Cross-Language Spoken Document Retrieval Using HMM-Based Retrieval Model with Multi-Scale Fusion

WAI-KIT LO, HELEN MENG, and P. C. CHING
The Chinese University of Hong Kong

Cross-language spoken document retrieval (CL-SDR) is the technology that facilitates automatic retrieval of relevant information from a collection of spoken documents in a language that is different from that used in the queries. Information sources that are in different languages can then be retrieved automatically with CL-SDR, and the number of searchable information sources will increase significantly. The HMM-based retrieval model is a probabilistic formulation for the retrieval problem. Extensions to this retrieval model can be made by taking advantage of its probabilistic nature. Specifically, we have incorporated the translation component to make it possible to perform cross-language information retrieval (CLIR). In addition, this HMM-based CLIR retrieval model is also extended for retrieval at subword scales.

In this work the extended HMM-based retrieval model has been applied to an English-Mandarin CL-SDR task, which is to search the Mandarin spoken document collection with English queries at word and subword scales. Retrieval results obtained from these indexing scales are then fused for multi-scale CL-SDR. Experimental results demonstrate that improvement in CL-SDR retrieval performance can be achieved by fusion of word and subword scales.

Categories and Subject Descriptors: H.4.0 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval model*

General Terms: Experimentation, Performance, Theory

Additional Key Words and Phrases: Cross-language information retrieval, Multi-scale data fusion, Spoken document retrieval

1. INTRODUCTION

Cross-language information retrieval (CLIR) is the technology that aims at retrieving relevant information from a document collection in one language using queries in another language. CLIR technology can increase the amount of useful information by enabling access to information sources in other languages.

Authors' addresses: Wai-Kit Lo, Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. He is now with the Spoken Language Translation Research Laboratories, Advanced Telecommunications Research Institute International, 2-2-2 Keihanna Science City, Kyoto 619-0288, Japan; Helen Meng, Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; P. C. Ching, Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

Potential applications of CLIR have drawn much attention to this problem, and several languages have been investigated for text-based CLIR. These include the early work in European languages (such as English, German, Spanish etc.) [Schauble and Sheridan 1997; Braschler et al. 1998], and recently searching Chinese documents using English queries [Kwok 1999; Gao et al. 2000; Chen 2001].

In CLIR, translation is an important component, and many different approaches are applicable. The translation process can be carried out on the query and/or the document side. Since there are usually many documents, adopting the document translation approach will induce significant overhead when there is any change in the translation process. So query translation is adopted because it is simple [Kwok 1999; Chen 2001; Gao et al. 2001]. However, if the translation process seldom changes and the document collection is fixed, the document translation approach has its merit in that the translation is performed only once. Therefore, the choice of translation approach should be made according to the requirements of a particular system. Translation for CLIR can be performed using machine translation systems, parallel corpora, or bilingual dictionaries [Sheridan et al. 1997; Grefenstette 1998; Nie et al. 1999]. In particular, dictionary-based translation [Hull and Grefenstette 1996; Chen et al. 2000; Pirkola et al. 2001] is usually adopted due to the limited availability of translation resources and the simplicity of this approach. In addition, phrasal translation can also be employed to enhance translation accuracy by reducing translation ambiguity [Ballesteros and Croft 1997; Chen 2000; Levow and Oard 2000; Resnik et al. 2001].

With the advent of multimedia technology, much information is archived in the form of multimedia data. Specifically, a large amount of information is archived in spoken form, e.g., recordings of broadcast news, conferences, presentations, meetings etc. These spoken document archives can also be spoken in different languages. In order to effectively retrieve information from these sources, cross-language spoken document retrieval (CL-SDR) is needed. Experience from cross-language text retrieval can be leveraged for achieving CL-SDR, including work that searches among English and German [Sheridan and Ballerini 1996; Sheridan et al. 1997a, 1997b], Serbian and Croatian [Hauptmann et al. 1998] and English-Chinese [Meng et al. 2000, 2001, Lo et al. 2001; Wang et al. 2001]. There are also large-scale investigations initiated by the TREC[1] and TDT[2] workshops.

In spoken document retrieval, indexing can be performed at word and subword scales. Words are lexically-based and subwords are building components for words. Since indexing units at the word scale are lexically oriented, the lexical information can improve specificity of indexing units for retrieval. However, indexing at the word scale is susceptible to the out-of-vocabulary (OOV) problem, since new words are introduced continuously. The relevance feedback technique has been applied to English SDR to handle the OOV problem [Woodland et al. 2000]. It is believed that missing terms due to OOV may be

[1]Text and REtrieval Conferences. http://trec.nist.gov.
[2]Topics Detection and Tracking Workshops. http://www.itl.nist.gov/iad/894.01/tests/tdt/index.htm.

