Multi-Scale Document Expansion in English-Mandarin Cross-Language Spoken Document Retrieval

Wai-Kit Lo[1], Yuk-Chi Li[1], Gina Levow[2], Hsin-Min Wang[3] and Helen Meng[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]University of Chicago, Chicago, IL, USA
[3]Academia Sinica, Taiwan, ROC
{wklo,ycli,hmmeng}@se.cuhk.edu.hk, levow@cs.uchicago.edu, whm@iis.sinica.edu.tw

Abstract

This paper presents the application of document expansion using a side collection to a cross-language spoken document retrieval (CL-SDR) task to improve retrieval performance. Document expansion is applied to a series of English-Mandarin CL-SDR experiments using selected retrieval models (probabilistic belief network, vector space model, and HMM-based retrieval model). English textual queries are used to retrieve relevant documents from an archive of Mandarin radio broadcast news. We have devised a multi-scale approach for document expansion – a process that *enriches* the Mandarin spoken document collection in order to improve overall retrieval performance. A document is expanded by (i) first retrieving related documents on a character bigram scale, (ii) then extracting word units from such related documents as expansion terms to augment the original document and (iii) finally indexing all documents in the collection by means of character bigrams and those expanded terms by within-word character bigrams to prepare for future retrieval. Hence the document expansion approach is multi-scale as it involves both word and subword scales. Experimental results show that this approach achieves performance improvements up to 14% across several retrieval models.

1. Introduction

With the advent of multimedia and Internet technologies, there is an increasing amount of multimedia information available. Examples include recordings of broadcast news, conferences, meetings, and presentations. In addition, these sources of information may be in different languages. In order to search for relevant information from these archives effectively, automatic retrieval is highly desirable. Cross-language spoken document retrieval (CL-SDR) is the technology that enables efficient search for relevant information from archives of speech recordings based on a user's request specified in a different language.

Previous work in cross-language speech retrieval includes retrieval from Serbo-Croatian using English [1], from German using French [2], and from Mandarin using English [3]. The MEI project [4] has further investigated in a multi-scale paradigm for retrieving Mandarin broadcast news using English queries. In this work, we will explore the application of multi-scale document expansion techniques to CL-SDR using different retrieval models.

1.1. Multi-scale units

Multi-scale refers to the use of word and subword indexing units. Words carry lexical meaning, and they are used as indexing terms for retrieval in many languages. In Chinese, there is no explicit word delimiter, and the definition of word is ambiguous. Hence character bigrams are applied to Chinese textual information retrieval.

In spoken document retrieval, spoken documents are automatically transcribed into textual representations using speech recognition. Since a pre-defined vocabulary is used in speech recognition, any word unknown to the recognizer will be treated as out-of-vocabulary (OOV). In the transcription, OOV words may either be deleted or replaced by other in-vocabulary words. Hence retrieval on the word scale is susceptible to the OOV problem. Previous work [5] has proposed using phoneme n-gram units for retrieval to handle OOV since phonemes can provide full phonological coverage. Relevance feedback [6] has also been used for SDR so that other useful terms can be included to compensate for those OOV words. For the Chinese language, since full textual coverage can be obtained using character-based units, the use of character bigrams for retrieval can circumvent the problem of OOV words. In this work, both word and subword scale indexing units have been used for our CL-SDR task.

1.2. The MEI project

The MEI (Mandarin-English Information) CL-SDR project started in the Johns Hopkins University 2000 summer workshop. It investigated the use of a multi-scale approach for English-Chinese spoken document retrieval – where English news articles are used as query examples to retrieve Mandarin broadcast news audio. Queries and documents are represented on both the word scale and subword scale. Retrieval proceeds on both scales being coupled together [4]. On the document side, recordings of the Mandarin broadcast news are first transcribed by automatic speech recognition [7] and then indexed on both word and subword scales. On the query side, single word and multi-word terms are identified in the query exemplars. Terms are then translated into Chinese based on the CETA [8] dictionary and Linguistic Data Consortium's Chinese-English term list. For terms with multiple Chinese translations, the "balanced" query formulation [9] is performed to average the contributions of different translations. Within the MEI framework, further investigations have also been carried out, including the use of N-gram features [10], an HMM-based CL-SDR model [11] and document expansion [12] [13].

