Document Expansion using a Side Collection for Monolingual and Cross-language Spoken Document Retrieval

Yuk-Chi LI and Helen M. MENG

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Email: {ycli,hmmeng}@se.cuhk.edu.hk

ABSTRACT

This paper presents a method of document expansion using a side collection for improving the overall performance in retrieving spoken documents using text queries. This method is applied to Chinese spoken document retrieval (SDR) tasks where a series of experiments have been carried out for both monolingual and cross-language SDR systems. In our monolingual retrieval experiments, Cantonese broadcast news documents are retrieved using a multi-scale syllable-based approach. Experimental results show that application of document expansion can achieve an improvement of 56% in average inverse rank (AIR). For the cross-language spoken document retrieval (CL-SDR) task where Mandarin broadcast news is retrieved using English textual queries, experimental results show that the use of document expansion has brought 14% relative improvement in retrieval performance.

Keywords: document expansion; Chinese spoken document retrieval; multi-scale syllable-based retrieval; cross-language spoken document retrieval.

1 INTRODUCTION

Information retrieval (IR) has become an important problem for searching and managing massive amounts of information available on the Internet [1]. Most previous effort in IR has been devoted to textual information. Since information comes in multimedia forms, retrieval of audio information, especially speech, has become a hot topic of interest [2]. Spoken document retrieval (SDR) involves indexing of the speech data and retrieval of information relevant to user's queries. The prevalent approach to SDR (e.g. in the recent TREC systems [3]) divides the task into two sub-tasks: automatic transcription of the spoken documents using large-vocabulary continuous speech recognition (LVCSR) and retrieval based on these transcriptions using text-based retrieval techniques.

The use of LVCSR for indexing spoken documents is subjected to the out-of-vocabulary (OOV) problem. For example, spoken documents in the news domain do not have a fixed vocabulary. In case there are words not covered by the recognizer's vocabulary, these words will be ignored or substituted by other words within the vocabulary. These recognition errors tend to cause retrieval errors.

Indexing spoken documents by means of subword units can circumvent the OOV problem. Audio data can be fully indexed by using subword units. This approach has been investigated by Ng et al. [4]. By virtue of the syllabic nature of the Chinese language, the syllable is a suitable subword unit for Chinese. Wang et al. [5][6] and Meng et al. [7] have used the syllable unit to index Chinese spoken documents in Mandarin and Cantonese respectively.

Since textual queries are used to retrieve spoken documents, a pronunciation dictionary is used to map all textual queries into syllables and this procedure is free from recognition errors. However, automatic transcription of spoken documents by speech recognition may introduce recognition errors during audio indexing. In order to bridge the gap between the *error-free* queries and erroneous documents, we have previously applied a *query expansion technique* that augments the syllable-based query representations by recognition confusions [8]. In this paper, we apply a *document expansion* technique that augments the syllable-based document representations with selected terms from a side collection.

2 MONOLINGUAL CANTONESE SPOKEN DOCUMENT RETRIEVAL (BASELINE)

2.1 Vector-space Retrieval Model

We have developed a retrieval engine based on the vector-space model [9]. The query and document term-weighing equations are shown in Equations (1) and (2) respectively:

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

where $tf_q[i]$ is the frequency of term i in query q, N is the total number of documents, and n_i is the number of queries contains term i.

$$d[i] = \frac{\ln(tf_d[i])+1.0}{(1-s)\times length_{doc} + s\times length_{coll\_average}} \quad \text{.. (2)}$$

where $tf_d[i]$ is the frequency of term i in document d, $length_{doc}$ is the length of the current document in terms of byte size of that document and $length_{coll\_average}$ is the average document length across the collection. The value of s ranges between 0 to 1 which controls the proportion of contribution

between $length_{doc}$ and $length_{coll\_average}$. We have set the value of s to 0.45 for our Cantonese SDR experiments to put a little more emphasis on the length of current document than the average length of the documents in the collection. It has previously been shown that the use of document length normalization (DLN) instead of cosine normalization can give better performance when document indexing involves imperfect recognition [10].