Fig. 1. This example shows the ambiguity in Chinese word segmentation. The character sequence can be segmented into different sequences of words that are syntactically valid and semantically meaningful.

recaptured from relevance feedback. Furthermore, application of relevance feedback to queries and documents using parallel corpus (query expansion and document expansion, respectively) also improve retrieval performance by possible reintroduction of OOV terms. Investigations from the TREC workshops have also shown that as the recognition accuracy of the speech recognition process continues to improve, SDR performance also improves [Garofolo et al. 1997, 1998, 1999]. Experimental results also show that a small improvement in retrieval performance is obtainable when the OOV rate is enhanced. On the other hand, since all words are made up of a finite amount of subword units, indexing at subword scales helps to solve the OOV problem. For example, phoneme n-grams are used as subword indexing units [Ng 2000] to achieve full phonological coverage for spoken documents in English. For Chinese SDR, character and syllable n-grams are used [Bai et al. 2000; Meng et al. 2000; Wang and Chen 2001]. In contrast to the degradation in performance when applying subword retrieval to English SDR (accuracy in phone recognition is generally lower than large vocabulary word recognition) [Mateev et al. 1997; Garofolo et al. 1997], previous experience demonstrated that the use of subword n-grams in Chinese textual information retrieval [Kwok 1997; Nie and Ren 1999] and SDR [Bai et al. 2000; Meng et al. 2000; Wang and Chen 2001] can achieve comparable performance to retrieval based on word units.

Out-of-vocabulary is a common problem in processing Chinese, due to the fact that the definition of “word” in Chinese is ambiguous. In Chinese textual material, there is no explicit word delimiter (such as space in English). Therefore, processing Chinese documents at the word scale requires an explicit word segmentation process. Word segmentation ambiguity in Chinese is shown in Figure 1, where the same character sequence can be segmented into three different word sequences that are syntactically valid and semantically meaningful. Furthermore, the word segmentation process is also based on a given vocabulary. Together with the vocabulary used in the automatic speech recognition process for transcription of spoken documents, the OOV problem has a significant effect on Chinese SDR. The solution to word-scale Chinese SDR does not only require improvement in speech recognition technology to give accurate word outputs, but also improvement in word segmentation technology to segment character sequences into word units that match the vocabulary in the

transcriptions. As a result, subword-based indexing units are commonly used in Chinese SDR [Bai et al. 2000; Meng et al. 2000; Wang and Chen 2001]. Since there is a finite number of characters and syllables in Chinese, the use of character and syllable n-grams as indexing units can always achieve full textual and phonological coverage. In order to take advantage of the robustness of subword indexing units as well as the word-scale lexical information, multi-scale fusion is also adopted for CL-SDR [Meng et al. 2000; Lo et al. 2001; Meng et al. 2001]. In this paper, we try to enhance the HMM-based retrieval model for CL-SDR at both word and subword scales. We also apply the multi-scale fusion technique in Meng et al. [2000] to further improve retrieval performance.

In this paper, we first give a brief introduction to the HMM-based retrieval model in Section 2. This model is then extended for performing cross-language information retrieval (CLIR) in Section 3. The extended model is further enhanced to facilitate CLIR at subword scale. The multi-scale fusion approach used in this work is then introduced in Section 5, and the extended retrieval model is applied to a CL-SDR task. In Section 6, details of the CL-SDR task are given, including task formulation, the experimental corpus, as well as the experimental setup. The CL-SDR results are then presented together with some analysis. Finally, the paper concludes in Section 8.

2. HMM-BASED RETRIEVAL MODEL

The HMM-based retrieval model is a formulation of the information retrieval problem using the hidden Markov model (HMM) [Berger and Lafferty 1999b; Song and Croft 1999a; Miller et al. 1998; Makhoul et al. 2000]. Information retrieval using the HMM-based retrieval model is achieved by finding the most relevant document for a given query in a probabilistic manner. By virtue of several assumptions, this probabilistic formulation for finding a matching document can be reformulated as a query generation process. The probabilities of generating a given query by documents in the collection are used as the retrieval scores. Details of the formulations for HMM-based retrieval models are given in the following sections.

2.1 Basic Formulation

In the probabilistic framework of the HMM-based retrieval model, the objective is to find the document that is likely to be the most relevant to a given query Q as shown in Eq. (1).

$$\arg\max_{D_i} p(D_i|Q) \tag{1}$$

Given a collection of documents, the retrieval process is to calculate the probabilities $p(D_i|Q)$, for every document D_i in the collection that is relevant to the given query Q. By applying Bayes rule, Eq. (1) can be expressed as

$$\begin{aligned} \arg\max_{D_i} p(D_i|Q) &= \arg\max_{D_i} \frac{p(Q \cdot D_i)}{p(Q)} \\ &= \arg\max_{D_i} \frac{p(Q|D_i)p(D_i)}{p(Q)} \end{aligned} \tag{2}$$

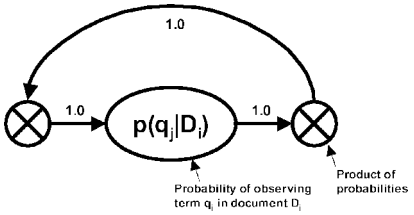

Fig. 2. The basic HMM-based retrieval model for generating query Q given document D_i, for all $q_j \in Q$. The generation probability of query Q by document D_i is determined by the product of all occurring probabilities of the terms $q_j \in Q$ in document D_i.