1.3. Document expansion

Document expansion techniques have been successfully applied to spoken document retrieval [14] to overcome the effects of misrecognitions and OOV terms [6]. These techniques attempt to enrich a document by augmenting it with related terms identified from a side collection (also known as expansion collection). The enrichment aims to

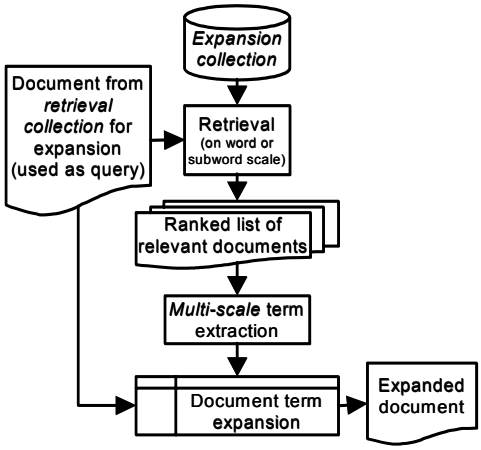

Figure 1. The block diagram illustration of multi-scale document expansion process.

improve retrieval performance by increasing the possible degree of match between document and query terms.

In this work, we attempt to devise and apply a *multi-scale document expansion* approach to CL-SDR. Documents are expanded with terms selected based on information from both the word and subword scales. The related terms added through document expansion can potentially improve retrieval performance not only by alleviating the OOV problem, but also by compensating for translation gaps or differences in lexical choice.

Figure 1 summarizes the general process of document expansion. Each document from the targeted collection (*retrieval collection*) is used as a query for retrieving related documents from an *expansion collection*. The expansion collection can either be a side collection or the retrieval collection itself. The retrieval process returns a ranked list of relevant documents. Multi-scale term extraction is then carried out using the top-ranking documents. Terms extracted can either be words, subword bigrams, or subword bigrams formed from selected words. Finally, the document is augmented with the selected terms.

2. Experimental Setup

Our CL-SDR experiments with document expansion use the queries formed by the MEI project and documents from the TDT-2 collection. Retrieval performance is evaluated using a modified version of mean average precision. In order to investigate the robustness of the document expansion approach, our retrieval experiments employ three popular retrieval models – probabilistic belief network, vector space model, and HMM-based retrieval model. Details of the experimental setup will be given in the following subsections.

2.1. Topic Detection and Tracking phase 2 (TDT-2) collection

The query exemplars and document collections used in this work are selected from the Topic Detection and Tracking evaluation collections (TDT) [3]. The TDT multi-lingual collection includes English and Chinese newswire texts as well as Mandarin (audio) broadcast news. Most of the audio data are furnished with word transcriptions produced by the Dragon automatic speech recognition system [7]. All news stories are tagged with topic labels, which serve as the relevance judgments for performance evaluation of our CL-SDR work.

2.1.1. Query exemplars

Query exemplars are documents that are used as query in retrieval experiments. The English news articles selected as query exemplars from the TDT-2 collection are contemporaneous with the retrieval collection. Details of the query exemplars are given in Table 1.

2.1.2. Retrieval collection

Retrieval collection is the document collection to be searched. The retrieval collection used in this work is Mandarin broadcast news from TDT-2. Details of the retrieval collections are also given in Table 1.

Table 1. Summary of documents selected from the TDT collections for use as queries and documents.

	Query	Document
Source	English text (NYT or APW newswire)	Mandarin audio (VOA news)
Quantity	17 topics, each with variable no. of exemplars	2265 manually segmented stories, 46 hours of audio
Period	Mar-Jun, 1998	Jan-Jun, 1998

2.1.3. Expansion collection

For document expansion, we have made use of two Chinese newswire text corpora from TDT-2 as a side collection. The side collection is also contemporaneous with the retrieval collection to promote topical coherence. Details of the side collection are summarized in Table 2.

Table 2. Summary of documents selected from the TDT collections for document expansion.

	Side collection	
Source	Chinese text (Xinhua news)	Chinese text (Zaobao news)
Quantity	11,277 articles	5,170 articles
Period	Jan-Jun, 1998	Jan-Jun, 1998

2.2. Evaluation measure

The evaluation measure is a modified non-interpolated mean average precision (mAP). It is modified in order to take into consideration the number of topics and variable number of exemplars for each topic. The following equation summarizes the modified mAP.

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

L: no. of topics (17 in this work)
M_i: no. of exemplars on topic i (up to 12 in this work)
N_i: no. of relevant documents on topic i
$rank_k(i,j)$: rank of kth relevant document for query exemplar j of topic i

2.3. Retrieval Engines

2.3.1. Probabilistic belief network (BN)

The InQuery retrieval engine employs a probabilistic belief network framework and was developed by the University of Massachusetts [15].

2.3.2. Vector space model

We have implemented the vector space model based on [16]. Query and document terms are weighted according to Equations (2) and (3) respectively.