The similarity $S(q, d)$ between a query q and document d is measured by the inner product as shown in Equation (3).

$$S(q,d) = q \cdot d \quad \text{.....(3)}$$

2.2 Experimental corpus

The corpus used for our Cantonese SDR experiments is derived from local television news broadcast and constitutes part of the AoE-IT Web Repository [11]. Details are summarized in Table 1.

Content	TV news broadcasts
Period	June 1997 – February 1998
# of stories	1801
Duration	1.5 min. (range: 11 sec. to 24 min.)
Format	Real Media

Table 1. Details related to the Cantonese television broadcast news corpus used in the monolingual SDR experiments.

Each news clip is accompanied by a textual summary and a headline. The summary is a brief description of the news story transcribed by human but it is not a verbatim transcription of the audio track of the corresponding news clip [7]. The headline is a short sentence (about 10 characters) capturing the gist of the news story.

Audio indexing is performed by a Cantonese speech recognizer. The audio tracks are extracted from the news clips and passed to the recognizer for automatic transcription. Details on the audio indexing were described in [7]. The transcriptions are in the form of base syllable sequences and they are used to form document vector representations for our SDR experiments.

2.3 Experiments

The Cantonese SDR experiments focus on the use of textual queries to retrieve relevant spoken documents. The headlines of our news stories are used as textual queries. There are three types of documents derived from our Cantonese broadcast news archive: (i) textual documents (i.e. the textual summaries) in the form of Chinese characters; (ii) documents in the form of syllables converted from (i) by dictionary lookup (hence approximates the case of perfect speech recognition); and (iii) documents in the form of syllables obtained from speech recognition.

We have formulated a known-item-retrieval (KIR) task for our Cantonese SDR experiments. It is assumed that for each of the 1801 queries, there is only one single "relevant" document, i.e. the corresponding story's textual summary of the audio track. The evaluation metric for the KIR task is the average inverse rank (*AIR*) as shown in Equation (4).

$$AIR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad \text{.....(4)}$$

where N is the total number of queries and $rank_i$ is the rank of document i when retrieved using query i.

Previous work [7] found that the best performing indexing unit is the overlapping character/syllable bigram. In this work, our retrieval performance will be reported for this unit only. The textual summaries are represented either as overlapping character bigrams in their document vectors, or as overlapping syllable bigrams where characters are converted to syllables by pronunciation dictionary lookup. Table 2 shows the retrieval results for our baseline experiments.

	Textual Documents	**Text Converted Syllables**	**Recognized Syllables**
AIR	0.830	0.826	0.479

Table 2. Retrieval performances in AIR for three types of documents based on overlapping character/syllable bigrams indexing.

2.4 Analysis

The performance degradation resulting from using overlapping character bigrams (AIR=0.830) to using overlapping syllable bigrams (AIR=0.826) indicates that there is a reduction in lexical knowledge as we migrate from characters to syllables.[1] The drop in retrieval performance when using syllables transcribed by the speech recognizer (AIR=0.479) shows that recognition errors degrade the retrieval performances significantly.

In order to improve the retrieval performance for spoken documents indexed by speech recognition, we attempt to apply some robust retrieval techniques. The query expansion technique reported in [8] augments the query vector by common syllable confusions derived from the recognition confusion matrix. This technique improved retrieval performance slightly, from AIR=0.479 to 0.489. In our current work, we devise a *document expansion* technique that augments the syllable-based document vectors with selected terms from a side collection. The overall goal is to bridge the gap between the query vectors that do not have recognition errors and the document vectors that contain recognition errors.

3 DOCUMENT EXPANSION FOR MONOLINGUAL SDR

Document expansion tries to improve the retrieval performance by augmenting the document vectors with additional indexing terms. These terms are selected from a

[1] In Chinese, more than one character may have the same syllable pronunciation. The use of syllable-based units in place of character-based units will reduce the amount of inherent lexical knowledge.

side textual collection. Hence our approach aims to select expansion terms that (i) do not contain recognition errors and (ii) relevant to the query.

3.1 The Side Collection

Recall from section 2 that each audio news story has a corresponding textual summary. While the textual summary is not a verbatim transcription of the audio, it contains the key information related to the audio news story. We use the batch of textual summaries as our side collection for document expansion. Hence this set of experiments can help gauge an upper bound for the performance of document expansion.