Since the optimization is independent of the denominator $p(Q)$, Eq. (2) becomes

$$\arg\max_{D_i} p(D_i|Q) = \arg\max_{D_i} p(Q|D_i)p(D_i) \tag{3}$$

If we assume that the *a priori* probability $p(D_i)$ is equal for every document, it can be eliminated, and Eq. (3) becomes

$$\arg\max_{D_i} p(D_i|Q) = \arg\max_{D_i} p(Q|D_i) \tag{4}$$

Equation (4) is essentially a generation model for query Q, given the specific document D_i. Relevant documents are supposed to give higher generation probabilities than irrelevant ones. In practical implementation, the *a priori* probability $p(D_i)$ in Eq. (3) may also be used for different purposes. For example, it can be applied to down-weight documents from a particular subset of documents (e.g., out-dated documents). In this work we assume that all documents are equally probable.

In order to find the probability $p(Q|D_i)$ of generating a query Q from document D_i, it is also assumed that every query term q_j is independent of other terms. The probability $p(Q|D_i)$ is then determined by the joint probabilities of observing all of the query terms q_j in document D_i. As a result, Eq. (4) can be written as

$$\arg\max_{D_i} p(D_i|Q) = \arg\max_{D_i} \prod_{q_j \in Q} p(q_j|D_i). \tag{5}$$

The probabilities $p(q_j|D_i)$ are document-dependent, and are known as the document model probabilities. Figure 2 shows the schematic diagram for this basic HMM-based model for information retrieval.

2.2 Smoothing with the General Language Model

The basic HMM-based retrieval model described by Eq. (5) has a major drawback—when there is any mismatch between the set of terms in the query and those in the document, the probabilities for the unseen terms $p(q_{unseen}|D_i)$ will be undefined. Since the retrieval score is the product of the document probabilities for the terms, any undefined document model probability will make the resulting retrieval score zero. In order to overcome this problem, a general language model is introduced to augment the document model [Makhoul et al. 2000; Chen et al. 2001; Berger and Lafferty 1999a; Song and Croft 1999b,

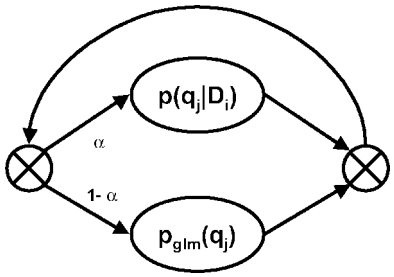

Fig. 3. The HMM-based retrieval model interpolated with the general language model. $p(q_j|D_i)$ represents the document model and $p_{glm}(q_j)$ is the general language model. The query-generation probability by D_i is obtained by taking the product of the weighted sum of document model and general language model for all terms in the query.

1999a; Zhai and Lafferty 2001]. The idea is to include an alternative path to the document models and provide probability estimates based on the specific language for the unseen terms.

Ideally, a probability given by the general language model represents the probability of occurrence for a term in the language. There should be probability estimates for all potential terms. However, exact estimation of these probabilities is impossible, and approximations are always used. One approach is to obtain maximum likelihood estimates from large corpora. For example, the collection of documents used for retrieval can be used to estimate term probabilities in the general language model.

After introducing the general language model to Eq. (5), the HMM-based retrieval model becomes

$$p(Q|D_i) = \prod_{q_j \in Q} (w_{doc} \cdot p(q_j|D_i) + w_{glm} \cdot p_{glm}(q_j)) \tag{6}$$

where w_{doc} and w_{glm} are the weights for the document model and general language model, respectively. The sum of w_{doc} and w_{glm} is usually constrained to 1 and w_{doc} is represented as α and w_{glm} as $(1-\alpha)$.

The linear combination of the document model and the general language model effectively smooths the document models by interpolation. When a query term is not contained in a document, the nonzero general language model will be interpolated with the zero document model probability of the missing term. A nonzero retrieval score can then be obtained. Figure 3 is a schematic diagram for the smoothed HMM-based retrieval model. This smoothing method is known as the Jelinek and Mercer (or interpolation-based) smoothing method in Zhai and Lafferty [2001]. It was shown that this smoothing method give better performance than other smoothing methods for the retrieval task using long queries. Therefore, this interpolation-based smoothing method is used in our experiments and $p(Q|D_i)$ from Eq. (6) are used as the retrieval score.