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

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 documents with term i.

$$d[i] = \frac{\ln(tf_d[i]) + 1.0}{(1 - slope) \times length_{coll\_average} + slope \times length_{doc}} \quad (3)$$

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 *slope* ranges between 0 to 1 which controls the proportion of contribution between $length_{doc}$ and $length_{coll\_average}$.

We have set the value of *slope* in Equation (3) to 0.5 to place equal importance between the current document length and the average document length. The use of document length normalization instead of cosine normalization has been shown to give better performance when document indexing involves imperfect recognition [17].

The similarity $S(q, d)$ between a query vector q and document vector d is measured by the inner product.

$$S(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} \quad \text{..... (4)}$$

2.3.3. HMM-based retrieval model

For the HMM-based retrieval experiments, we have implemented an HMM-based retrieval model augmented by a general language model [18]. The equation for this model is shown below.

$$p(Q \mid D_i) = \prod_{q_j \in Q} [\alpha \cdot p(q_j \mid D_i) + (1 - \alpha) \cdot p_{glm}(q_j)] \quad (5)$$

where Q is the query containing terms q_j, D_i is the ith document in the collection, α is the weight for the document model; 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 Equations (6) and (7) respectively.

$$p(q_j \mid D_i) = \frac{count(q_j \in D_i)}{count(all\ terms\ in\ D_i)} \quad \text{..... (6)}$$

$$p_{glm}(q_j) = \frac{count(q_j\ in\ all\ documents)}{count(all\ terms\ in\ all\ documents)} \quad \text{..... (7).}$$

The value of α in Equation (5) is obtained empirically by searching from 0 to 1 in steps of 0.1 and the optimal value is found to be around 0.3.

3. Multi-Scale Document Expansion

Each document (from VOA) is used as a query to retrieve *related* documents from the expansion collection. We extract expansion terms from the top five documents[1] in the ranked retrieval list.

Expansion terms may be words or character bigrams and they are ranked according to the TF-IDF (term frequency * inverse document frequency) value computed according to Equation (8). 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 tend to be overly specific.

$$TF - IDF = n_{occ} \ln\left(\frac{N+1}{n_i}\right) \quad \text{..... (8)}$$

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

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.

3.1. Multi-scale expansion procedure

In this work, we extend the multi-scale paradigm [19] to multi-scale document expansion. Both word and subword scale document expansion will be described below. Retrieval from the expansion collection is performed using InQuery to enable direct comparison.

3.1.1. Document expansion based on words

For document expansion on the word scale, we use tokenized Chinese words as the retrieval unit during document expansion. Word tokenization for both the retrieval collection and the expansion collection is performed by a maximum match procedure that references the vocabulary of the VOA documents. Each document from the retrieval collection will be used as a query to retrieve documents from the expansion collection on word scale. After retrieval, terms on the word scale are extracted from the top-ranking documents as described in Section 3.1 and added to the document to be expanded.

3.1.2. Document expansion based on character bigrams

For document expansion on the subword scale, we use overlapping character bigrams as the retrieval unit during document expansion. In this case, documents from both the retrieval collection and the expansion collection are converted into overlapping character bigrams. Retrieval is then performed on the subword scale for each of the documents from the retrieval collection over the expansion collection.

After retrieval, the document is then expanded by adding extracted overlapping character bigrams. The overlapping character bigrams for expansion can be obtained in two different ways: 1) using character bigrams directly from the retrieved expansion collection documents (raw overlapping character bigrams). 2) using character bigrams created within tokenized words from the retrieved side collection documents (overlapping character bigrams within word).

[1] Previous work [20] indicated that it is safer to use only the top few documents for expansion when there is no relevance judgment between the document to be expanded and the documents in the side collection.

4. Results and Discussion

Retrieval experiments are performed without document expansion (baseline) and with expansion. Table 3 summarizes our experimental results.

Table 3. Comparison of CL-SDR performance to that obtained with document expansion. Relative improvements over baseline are given in parentheses.

	Baseline		With document expansion		
	character bigrams	words	using raw character bigrams	using character bigrams within words	using words
Probabilistic BN	**0.497**	**0.451**	**0.560 (12.6%)**	**0.569 (14.5%)**	**0.499 (10.6%)**
VSM	**0.540**	**0.483**	**0.540 (0%)**	**0.579 (7.2%)**	**0.507 (5.0%)**
HMM	**0.490**	**0.471**	**0.515 (5.1%)**	**0.540 (10.2%)**	**0.494 (4.9%)**

From Table 3, it can be seen that document expansion improves retrieval performance for most cases. Among the different expansion approaches, the best performance is obtained from multi-scale document expansion where the character bigrams for expansion are extracted from within words. The relative improvement over baseline character bigram retrieval performance is between 7.2% and 14.5%. The advantage of this expansion strategy is that the new character bigrams are derived from lexically meaningful words. Expanding documents with these terms can avoid introduction of "noisy" character bigrams such as those formed across word boundaries.