3.2 Document Expansion Procedures

We devised the following procedures for document expansion:

i. Select a single spoken document (audio file) that has been automatically transcribed by the Cantonese syllable recognizer. Syllable bigrams from the transcription are used to form the *query vector*.

ii. All textual summaries (in the side collection) are transformed into syllable representations by pronunciation dictionary lookup. Syllable bigrams formed from these syllable representations are used for building *document vectors*.

iii. The query is used to retrieve a list of documents ranked according to the similarity measure. The top *N (N=1)* retrieved documents are selected for term extraction in the next step.

iv. For each terms in the side collection, we compute its TF-IDF value as shown in Equation (5). All terms extracted from the retrieved document in the preceding step are ranked in descending order of their TF-IDF values. This forms a term list *L*. The more "important" terms are placed towards the top of the list. Then we count the number (*M*) of indexing terms in the spoken document. The top *M* terms in *L* are then extracted and added to the spoken document. This selection avoids over diluting the original terms in the spoken document.

$$TF - IDF = n_{occ} \ln(\frac{N+1}{n_i}) \quad \text{..... (5)}$$

where n_{occ} is the number of occurrences of the term under consideration, N is the total number of documents in the collection and n_i is the number of documents containing this term.

3.3 Results and Analysis

	AIR
Baseline	0.479
With Document Expansion	0.747
Textual Summaries	0.826

Table 3. Performance measured in terms of the average inverse rank (AIR) for our Cantonese monolingual spoken document retrieval task. The performance on textual summaries is for reference with the "perfect speech recognition benchmark". The performance corresponding to document expansion is an approximate upper bound of the effect of document expansion, since the side collection consist of the textual summaries.

Table 3 shows the SDR results with document expansion. Recall that the baseline SDR performance using overlapping syllable bigrams is AIR=0.479. When we incorporate document expansion, AIR improves to 0.747. This improvement is based on the expansion terms from the side collection (and has no recognition errors). In some cases, the recognizer transcribed certain terms with errors, but document expansion added the *error-free* version back into the document. Table 4 provides an example for illustration.

Correct Content	/lap faat wui zyu zik faan ceoi lai taai biu si gaa jyu/ 立法會主席范徐麗泰表示假如
Erroneous Content due to recognition (underlined)	/lap loeng faat wui syut dyut faan ngai hai biu si gaa jyun/ 立兩法會說奪范藝系表示假原
Error-free bigrams introduced through document expansion	/laap/-/faat/ /zyu/-/zik/ /faan/-/ceoi/ /ceoi/-/lai/ /lai/-/taai/ /gaa/-/jyu/ 立法，主席，范徐，徐麗，麗泰，假如

Table 4. Example illustrating the presence of recognition errors in the spoken documents indexed by automatic speech recognition of Cantonese base syllables. However, certain terms that were misrecognized may be have their correct versions introduced back into the document by means of the document expansion technique. The Chinese characters are mapped into their syllable pronunciations by pronunciation lookup.

We also found that during document expansion, 91% of the queries manage to retrieve their corresponding textual summary for term extraction. Hence the remaining queries were expanded incorrectly. Nevertheless, results in Table 3 suggest that document expansion for Cantonese monolingual SDR is promising.

The "perfect speech recognition benchmark" for retrieval performance may be estimated by converting the textual summaries into overlapping base syllable bigrams and retrieving with the same queries. The AIR obtained in this case is 0.826. The observed discrepancy (0.747 versus 0.826) is due to incorrect document expansion as explained above, as well as the erroneous indexing terms that remain in the spoken documents.

4 CROSS-LANGUAGE SPOKEN DOCUMENT RETRIEVAL

We also applied the document expansion technique to a cross-language SDR task to explore its generalizability. In this task, we adopted a query-by-example formulation in which complete English news articles (text) are presented as queries for retrieving relevant Mandarin broadcast news

stories (audio). We have also adopted a query translation approach in which the English queries are translated to Chinese for retrieval. The remaining parts of the retrieval process are similar to monolingual Chinese SDR.

