Multi-Scale Retrieval in MEI: An English-Chinese Translingual Speech Retrieval System

Wai-Kit LO[1], *Patrick SCHONE*[2] *and Helen MENG*[1]

[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

[2]Department of Defense, 7715 Patuxent Oak Ct., Elkridge, MD 21075, USA

wklo@ee.cuhk.edu.hk, pjs500@afterlife.ncsc.mil, hmmeng@se.cuhk.edu.hk

Abstract

This paper presents a multi-scale retrieval approach in MEI (Mandarin-English Information), an English-Chinese cross-lingual spoken document retrieval (CL-SDR) system. It accepts an entire English news story (from newspaper text) as the input query, and automatically retrieves "relevant" Mandarin news stories (from broadcast audio). This allows the user to search for personally relevant content across the language and media barriers – a *cross-lingual* and *cross-media* retrieval task. MEI advocates a *multi-scale paradigm* for the retrieval task. Multi-scale refers to the use of *both* words and subwords (Chinese characters and syllables) for retrieval. Words offer lexical knowledge to enhance precision, and subwords can potentially alleviate some prevailing problems in CL-SDR, e.g. open vocabularies in translation and recognition, out-of-vocabulary words in audio indexing, and ambiguities in Chinese homophones and word tokenizaiton. We present techniques for word-subword fusion, which improved retrieval performance in our experiments with the Topic Detection and Tracking collection.

1. Introduction

This paper presents a multi-scale retrieval approach for cross-lingual and cross-media information retrieval. We focus on the use of English textual queries to retrieve Mandarin spoken documents, i.e. and English-Chinese cross-lingual spoken document retrieval (CL-SDR) task. With the growing multi-media and multi-lingual content in the global information infrastructures, CL-SDR technologies are potentially very powerful, as they enable the user to search for personally relevant audio content, (e.g. recordings of meetings, lectures or radio broadcasts), across the barriers of language and media.

MEI (Mandarin-English Information) [1] is an English-Chinese CL-SDR system developed during the Johns Hopkins University Summer Workshop 2000. Mandarin is the key dialect of Chinese. MEI accepts an entire English textual story (from newspapers) as the input query, and automatically retrieves "relevant" Mandarin audio stories (from radio broadcasts). English and Chinese are two predominant languages used by the global population. They are very different linguistically, hence English-Chinese CL-SDR presents unique research challenges.

MEI advocates a multi-scale paradigm for the English-Chinese CL-SDR task. This includes multi-scale query processing, multi-scale audio indexing and multi-scale retrieval. The multi-scale paradigm includes word-based retrieval, since words possess lexical knowledge which enhances precision. However, the paradigm also includes subword-based retrieval, which aims to alleviate problems related to English-Chinese CL-SDR, such as:

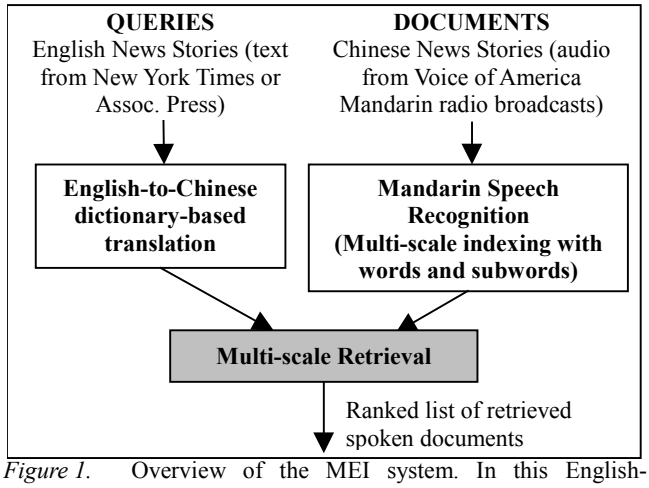

Figure 1. Overview of the MEI system. In this English-Chinese spoken document retrieval task, the query is formed from an entire English news story (text) from the New York Times or Associated Press. The spoken documents are Mandarin news stories (audio) from Voice of America news broadcasts. Multi-scale retrieval of the spoken documents (shaded box) is the focus of this paper. System performance is evaluated based on the relevance of the ranked list of spoken documents retrieved for each query [2].

Multiplicity in translation – dictionary-based term-by-term translation may produce multiple translation alternatives, or no translations, e.g. proper names. The use of phrases can often resolve translation ambiguity, e.g. “human rights”. The use of phonetic transliteration can help address the out-of-vocabulary problem in translation, e.g. Kosovo becomes /ke suo fu/ (科索沃), and the pinyin transcription can be utilized for SDR.