3. FORMULATION OF THE HMM-BASED RETRIEVAL MODEL FOR CLIR

In this section the HMM-based retrieval model is extended for the CLIR task. As mentioned in the previous section, the HMM-based retrieval model formulated

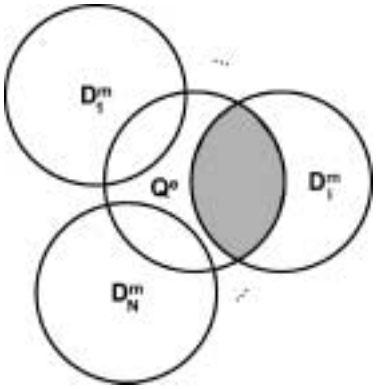

Fig. 4. The cross-language retrieval problem using HMM. The probability space for the queries Q^e and the documents D_i^m are shown. If the probabilities of all documents are equal, the problem is to find the one with the largest overlap with the query.

for monolingual retrieval has been investigated extensively. However, the original formulation of the HMM-based retrieval model is not capable of performing cross-language retrieval. As mentioned in other work [Hiemstra 2000; Gao et al. 2000; Xu and Weischedel 2000], the probabilistic formulation of the HMM-based retrieval model makes it possible to extend it for the use of CLIR. By making several assumptions, we incorporate the translation component into the HMM-based retrieval model to facilitate CLIR.

3.1 Derivation for HMM CLIR from the Basic Formula

Suppose that there is a collection of documents in language m. If it is required to retrieve relevant documents from the collection using queries given in another language e, one solution is to translate the queries into the language of the documents using an available machine translation system and then perform monolingual document retrieval using the translated queries. In this section we are going to extend the HMM-based retrieval model with the cross-language translation component incorporated to perform CLIR.

By substituting the query Q^e in language e and document D_i^m in language m into the original HMM formulation, the basic formula for cross-language retrieval is

$$\arg\max_{D_i^m} p\left(D_i^m | Q^e\right) \tag{7}$$

The goal is to find the most probable document D_i^m in language m from the document collection for the given query Q^e in language e.

Similar to the monolingual formulation, Eq. (7) can be re-established as finding the most probable document D_i^m that generates the query Q^e in language e. Equation (8) shows the resulting formulation, and Figure 4 is an illustration of the probabilistic cross-language retrieval problem where the document D_i^m with the maximum intersecting area with the query Q^e is the required solution.

$$\arg\max_{D_i^m} p\left(Q^e | D_i^m\right) \tag{8}$$

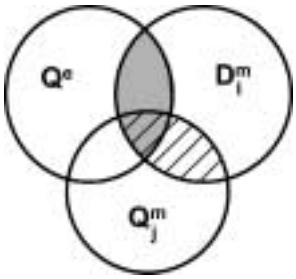

Fig. 5. The probability space for the queries Q^e and document D_i^m together with the immediate query Q_j^m. The patched area is $p(Q_j^m D_i^m)$.

In order to connect the two languages, intermediate query Q^m in the language of the document is introduced. An illustration of the CLIR problem with the intermediate query is shown in Figure 5. It is assumed that every query Q^e in language e can be translated into language m to give the query Q^m.[3]

In addition, we also assume that

$$p\left(Q^e|D_i^m\right) \approx \sum_{\forall Q^m} p\left(Q^e Q^m|D_i^m\right) \tag{9}$$

Substituting Eq. (9) for Eq. (8), we have

$$\begin{aligned}
\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) &\approx \arg\max_{D_i^m} \sum_{\forall Q^m} p\left(Q^e Q^m|D_i^m\right) \\
&= \arg\max_{D_i^m} \sum_{\forall Q^m} \frac{p\left(Q^e Q^m D_i^m\right) \cdot p\left(Q^m D_i^m\right)}{p\left(Q^m D_i^m\right) \cdot p\left(D_i^m\right)} \\
&= \arg\max_{D_i^m} \sum_{\forall Q^m} \left[p\left(Q^e|Q^m D_i^m\right) \cdot p\left(Q^m|D_i^m\right)\right] \qquad (10)
\end{aligned}$$

In Eq. (10), $p(Q^m|D_i^m)$ represents the probability of the query Q^m being generated by document D_i^m; $p(Q^e|Q^m D_i^m)$ is the probability of generating query Q^e in language e given that the document is D_i^m and the query in language m is Q^m.

3.2 Cross-Language Translation Using Translated Queries

To retrieve using machine-translated queries, Eq. (10) can be applied with some modifications. For a given set of top-N hypotheses from machine translation systems or translations from N different systems, these translations can be used as queries in the space of intermediate queries, Q^m. Assuming that the translation is independent of D_i^m, Eq. (10) can be modified to

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \sum_{j=1}^{N} \left[p\left(Q^e|Q_j^m\right) \cdot p\left(Q_j^m|D_i^m\right)\right].$$

[3]However, there could be untranslatable queries in practice. In such cases, the translation and thus the retrieval will fail.