4.1 Experimental corpus

A subset of the Topic Detection and Tracking phase 2 Collection (TDT-2) [12] is used in our experiments. Table 5 shows a summary of data used.

	Audio Document	Textual Query	Side Collection
Source	Voice of America	New York Times, Associated Press	Xinhua, Zaobao
# of stories	2,265	195 (randomly selected)	11,277 + 5,170
Time spans	Mar-Jun, 1998	Jan-Jun, 1998	Jan-Jun, 1998

Table 5. Corpus for the cross-language spoken document retrieval task.

4.1.1 Spoken Document Collection

The spoken documents for our task are Voice of America (VOA) Mandarin broadcasts. The audio has been transcribed by a research grade Mandarin large-vocabulary continuous speech recognition from Dragon [13]. The transcriptions with recognition errors are then segmented into word sequences.

4.1.2 Translingual Queries

Articles from the New York Times (NYT) and Associated Press (AP) newswires are used as queries for retrieval. The 195 queries are divided into 12 batches where each batch contains 16 to 17 queries from different topics. The translated queries from [14] are used as our queries.

4.1.3 Side Collection

The side collection used for expansion is obtained by combining articles from Xinhua and Zaobao in the TDT-2 collection. Both sources are Chinese text (in GB encoding) without explicit word boundaries. The side collection has a total of 16,447 articles. This is approximately seven times that of the size of the VOA corpus and forms a rich collection for document expansion.

4.2 Machine Translated Queries

The queries derived from NYT and AP articles are translated prior to retrieval. We combined phrase-based translation with word-based translation. A bilingual term list with over 200K words is used as the translation dictionary. English words having multiple translations are considered. For example, translations for the query terms "stock exchange of thailand" are illustrated in Table 6.

stock (16)	exchange (26)	of (7)	thailand (2)
股票，股份，原料	交流，交換，兌換 ...	的，關於，之中 ...	泰國，泰

Table 6. Examples of multiple translations in query terms. The numbers inside parentheses are the number of translations for each corresponding word.

4.2.1 Unbalanced Queries

The unbalanced query refers to the formation of a translated query by simply including all translations. For the example in Table 6, each English word in the original query is weighted equally after translation. In an unbalanced query, all translations are weighted equally. Hence a word such as "exchange" which has 26 translations (see Table 6) will be weighted much more heavily by the vector-space model when compared to the word "Thailand" that has only two translations. This effect is undesirable because words with many translations actually introduce more ambiguity and they should not be weighted more heavily.

4.2.2 Balanced Queries

In a *balanced query*, each Chinese translation alternative is re-weighted so that the combined weight will be equal to that of the original English term, as shown in Equation (6).

$$weight_{alternatives} = \frac{1}{no.of\ translation\ alternatives} \quad \text{..... (6)}$$

An example is shown in Table 7 for the translation of the word "Thai".

Terms	Thailand	泰國	泰
Weight	1	0.5	0.5

Table 7. An example of reweighing Chinese translation alternatives in balanced queries.

After translation, the English word "Thailand" becomes two words "泰國" and "泰". According to Equation (6), each alternative should have a weight of 0.5 in this case. Withthis method, each translated alternative from the word "stock", "exchange" and "of" in Table 6 will be equal to 1/16, 1/26 and 1/7 respectively.

When the query vector is formed from subword units, such as overlapping character bigrams, a further level of balancing should be applied. Longer (translated) Chinese terms will generate more bigrams than shorter ones. For example, the word "exchange" with term weight 1/26 translates to the Chinese word "兌換手續費". This produces four overlapping bigrams as illustrated in Table 8.

Word	兌換手續費
Bigrams	兌換，換手，手續，續費

Table 8. An example of overlapping bigrams formed from a translated Chinese term.