Open vocabulary in recognition – indexing spoken documents with word-based speech recognition is constrained to the recognizer’s vocabulary. Out-of-vocabulary words cannot be indexed. Since the Mandarin Chinese can be fully represented by about 400 base syllables or 6000 characters, we can obtain full phonological / lexical coverage of the spoken documents using syllables / characters for indexing.

Ambiguity in Chinese homophones – each Chinese character is pronounced as a single syllable, and the mapping is many-to-many. Hence there is a large number of Chinese homophones, which can cause word-level confusions in SDR. This problem can be addressed by retrieval based on syllables.

Ambiguity in Chinese word tokenization – the Chinese word contains one to multiple characters, with no word delimiter. Word tokenization has much ambiguity, which can cause word-level mismatches between queries and documents

in retrieval. This problem can be addressed by retrieval based on characters.

This work describes our *word-subword fusion strategies* to facilitate multi-scale retrieval for English-Chinese SDR. To the best of our knowledge, the MEI system is the first-of-its-kind in English-Chinese SDR. However, we draw from the vast experiences in previous work such as English-French CL-SDR [3], monolingual Chinese SDR [4,5], English subword-based SDR [6,7] and monolingual retrieval using multiple indexing sources [8,9].

2. The TDT Collection

2.1 Topic Detection and Tracking

Topic Detection and Tracking (TDT) is a DARPA-sponsored program where participating sites tackle tasks such as identifying the first time a story is reported on a given topic; or grouping similar topics from audio and textual streams of newswire data. In recent years, TDT has focused primarily on performing such tasks in both English and Mandarin Chinese. The task that we tackle in the MEI project is not part of TDT, but the TDT collection serves as a valuable resource for our work. Therefore, for the sake of the reader, we include a brief description of that portion of TDT that we apply to the task of English-Chinese CL-SDR.

2.2 Collection

The TDT multi-lingual collection includes English and Mandarin news texts as well as (audio) broadcast news. Most of the audio data are furnished with word transcriptions produced by the Dragon automatic speech recognition system [10]. All news stories are tagged with topic labels, which serves as the relevance judgments for performance evaluation of our CL-SDR work. We use the TDT-2 corpus as our development set, and TDT-3 as our evaluation set. Table 1 describes the content in these collections.

	TDT-2 (Dev set)	TDT-3 (Eval set)
English NYT/AP news text (queries)	17 topics, variable # of exemplars	56 topics, variable # of exemplars
Mandarin VOA audio (docs)	2265 stories, 46.0hrs of audio	3371 stories, 98.4hrs of audio

Table 1 Statistics of TDT-2 and TDT-3: our development and evaluation data sets. The Mandarin audio documents are furnished with recognized words from the Dragon system [2].

3. Multi-Scale Retrieval

3.1 Words and Subwords

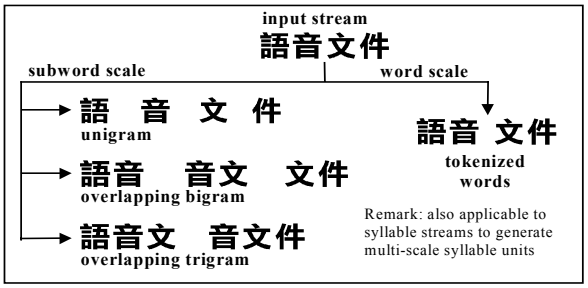

Figure 2. Multi-scale units for document retrieval. The example Chinese phrase is made up from 2 two-character words that means "spoken document".

As shown in Figure 1, English news stories are used as query exemplars for a topic. Key terms are selected and then translated to Chinese using a hybrid of phrase-translation and dictionary-based term-by-term translation [1]. Our audio documents are indexed according to Dragon's word hypotheses. Therefore, a straightforward method for retrieval is to match the translated Chinese words in the query with the recognized words from the audio documents.

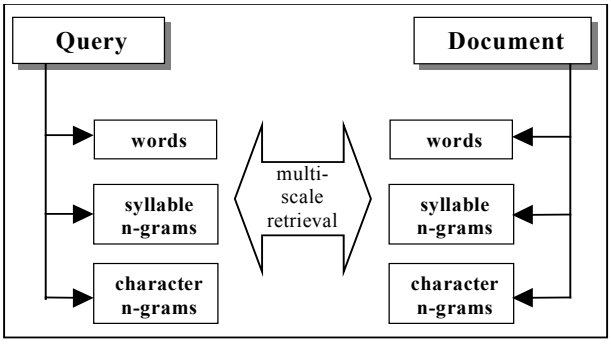

Figure 3. Multi-scale retrieval for Chinese spoken documents.