It is also assumed that

$$p\left(Q_j^m|D_i^m\right) = \prod_{\forall q^m \in Q_j^m} p\left(q^m|D_i^m\right).$$

Therefore, we have

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \sum_{j=1}^{N} \left[p\left(Q^e|Q_j^m\right) \cdot \prod_{\forall q^m \in Q_j^m} p\left(q^m|D_i^m\right) \right]. \quad (11)$$

Since $p(Q^e|Q_j^m) = \frac{p(Q_j^m|Q^e)\cdot p(Q^e)}{p(Q_j^m)}$, $p(Q^e|Q_j^m)$ in Eq. (11) can be replaced by $p(Q_j^m|Q^e)$ by assuming that $p(Q^e)$ and $p(Q_j^m)$ are equally probable. As a result, the CLIR formulation becomes

$$\arg\max_{D_i^m} p(Q^e|D_i^m) \approx \arg\max_{D_i^m} \sum_{j=1}^{N} \left[p\left(Q_j^m|Q^e\right) \cdot \prod_{\forall q^m \in Q_j^m} p\left(q^m|D_i^m\right) \right] \quad (12)$$

where $p(Q_j^m|Q^e)$ is the confidence of the hypothesis given by the machine translation system and $p(q^m|D_i^m)$ is the document model as in the monolingual formulation.

When there is only *one* translation Q_1^m for the query Q^e, the summation in Eq. (12) will be made over a single element set, and hence the formulation is degenerated to a monolingual retrieval task using the translated query.

3.3 Cross-Language Translation by Incorporating Translation Components

CLIR formulation using HMM can be derived by generalization of the monolingual formulation. Additional probabilistic translation terms are introduced to connect terms in the query and those in the document. These probabilities are then summed over all possible indexing terms.

The CLIR formulation described by Eq. (10) can be modified to incorporate the appropriate translation component to make use of the translation resources such as parallel corpora and translation dictionary. It is assumed that query terms q^e in Q^e are independent of each other. Therefore,

$$p\left(Q^e \cdot D_i^m\right) \approx \prod_{q^e \in Q^e} p\left(q^e \cdot D_i^m\right)$$

and

$$\arg\max_{D_i^m} p\left(D_i^m|Q^e\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall Q^m} \left[p\left(q^e|Q^m D_i^m\right) \cdot p\left(Q^m|D_i^m\right) \right] \quad (13)$$

Since it is impossible to iterate over all intermediate queries Q^m, it is simpler to iterate over all possible indexing terms q^m. As a result, the CLIR formulation becomes

$$\arg\max_{D_i^m} p\left(D_i^m|Q^e\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q^m} \left[p\left(q^e|q^m D_i^m\right) \cdot p\left(q^m|D_i^m\right) \right] \quad (14)$$

In Eq. (14), the term $p(q^m|D_i^m)$ can be obtained using maximum likelihood estimation over each of the documents D_i^m. For the cross-language translation term $p(q^e|q^m D_i^m)$, there are several approaches for estimation that are discussed in the following sections.

3.3.1 *Simplification for Context-Dependent Translation.* In practice, reliable estimation for the cross-language translation probability, $p(q^e|q^m D_i^m)$, is difficult. One solution to this problem is to make use of the contextual information for context-dependent term-based translation. The translation component can be simplified to

$$p\left(q^e|q^m D_i^m\right) \approx p(q^e|q^m context(q^m)).$$

Terms in its proximity can be used as the context of q^m, such as

$$p\left(q^e|q^m D_i^m\right) \approx p\left(q^e|q_{-n}^m \cdots q_{-1}^m q^m q_1^m \cdots q_n^m\right).$$

where n is the desired contextual length and q_i^m is the term with an offset of i term from the term q^m.

In this framework, the translation probability can be obtained from a context-dependent translation dictionary or derived from aligned parallel corpora. Thus, the CLIR formulation can be rewritten as

$$\begin{aligned}&\arg\max_{D_i^m} p\left(Q^e|D_i^m\right)\\ &\approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q^m \in D_i^m} \left[p\left(q^e|q_{-n}^m \cdots q_{-1}^m q^m q_1^m \cdots q_n^m\right) \cdot p\left(q^m|D_i^m\right)\right].\end{aligned}$$

3.3.2 *Simplification for Domain-Specific Translation.* The translation probability can also be simplified to make use of the domain-specific translation resources (e.g., domain-dependent translation dictionary) by assuming that

$$p\left(q^e|q^m D_i^m\right) \approx p\left(q^e|q^m domain\left(D_i^m\right)\right).$$

This is the probability of observing query term q^e, given query term q^m and the domain of the document is $domain(D_i^m)$. Domain-specific translation dictionaries can be used to obtain the translation probability whenever the domain is known. As a result, the CLIR formulation becomes

$$\begin{aligned}&\arg\max_{D_i^m} p\left(Q^e|D_i^m\right)\\ &\approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q^m \in D_i^m} \left[p\left(q^e|q^m domain\left(D_i^m\right)\right) \cdot p\left(q^m|D_i^m\right)\right].\end{aligned}$$

3.3.3 *Simplification for Context-Free Translation.* In case the domain is unknown or only domain-independent translation resources are available, the cross-language translation probability can also be simplified to a context-free domain-independent translation probability

$$p\left(q^e|q^m D_i^m\right) \approx p(q^e|q^m).$$

In this simplification, it is assumed that q^e is solely determined by q^m and is independent of D_i^m. To estimate this simplified translation probability, we can make use of the general domain bilingual dictionaries. This approach is adopted here, and the probability is calculated based on a translation dictionary in the form of a bilingual term list (details are given in Section 7). The resulting HMM CLIR formulation is given by

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q^m} \left[p(q^e|q^m) \cdot p\left(q^m|D_i^m\right)\right]. \quad (15)$$

If we let q^e and q^m span the same space and assume that there is no inter-dependency among these terms (i.e., $p(q_i^e|q_j^m) = 1$ for $q_i^e = q_j^m$, otherwise 0), the CLIR formulation in Eq. (15) can be degenerated to a monolingual IR formulation.