In this example, each bigram is further down-weighted by 1/4 to produce the final weight of 1/4 × 1/26 = 1/104. The reweighing function for subword-based queries is shown in Equation (7).

$$weight_{bigrams} = weight_{alternatives} \times \frac{1}{no.of\ bigrams} \quad \text{.. (7)}$$

4.2.3 Results on the Weight Balancing Scheme

We performed CL-SDR experiments with both balanced and unbalanced queries. Results are shown in Table 9. Balancing queries brings a relative improvement of 75.2% in retrieval performance. Therefore, we will focus in using balanced queries for our subsequent experiments.

	mAP
Unbalanced Query	0.294
Balanced Query	0.514

Table 9. The effect of using balanced queries on cross-language spoken document retrieval.

5 DOCUMENT EXPANSION FOR CROSS-LANGUAGE SDR

The document expansion technique is also applied to CL-SDR. The side collection consists of Xinhua and Zaobao news articles. Expansion procedures are similar to those described in Section 3 except for the last step. Since we do not have any relevance judgments between the VOA documents (used as queries in document expansion) and the Xinhua/Zaobao articles (used as the side collection in document expansion), we set up the document expansion in a conservative manner, as will be described in the following.

i. We use tokenized Chinese words as the retrieval unit during document expansion. Tokenization is based on the vocabulary in the VOA documents. Each VOA document is used as a query to retrieve *related* documents from the side collection. The retrieved documents are ranked in descending order according to the degree of match.

ii. We extract expansion terms from the top five documents in the ranked retrieval list. Previous work has stated that it is safer to use only the top few documents for expansion when there is no relevance judgement between the document to be expanded and the documents in the side collection [15].

iii. Expansion term extraction is based on the TF-IDF value computed according to Equation (5). Terms with higher TF-IDF values tend to have stronger discriminative power. We extracted terms whose TF-IDF values are within the mid-to-high range (7 to 12). We found that terms with TF-IDF exceeding 12 tends to be overly specific.

iv. To control the degree of expansion, the maximum number of expansion terms introduced to each spoken document is limited to 80% of the number of terms in the original document. This avoids over-dilution of the original terms in the spoken documents.

5.1 Performance Evaluation and Analysis

Since the TDT collection is manually annotated with relevance judgments, there can be multiple relevant documents for each of the queries. In this context, a suitable evaluation criterion is the mean average precision (mAP) as shown in Equation (8).

$$mAP = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{M_i}\sum_{j=1}^{M_i}\left[\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_k(i,j)}\right]\right\} \quad (8)$$

where L (=17) is the number of topics, M_i (=16 or 17) is the number of exemplars on topic i, N_i is the number of relevant documents on topic i, and $rank_k(i,j)$ is the rank of the k^{th} relevant document for query exemplar j of topic i.

Performance evaluation is based on the 195 translated queries. Retrieval results are shown in Figure 1 for baseline experiments and also that with document expansion.

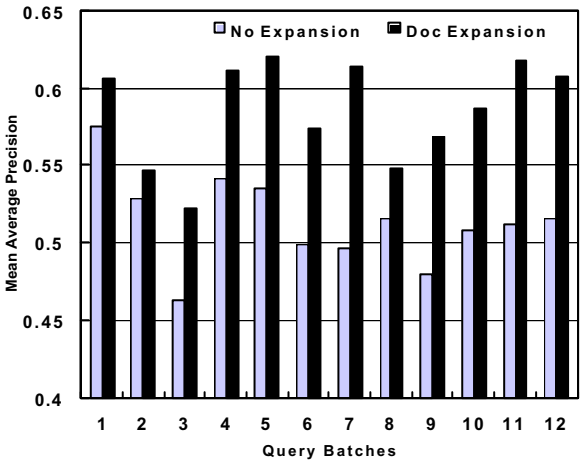

Figure 1. Comparison of the retrieval performance for the baseline system and after document expansion.

Retrieval performance without document expansion gave average precisions between 0.462 and 0.576. The mean average precision (mAP) over 12 query batches is 0.514. With document expansion, we observe consistent improvements across all query batches. Average precision ranges between 0.522 and 0.620 with a relative improvement in retrieval performance is 13.8%. Table 10 shows the CL-SDR results of document expansion.

	mAP
Baseline	0.514
With Document Expansion	0.585

Table 10. Mean average precisions over 12 query batches with and without document expansion for the cross-language spoken document retrieval task.

