on syllable-syllable confusions:

(i) the frequency of occurrence of the reference syllable (hereafter referred to as *cutoff*).

(ii) the percentage of confusions with a specified hypothesize syllable (hereafter referred to as *confusion%).*

By adjusting these two thresholds, we can control the degree of expansion for our queries. The degree of expansion increases with decreasing values of *cutoff* and *confusion%.*

Ref syllable[4]	Hyp syllable	Sustitution counts	Ref syl counts	% of Substitution
cau	sau	4	16	25
cuk	zuk	5	11	45
jyut	jyu	7	34	21
lam	nam	2	8	25
mui	wui	4	14	29
muk	mou	2	8	25
paai	taai	5	23	22
pui	fui	3	8	38
seon	san	6	25	24
sou	zou	10	44	23

Table 1. Examples of syllable-syllable confusions produced by our speech recognizer. The columns from left to right show the reference syllable, the hypothesized syllable, substitution counts, reference syllable occurrences and probability of substitution of the hypothesized syllable for the reference syllable.

A higher degree of query expansion generally improves retrieval performance. The tradeoff, however, is that greater query expansion increases the number of query terms. We also tried to prevent excessive query expansion by taking reference to the pool of document terms when we expand into new query terms. New query terms will be expanded only if they are present in the pool of document terms. Our approach to query expansion is as follows:

For each syllable in the query syllable sequence, we search for its possible confusable counterparts in the confusion matrix. We generate a new query by adding every combination of possible confusions to the original query. For example, Figure 3a shows a textual query and its correct syllable sequence. The two possible confusions that apply here are: *gok* versus *go*; as well as *hong* versus *hang*. This produces five more terms if we use bigrams indexing and four more terms if we use skipped bigrams indexing, as shown in Figures 3b and 3c respectively.

市 面 各 項 慶 祝 回 歸 活 動 準 備 就 緒

si min gok hong hing zuk wui gwai wut dung zeon bei zau seoi

Figure 3a. A sample of textual query with its corresponding mapped syllable.

[4] We adopt the standard Cantonese syllable inventory set by LSHK (Linguistic Society of Hong Kong).

Original terms: si min *gok* *hong* hing zuk wui gwai wut dung zeon bei zau seoi

Syl. Confusions: *gok*→*go; hong*→*hang*

Newly added bigrams: *min-go go-hong go-hang gok-hang hang-hing*

Figure 3b. Query expansion produces five new terms from the original query when using syllable bigrams for indexing.

Original terms: si min *gok hong* hing zuk wui gwai wut dung zeon bei zau seoi

Syl. Confusions: *gok*→*go; hong*→*hang*

Newly added skipped bigrams: sigo gohing minhang hangzuk

Figure 3c. Query expansion produces four new terms from the original query when using skipped bigrams indexing.

All the newly created terms are appended together with the original query to form a new query for retrieval.

We have "optimized" the parameter settings for *cutoff* and *confusion%*, in order to maximize retrieval performance. The process involves selecting 600 queries randomly as our training set. The "optimized" parameter settings according to the training data were *cutoff*=1 and *confusion%*=9. (See Table 2). This produces an AIR value of 0.604 (with query expansion), which is higher than the baseline AIR of 0.573 (no query expansion). Using these selected settings, we achieved an improvement to 0.532 (with query expansion) from AIR=0.529 (baseline) for the remaining 1201 queries. Overall performance on the entire 1801 queries are 0.489 (with query expansion) from AIR=0.477 (baseline) respectively with syllable bigrams as indexing terms.

	600 Queries (Training)	1201 Queries (Testing)	1801 Queries (Overall)
AIR (Baseline)	0.573	0.529	0.477
AIR (Query Expansion)	0.604	0.532	0.489

Table 2. Average Inverse Rank (AIR) of expanded query retrieval for the randomly chosen 600 training queries, 1201 testing queries and the whole 1801 query set. Query expansion parameters are *cutoff=1* and *confusion%=9* respectively.

We have also used syllable bigrams augmented with skipped syllable bigrams as our indexing terms. The baseline result without query expansion is AIR=0.601. The "optimal" parameter settings computed from the training data are *cutoff*=22 and *confusion%*=10 respectively. Query expansion gave AIR=0.612 (performance on the 600 training queries); AIR=0.527 (performance on the 1201 testing queries) and AIR=0.491 (performance on the entire 1801 query set) respectively.

5. CONCLUSIONS

This paper reports on an approach for robust retrieval of Cantonese television news broadcasts. We used syllable-based units for audio indexing by speech recognition, where the syllables provide full phonological coverage of the Cantonese speech data, and circumvent the out-of-vocabulary problem. Our retrieval task involves textual queries and spoken documents. The textual queries are mapped into syllables by pronunciation dictionary lookup. Hence the query syllables are error-free, while the document syllables are susceptible to speech recognition errors. In order to bridge this gap between queries and documents, we applied a query expansion technique based on syllable-syllable confusions from our recognition outputs. Our objective is to achieve robustness to improve retrieval performance. The best baseline result is achieved by indexing with syllable bigrams and skipped syllable bigrams, giving an AIR=0.481. The performance is improved to AIR=0.491 when query expansion is applied.

6. ACKNOWLEDGMENTS

We would like to thank Television Broadcasts Limited for granting us permission to use their Cantonese news archives in our research.

7. REFERENCES

[1] Oard, D. W. and Dorr, B.J., “A Survey of Multilingual Text Retrieval”, *Technical Report UMIACS-TR-96-19 CS-TR-3615 of the Electrical Engineering Department*, University of Maryland, USA, 1996.

[2] Hauptmann, A.G., Scheytt, P., Wactlar, H.D., and Kennedy, P.E., "Multilingual Informedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages," *Proceedings of the DARPA Workshop on Broadcast News Understanding Systems, Lansdowne*, VA, February, 1998.

[3] Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P., Stadford, V. M. and Lund, B. A., “1998 TREC-7 Spoken Document Retrieval Track Overview and Results”, Proceedings of the Seventh Text Retrieval Conference, 1998.

[4] Ng, K. and Zue, V., "Towards Robust Methods for Spoken Document Retrieval," *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998.

[5] Chien, L. F. and Wang, H. M., "Exploration of Spoken Access for Chinese Text and Speech Information Retrieval," *Proceedings of the International Symposium on Signal Processing and Intelligent Systems*, 1999.

[6] Meng, H., Lo, W. K., Li, Y. C. and Ching, P.C., “A Study on the Use of Syllables for Chinese Spoken Document Retrieval”, *Technical Report SEEM99-11 of the Department of Systems Engineering and Engineering Management*, the Chinese University of Hong Kong, Hong Kong, 1999.

[7] Lo, W. K., Lee, T. and Ching, P. C., "Development of Cantonese Spoken Language Corpora for Speech Applications", *Proceedings of the first International Symposium on Chinese Spoken Language Processing, Singapore,* 1998.

[8] Salton, G. and McGill, M., “Introduction to Modern Information Retrieval,” McGraw-Hill, New York 1983.

[9] Buckley, C., Allan, J., Salton, G. and Singhal, A., “Automatic query expansion using SMART : TREC 3.” *Proceedings of the Third Text Retrieval Conference (TREC-3),* pp. 69-80, April 1995.

[10] http://dsp.ee.cuhk.edu.hk/speech, the Chinese University of Hong Kong, Hong Kong SAR, China.

[11] Chen, B., Wang, H. M. and Lee, L. S., "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan Using Syllable-level Statistical Characteristics", *Proceedings of ICASSP-2000.*