4. EXTENSION OF THE HMM-BASED MODEL FOR SUBWORD SCALE CLIR

In this section, the CLIR formulation for the HMM-based retrieval model is extended to facilitate retrieval at subword scales. For monolingual retrieval tasks using the HMM-based retrieval model, retrieval at subword indexing scales can be performed by preprocessing queries and documents into desired scales before retrieval. Word and subword scale indexing units can then be processed in the same way. However, for the derived cross-language information retrieval formulation, the translation component has to be modified to facilitate retrieval at subword scales.

For subword scale CLIR, word scale terms q^m in Eq. (15) are replaced with terms in subword scale q_s^m. Therefore, the translation probabilities and the document model probabilities are rewritten in subword scale.

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q_s^m} \left[p\left(q^e|q_s^m\right) \cdot p\left(q_s^m|D_i^m\right)\right], \quad (16)$$

where $p(q_s^m|D_i^m)$ represents the document model probability for subword units and $p(q^e|q_s^m)$ is the subword scale cross-language translation probability.

For the subword scale translation probability, it can be approximated by

$$p\left(q^e|q_s^m\right) \approx \sum_{\forall q^m} p(q^e|q^m) \cdot p\left(q^m|q_s^m\right). \quad (17)$$

The subword scale translation probabilities is obtained by summing up the product of word translation probability $p(q^e|q^m)$ and subword-word probability $p(q^m|q_s^m)$ over all words. In Eq. (17) the subword-word probability represents the probability of observing the word scale term q^m given the subword scale unit q_s^m. By substituting Eq. (17) for Eq. (16), subword scale CLIR using the

HMM-based retrieval model is obtained as shown in Eq. (18).

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q_s^m} \left[p\left(q^e|q_s^m\right) \cdot p\left(q_s^m|D_i^m\right)\right]$$

$$= \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q_s^m} \left[\left[\sum_{\forall q^m} p(q^e|q^m) \cdot p\left(q^m|q_s^m\right)\right] \cdot p\left(q_s^m|D_i^m\right)\right] \quad (18)$$

5. MULTI-SCALE RETRIEVAL

It is known that retrieval at each indexing scale presents a set of unique advantages. At the word scale, indexing units are lexically meaningful entities. The lexical information is useful for retrieval by discriminating among confusing units. However, it has also been shown that retrieval at word scale is susceptible to out-of-vocabulary (OOV) problems (both from word segmentation of text and also transcription of spoken documents using automatic speech recognizers) [Ng 2000; Meng et al. 2000]. The use of indexing units at subword scales can help circumvent these OOV problems. Although subword indexing units (such as phoneme bigrams, or character bigrams) may introduce some ambiguity in the indexing units, the robustness of subword indexing has been demonstrated to improve retrieval performance for both textual [Kwok 1997; Nie and Ren 1999] and spoken document retrieval [Chen et al. 2000; Wang 2000; Meng et al. 2000; Chen et al. 2001] in Chinese.

In order to take advantage of word and subword scale indexing units, multi-scale retrieval can be applied [Meng et al. 2000; Lo et al. 2001]. Data fusion of retrieval results for performance improvement has been extensively investigated for textual information retrieval [Fox and Shaw 1993; Bartell et al. 1994; Belkin et al. 1995]. Ranked retrieval lists from different retrieval systems and/or query formulations are merged to improve retrieval performance. It is also stated in Vogt and Cottrell [1999] that fusion of retrieval ranked lists can only achieve improvement in certain cases. One possible way is to merge retrieval results that are correct and complementary. Improvement is obtained by boosting the rankings of relevant documents and suppressing those of irrelevant ones after combining several ranked retrieval lists. Specifically, multi-scale retrieval can be applied by data fusion to retrieval results obtained from different indexing scales. Performance improvement is believed to be achievable by fusing word and subword scales. In this work, we examine the CL-SDR retrieval performance at the word, character bigram, and syllable bigram indexing scales as well as the retrieval performance achieved by multi-scale fusion among these scales.