In this work, we performed *multi-scale retrieval*. This is illustrated in Figure 3. In addition to using the word unit for retrieval, we include the subword units as well, as motivated in Section 1. Indexing with subwords refers to the use of overlapping character n-grams, and syllable n-grams. We mainly use bigrams since previous work indicated that they are the most effective. We have also included trigrams in some of our experiments.

3.2 Retrieval Engine

We use InQuery as our retrieval engine. It is a state-of-the-art information retrieval engine developed by the University of Massachusetts [11]. InQuery employs a probabilistic belief network as the main data structure behind its querying capability. This belief network allows users to build various kinds of queries and have these be evaluated according to different probabilistic paradigms.

The paradigm that we make particular use of is the "balanced query" mechanism. For InQuery oficionados, this is referred to as the #sum operator. Suppose that we had a query given by #sum(T_1 T_2 ... T_n), where the T_i's represent terms. For a given document *D*, we denote the belief that T_i is satisfied by *D* as P_i. The balanced query (or #sum operator) suggests that if one wants to know the belief that *D* satisfies the query, they would need merely to compute the average of the P_i's. This function is particularly desirable when using cross-lingual retrieval. For example, if one does not know the proper Chinese translation for a given English term, one could wrap the #sum operator around a collection of possible translations, which indicates to InQuery that it should simply take the average. We will also show later that #sum can be beneficial for coupling.

3.3 Evaluation Criterion

In order to evaluate our retrieval performance, we use the non-interpolated mean average precision, the same metric adopted at Text Retrieval Conference (TREC).

The non-interpolated mean average precision is computed as follows: For a given query and its ranked list of retrieved documents, we proceed from the top downwards and calculate the precision for every relevant document retrieved. The average of all the precision is the average precision for that particular query. Taking another average over all queries

produce a single value as our evaluation metric. The following equation summarizes the process:

$$mAP_{\text{non-interpolated}} = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{\frac{1}{N_j}\sum_{k=1}^{N_j} precN_{Q_j}(k)\right\}\right\}$$

where $mAP_{\text{non-interpolated}}$ is the non-interpolated mean average precision, $precN_{Q_j}(j)$ is the precision for Q_j after j relevant documents are retrieved, L is the total number of batches, M is the total number of queries and N_j is the total number of relevant documents for query Q_j.

4. Multi-Scale Fusion Strategies

The various word and subword units contribute in different ways towards our CL-SDR task. In this work, we explore two major strategies in fusing the multi-scale units for retrieval. The first strategy is to perform retrieval for every scale individually, and obtain a ranked list of retrieved documents in each case. Then all the ranked lists are combined together – this is termed as *loose coupling*. The second strategy is to integrate the multi-scale units in the query / document representations prior to retrieval, which then produces a single ranked list – this is termed as *tight coupling.*

4.1 Loose Coupling

For loose coupling, we need to integrate multiple ranked lists of retrieved documents, each obtained by using a single type of unit for retrieval. Each entry in the ranked list also has a corresponding score, which is a value reflecting the similarity between the query and the retrieved document. This similarity measure can be the retrieval *score*, or the *rank* of the retrieved document. We can then apply integration functions such as summation, maximum-of, etc. to these similarity scores. In the MEI project, we adopted the linear combination, defined by the equation below. The merged score is then used to re-rank all retrieved documents to produce a final ranked list.

$$S_{merged}(Q_i, D_j) = \sum_{k \in K} w_k S_k(Q_i, D_j)$$

where w_k is the weighting factor for unit k, K is the set of multi-scale units, S_k is the score for query i retrieving document j and S_{merged} is the merged score.

4.2 Tight Coupling

The notion behind loose coupling is straightforward and, given that there is a development set for training each w_k, it may provide significantly better performance than any single system. Although we do have such a development set, one might still like to experiment with coupling approaches that do not rely upon a development set. Due to the balanced query capability of InQuery, tight coupling may serve as an appropriate alternative.