Another observation is that expansion on the word scale has the weakest performance. This can be attributed to poorer retrieval on the word scale than on the subword scale. As a result, the retrieval from the expansion collection returns less useful documents (when compared to retrieval using character bigrams) and hence fewer useful expansion terms.

As seen in Table 3, retrieval with document expansion in the VSM provides the least improvement in retrieval performance. A major reason for this is the higher baseline performance obtained from VSM. Among our experiments, the best retrieval performance (0.579) is achieved using VSM with character bigrams extracted within words. Another reason for less improvement using VSM is believed to be due to the introduction of a different retrieval engine (InQuery) for expansion. In order to look into this issue, we have also performed document expansion retrieval using the VSM and achieved improved retrieval performance, where raw character bigrams yield mAP of 0.561, character bigrams, 0.591, and words, 0.528. The best relative improvement is 9.4%.

5. Conclusion

Our experimental results demonstrate that document expansion yields consistent improvements in retrieval performance. For error-free documents, document expansion can enrich the documents with relevant terms to enhance the retrieval performance. Furthermore in SDR, those documents with OOV-related recognition errors can benefit from the addition of relevant terms. In addition, CL-SDR can also benefit from the introduction of terms that match the translated query. When document expansion is performed using character bigrams within words, our CL-SDR experiments show consistent relative improvement in retrieval performance from 7% to 14% for the investigated retrieval models.

6. References

[1] A. G. Hauptmann, et. al., "Multilingual Informedia: a demonstration of speech recognition and information retrieval across multiple languages," in *Proceedings of 1998 Broadcast News Transcription and Understanding Workshop*, 1998.

[2] P. Sheridan, et. al., "Cross language speech retrieval: establishing a baseline performance," in *Proceedings of SIGIR1997*, pp. 99–108, 1997.

[3] Topic Detection and Tracking, http:// www. nist. gov / speech / tests/ tdt/ index.htm

[4] H. M. 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.

[5] K. Ng, "Towards robust methods for spoken document retrieval," in *Proceedings of ICSLP1998*, Sydney, 1998.

[6] P. C. Woodland, "Effects of out of vocabulary words in spoken document retrieval," in *Proceedings of SIGIR2000*, pp. 372-374, 2000.

[7] P. Zhan, et. al., "Dragon Systems' 1998 broadcast news transcription system for Mandarin," in *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[8] Chinese-English Translation Assistance (CETA), licensed as Optilex by MRM Inc. 3910 Knowles Avenue, Kensington, MD 20895, U.S.A.

[9] A. Pirkola, "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval,", in *Proceedings of SIGIR1998*, pp55-63, 1998.

[10] Berlin Chen, et. al., "An HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval", in *Proceedings of Eurospeech2001*, pp 1045-1048, 2001

[11] W. K. Lo, "Information fusion for monolingual and cross-language spoken document retrieval," *Ph.D. Thesis, the Chinese University of Hong Kong*, 2000

[12] Gina A. Levow, "Multi-scale document expansion for Mandarin Chinese," in *Proceedings of the ISCA workshop on multilingual spoken document retrieval*, pp. 73-78, 2003.

[13] Y. C. Li and H. Meng, "Document expansion using a side collection for monolingual and cross-language spoken document retrieval," in *Proceedings of the ISCA workshop on multilingual spoken document retrieval*, pp. 85-90, 2003.

[14] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proceedings of SIGIR1999*, pp. 34-41, 1999.

[15] J. P. Callan, et. al., "The INQUERY Retrieval System," in *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78-83.

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

[17] A. Singhal et al., "Pivoted document length normalization," in *Proceedings of SIGIR1996*, pp. 21-29, 1996.

[18] A. Berger and J. Lafferty, "The Weaver system for document retrieval", *Proceedings of the 8th Text Retrieval Conference,* pp. 163-174, 1999.

[19] H. Meng, et. al., "Multi-scale audio indexing for Chinese spoken document retrieval," in *Proceedings of ICSLP2000*, pp. 101-104, 2000.

[20] C. Silverstein, et. al. "Analysis of a very large AltaVista query log", Technical Report, *Digital SRC Technical Note 1998-014*, October 1998.