The idea of multi-scale retrieval by data fusion of retrieval results is illustrated in Figure 6. Retrieval results in the form of ranked retrieval lists are merged by a fusion function. The fusion function returns a new retrieval score for every pair of queries and documents in the collection of documents. In this work, multi-scale fusion of retrieval results is accomplished by a linear combination of the retrieval scores obtained from Eq. (18). The general form of the

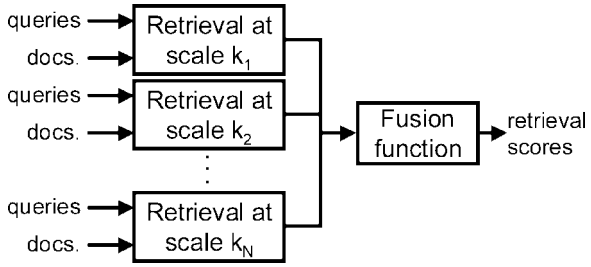

Fig. 6. Multi-scale retrieval achieved by fusion of retrieval results from composite scales. A fusion function is defined to derive new retrieval scores from the composite retrieval results.

fusion function is given as follows,

$$Score_{multi\text{-}scale}(q_i, d_j) = \sum_{k \in K} w_k \cdot Score_k(q_i, d_j) \tag{19}$$

where $Score_{multi\text{-}scale}(q_i, d_j)$ is the multi-scale retrieval score for retrieving document j with query i. $Score_k$ is the retrieval score obtained from retrieval at indexing scale k, K is the set of indexing scales to be fused, and w_k is the weight for scale k in the fusion. After multi-scale fusion of retrieval results, documents are then reranked according to the fused retrieval scores and re-evaluated.

6. RETRIEVAL TASK AND EXPERIMENTAL CORPUS

In our CL-SDR experiments, textual queries in English are used to search for relevant spoken documents in Mandarin. This is an English-Mandarin cross-language spoken document retrieval task. We have adopted a query-by-exemplar task formulation. News articles are used as queries to search for relevant documents in the collection.

6.1 TDT-2 Corpus

The Topic Detection and Tracking phase 2 corpus (TDT-2) [LDC 2000] is used in our CL-SDR experiments. In the TDT-2 corpus, there are news articles and broadcast news recordings collected from different sources. Specifically for our tests, 2265 recordings from the Mandarin radio news broadcast of Voice of America are used as the spoken document collection and 171 articles from the New York Times and Associated Press are used as the English queries.

6.1.1 *Topic Relevance Information.* In the TDT-2 corpus, all news articles and audio recordings are annotated with a topic together with a description on the level of relevance. The topics are chosen out of a pool of 100 predefined topics and the level of relevance is either “brief,” “yes,” or irrelevant. “Yes” means that the article is well matched with the topic and “brief” means that the topic is only loosely related.

For evaluation of retrieval performance, a relevance list for the queries and documents used in the experiments is needed. The list of relevant queries and documents is derived from these topic annotations. News articles and audio recordings in the corpus that are annotated with the same topic and assigned

Table I

Topic Assignment in the TDT-2 Corpus

Query/document ID	Topics	Level of relevance
VOA19980224.2100.3306	13	yes
VOA19980304.2300.2809	13	brief
NYT19980321.0123	13	brief
NYT19980404.0131	13	yes
	⋮	

⇓

Query-Document Relevance Assignment

Query/document ID	VOA19980224.2100.3306	VOA19980304.2300.2809
NYT19980321.0123	irrelevant	irrelevant
NYT19980404.0131	relevant	irrelevant

Illustration of the derivation process for query-document relevance of the data in the TDT-2 corpus. Articles used as queries or documents that are assigned with the same topic and have relevance levels "yes" are treated as relevant. All other articles are considered irrelevant.

with levels of relevance "yes" are considered relevant. All other articles and audio recordings are treated as irrelevant. To illustrate the derivation process for the query-document relevance list, an example is given in Table I.

6.1.2 *Evaluation Measure.* For performance evaluation, a modified non-interpolated mean average precision (mAP) is defined as

$$mAP_{noninterpolated} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_k(i,\, j)} \tag{20}$$

where L is the total number of topics, M_i is the number of query exemplars for each topic i, N_i is the total number of relevant documents for topic i, and $rank_k(i, j)$ is the rank of the kth relevant document in the ranked list for query j on topic i.

6.1.3 *Mandarin Spoken Documents.* The Mandarin spoken documents in this study consist of the daily evening news (19:00–20:00) collected from the Voice of America radio broadcast between February to June 1998. The broadcast is made by satellite link and the audio data encoded with MPEG compression can be collected digitally. Collected data is then converted to a sampling rate of 16 kHz with 16-bit resolution. Further details about data quality and collection process can be found in the TDT-2 document [LDC 2000].