This is best explained by example. Suppose that an English query contained the phrase "prime minister." This phrase would be translated to the Chinese phrase 以色列首相, pronounced */yi-se-lie shou-xiang/*. It has already been explained that we could query using the individual words themselves or using overlapping character or syllable *n*-grams. However, could we query both words, and say, syllable bigrams at the same time?

To do this, we would first need to set up the document structure so that it could handle either format. For every spoken document *S*, we need to create a S*'* which is double the size of *S*, where the first half of *S'* contains only the word-level representation of *S* and the second half contains the syllable bigram representation. Next, we formulate the queries so as to take simultaneous advantage of both words and syllables. Using InQuery's balanced query mechanism, we can form the query:

#sum(以色列,#sum(yi-se se-lie)) #sum(首相,#sum(shou-xiang))

Recall that the #sum operator signifies that the belief that a given document satisfies the query is the average of the beliefs that it satisfies the individual components. If we think of the #sum as a voting procedure, then InQuery is effectively being told that syllable bigrams get one agglomerative vote for their choice of documents to return, the corresponding word gets a vote unto itself, and both representations cast a single, collective vote for the documents they believe best satsify the term in general.

5. Experimental Results

5.1 Loose coupling

We chose to loosely couple retrieval ranked lists based on the word, character bigram and syllable bigram, since these are the units which performed well individually [1,2]. We experimented with *score-based loose coupling* (as mentioned in Section 4.1).

To optimize the weights, we tried all combinations within the constraint of (sum of weights = 1). Results are shown in *Figure 4*.

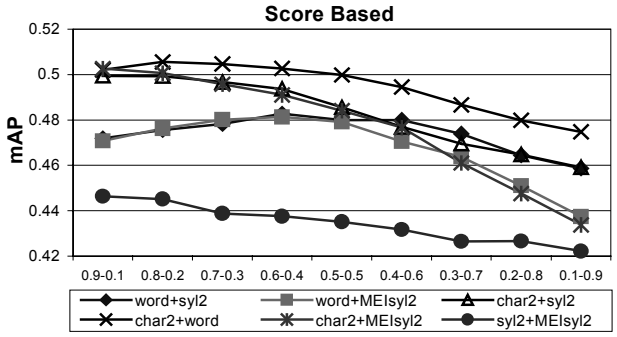

Figure 4. Score-based loose coupling weighting parameters sweeping result (Note: "char2" means character bigram, "syl2" means syllable bigram and MEIsyl stands for recognition results from the audio data using MEIsyllable recognizer [2]).

After the sweeping experiments, we have selected the best weighting values for the loose coupling and applied to the selected retrieval runs. The results are shown in Figure 5.

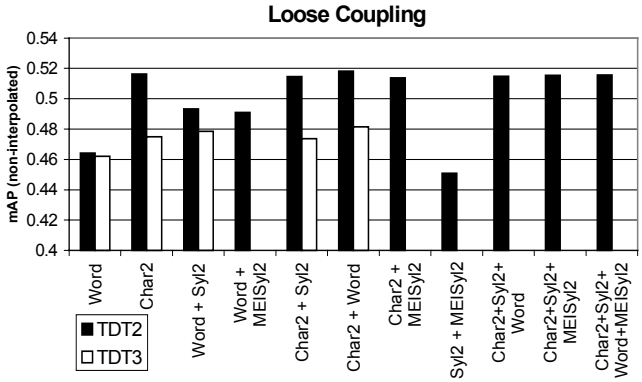

Figure 5. Loose coupling results from integrating different word and subword scale units.

Our loose coupling results (see Figure 5) show that the best retrieval performance is obtained by combining the word and character bigram. We observe a significant drop in performance when syllable bigram and MEIsyllable bigram are coupled. It is believed to be due to the reason that the information provided is mostly redundant.

5.2 Tight coupling

Our hypothesis for tight coupling was that there would be no need for using a development set in order to couple multiple query representations. Therefore, the results we provide below for TDT3 have *not* been optimized using information from TDT2. The coupled retrieval results are shown in Figure 6.

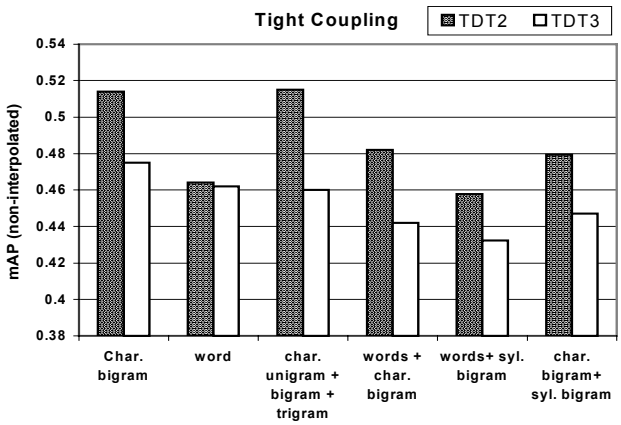

Figure 6. Tight coupling results for different indexing units.