Table 11 illustrates how document expansion help improve retrieval performance. The name "阿爾及利亞" and the verb "炸死" have been misrecognized as "鮑爾激勵要 "and "扎斯" respectively. With document expansion, the *error-free* versions were extracted during expansion and added to the spoken document to bring about correct retrieval.

Erroneous Contents	鮑爾激勵要交區發生了爆炸事件扎斯了兩個人
Added Error-free Terms	阿爾及利亞，炸死

Table 11. Example illustrating misrecognized terms in the spoken document and the extraction of the *error-free* versions during document expansion. The expansions terms are then added back to the spoken document to bring about correct retrieval.

6 CONCLUSIONS

This paper reports on an attempt to use document expansion to improve performance in spoken document retrieval. Our monolingual and cross language retrieval tasks involve textual queries and spoken documents. The translated textual queries are mapped into syllables by pronunciation dictionary lookup. Hence the query syllables have no recognition errors. The spoken documents are indexed automatically by speech recognition and hence the indices contain recognition errors. In order to bridge this gap between queries and documents, we apply document expansion in which terms are selected from a side collection and added to the spoken documents. This procedure aims to correct misrecognized terms that are wrongly introduced in the indexed spoken documents during automatic transcription. Document expansion aims to extract the *error-free* versions of such terms from the side collection and add the terms into the indexed spoken documents. Other expansion terms deemed relevant to the topic of interest are also added. Our SDR experiments on both monolingual and cross-language tasks show that document expansion led to relative improvements of 56% and 14% respectively.

ACKNOWLEDGMENTS

We thank the Television Broadcasts (HK) Limited for providing the Cantonese news video content. This work is partially supported by a grant from the Hong Kong SAR Government Research Grants Council (Project No. CUHK 4223/01E) and a grant from the Hong Kong SAR Government University Grant Council under Area of Excellence in Information Technology. We also thank Professor Gina-Anne Levow and Dr. Wai-Kit Lo for their comments and suggestions on this work.

References

[1] Oard, D. and Dorr, B., "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., Scheytt, P., Wactlar, H., 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., Voorhees, E., Auzanne, C., Stadford, V. and Lund, B., "1998 TREC-7 Spoken Document Retrieval Track Overview and Results", *Proceedings of the 7th Text Retrieval Conference*, 1998.

[4] Ng, K. and Zue, V., "Towards Robust Methods for Spoken Document Retrieval," *Proceedings of the 5th 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] 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 the 2000 International Conference in Acoustics, Speech, and Signal Processing,* vol. 3, pp. 1771-1774, 2000

[7] Meng, H., Lo, W. K., Li, Y. C. and Ching, P. C., "Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval", *Processings of 6th International Conference on Spoken Language Processing*, vol 4, pp. 101-104, Beijing, 2000.

[8] Li, Y. C., Lo, W. K., Meng, H., Ching, P. C., "Query Expansion Using Phonetic Confusion for Chinese Spoken Document Retrieval", *Proceedings of the 5th International Workshop on Informa*tion *Retrieval with Asian Languages,* pp. 83-93, Hong Kong, 2000.

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

[10] Singhal, A., Buckley, C., and Mitra, M., "Pivoted Document Length Normalization", *Proceedings of the 19th International Conference on Research and Development in Information Retrieval,* pp. 21-29, 1996.

[11] Area of Excellence in Information Technology (AoE-IT) Web Repository http://www.se.cuhk.edu.hk/~aoe/.

[12] Linguistic Data Consortium, Topic Detection and Tracking phase 2 collection (TDT-2), http://www.ldc.upenn.edu.

[13] Dragon System Inc. Dragon Systems Developer's Den. http://developer.dragonsys.com.

[14] Meng, H., Chen, B., Grams, E., Khudanpur, S., Lo, W. K., Levow, G., Oard, D., Schone, P., Tang, K., Wang, H. M., and Wang, J. Q., "Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval," *Final Report of the Mandarin-English Information group at Workshop 2000, Center for Language and Speech Processing, Johns Hopkins University*, Baltimore, Maryland, USA, 2000.

[15] Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. "Analysis of a Very Large AltaVista Query Log", *Technical Report, Digital SRC Technical Note 1998-014*, October 1998.