The first two categories of Figure 6 represent performance for words and character bigrams alone. The remaining columns indicate different coupling combinations. Much to our dismay, there were no modes of tight coupling that even attained the lower performance of words alone. Even when tight coupling was run on TDT2, performance never exceeded the best representation being coupled.

6. Discussions

Our loose coupling results show that by using optimized weights (from the development set), as well as the indexing units that perform well individually, we can obtain improved retrieval performance over the individual runs.

Our tight coupling results indicate that our initial assumption that one could query simultaneously across multiple input representations without the need for weighting was incorrect. As we saw in Figure 6, tight coupling results almost never exceeded those of the best single approach (namely, character bigrams). InQuery has an additional operator, #wsum, which makes possible the ability to compute beliefs using weighted averages. We could perhaps attempt a brute-force approach to use #wsum and learn weightings as we had done with loose coupling. However, due to time limitations, we have not conducted extensive study on this nor on other alternative approaches.

7. Conclusions

We have carried out several experiments in integration for multi-scale document retrieval for an English-Chinese cross-lingual spoken document retrieval task. Our experiments show that loose coupling provides reasonably consistent improvements over single-representation modes, whereas unweighted tight coupling shows severe degradation. Additional experiments with tight coupling may illustrate that it, too, can be made superior to single-representation modes. Nevertheless, the loosely coupled version is an excellent multiscale approach whose performance on cross-lingual retrieval begins to mirror that of monolingual results.

Acknowledgments

We wish to acknowledge our fellow MEI team members for their contributions: Berlin Chen, Erika Grams, Sanjeev Khudanpur, Gina Levow, Douglas Oard, Karen Tang, Hsin-min Wang and Jianqiang Wang. The project is conducted during the Johns Hopkins University Summer Workshop 2000 (an NSF Workshop). We thank the LDC for providing the TDT collections. We thank Charles Wayne, George Doddington, James Allan, John Garafolo, Hsin-Hsi Chen and Richard Schwartz for their help. We are grateful to Fred Jelinek and his staff at CLSP for organizing the Workshop.

References

1. H. Meng, B. Chen, E. Grams, S. Khudanpur, G. A. Levow, W. K. LO, D. W. Oard, P. Schone, K. Tang, H. M. Wang, and J. Q. Wang, "Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval," *Technical Report for the NSF summer workshop 2000, Johns Hopkins University*, 2000.
2. H. M. Wang, H. Meng, B. Chen, and W. K. Lo, "Multi-scale audio indexing for Translingual Spoken Document Retrieval," *submitted to ICASSP2001*, 2001.
3. P. Sheridan, M. Wechsler, and P. Schauble, "Cross-Language Speech Retrieval: Establishing a Baseline Performance," *Proceedings of ACM SIGIR97*, pp. 99-108, 1997.
4. H. Meng, W. K. Lo, Y. C. Li, and P. C. Ching, "Multi-scale audio indexing for Chinese spoken document retrieval," *Proceedings of ICSLP2000, vol-IV, pp. 101-4*, 2000.
5. H. M. Wang, "Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese," *Speech Communications*, vol. 32, pp. 49-60, 2000.
6. K. Ng, "Information Fusion for Spoken Document Retrieval," *Proceedings of ICASSP2000*, pp2405-8, 2000.
7. K. Ng, "Towards an integrated approach for spoken document retrieval," *Proceedings of ICSLP 2000*, 2000.
8. W. B. Croft, L. A. Smith, and H. R. Turtle "A Loosely-Coupled Integration of a Text Retrieval System and an Object-Oriented Database System," *Proceedings of ACM SIGIR92*, pp. 223-32, 1992.
9. G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Retrieving Spoken Documents by Combining Multiple Index Sources," *Proceedings of ACM SIGIR 96*, pp. 30-8, 1996.
10. P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 broadcast news transcription system for Mandarin," *Proceedings of the DARPA Broadcast News Workshop,* 1999.
11. J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY Retrieval System," Proceedings of the 3rd International Conference on Database and Expert Systems Applications, pp. 78-83.