In the TDT-2 corpus, the automatic speech recognition outputs from *a streamlined version of the research grade Dragon Mandarin ASR* [Zhan 1999; LDC 2000] are provided with the audio recordings in the form of Chinese word sequences. These Chinese word transcriptions are used as the word scale document representations. From these sequences of words, document representations at the character bigram scale are also derived. In addition, syllables for the words are obtained by looking up the CALLHOME [Huang et al. 1997] pronunciation lexicon. Overlapping syllable bigrams are then derived similarly from these syllable sequences for the document representations at the syllable bigram scale. Figure 7 is an example showing the derivation process

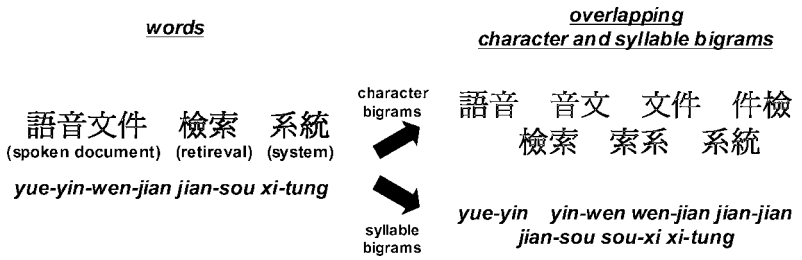

Fig. 7. An example showing the formation of overlapping character and syllable bigrams from a word. This word means "spoken document retrieval system."

for overlapping character bigrams as well as syllable bigrams from a word sequence.

6.1.4 *English Queries.* The English news articles from the TDT-2 corpus are used as the textual queries for our English-Mandarin CL-SDR experiments. We have chosen two sources of news articles: the New York Times and the Associated Press. These articles span the period from January to June 1998. These English articles are also marked for relevance to the 100 predefined topics.

Here, English queries are selected from 15 topics and a maximum of 12 articles are randomly selected from the TDT-2 corpus. If there are fewer than 12 relevant articles, all the relevant ones will be used. There are altogether 171 English news articles covering the 15 topics used as our English queries.

In addition, a phrase extraction process is applied to every selected article for locating boundaries where consecutive sequences of words can form lexical phrases. The English articles are first fed to a BBN Identifinder(TM) [BBN 2000] that marks all the named entities in the articles. The Identifinder outputs are then passed to a phrase extraction process. Phrase extraction is achieved by a left-to-right maximum matching algorithm [Meng et al. 2000] in conjunction with the English word list from the translation dictionary. As a result, the translation process for cross-language retrieval can be carried out on a phrase basis. After extracting the phrases, stop words are also removed from the articles by referring to a stop-word list (from SMART [Salton and McGill 1983]).

On the other hand, Chinese queries are also selected for the 15 topics with 3 articles on every topic to obtain a monolingual Chinese SDR performance reference. The Chinese queries are news articles included in the TDT-2 collection. Figure 8 gives a summary of the Chinese and English articles and the Mandarin spoken documents that are extracted from the TDT-2 corpus.

7. CL-SDR EXPERIMENTS USING THE HMM-BASED RETRIEVAL MODEL

In order to evaluate the extended HMM-based retrieval model for CL-SDR, English-Mandarin CL-SDR experiments are performed using the extended retrieval model at both word and subword scales. For word scale retrieval, Eq. (15) is applied where q^e are identified English phrases in queries and q^m are Chinese words in transcriptions of Mandarin spoken documents. In addition, a general language model is introduced as an alternative path to smooth the document

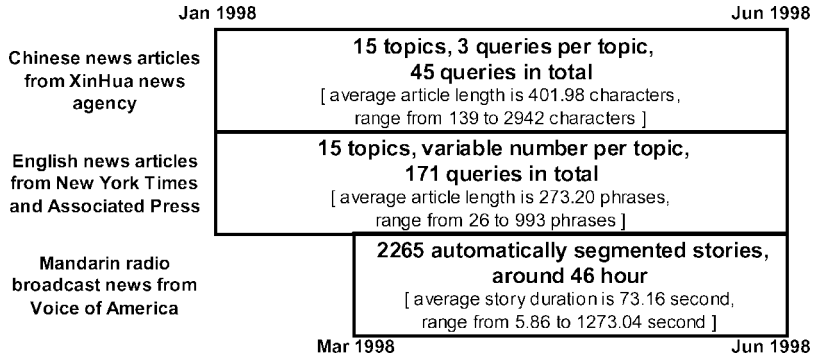

Fig. 8. Summary of the Mandarin spoken documents, Chinese and English queries from the TDT-2 corpus are used in the spoken document retrieval experiments.

models. As a result, Eq. (15) for word scale retrieval is modified to

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q^m} \left[p(q^e|q^m) \cdot \left[\alpha \cdot p\left(q^m|D_i^m\right) + (1-\alpha) \cdot p_{glm}(q^m)\right] \right] \tag{21}$$

where α is the document model weight, q^e are the English phrases in the queries, and q^m are the words in the documents.

For subword scales, retrieval is performed at character bigram and syllable bigram scales. General language models are also applied to Eq. (18) for smoothing. The operation for the subword scale CL-SDR using smoothed HMM becomes

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q_s^m} \left[\left[\sum_{\forall q^m} p(q^e|q^m) \cdot p\left(q^m|q_s^m\right) \right] \cdot \left[\alpha \cdot p\left(q_s^m|D_i^m\right) + (1-\alpha) \cdot p_{glm}\left(q_s^m\right)\right] \right]$$

where α is the document model weight, q^e are the English phrases in the queries, and q_s^m are the subword scale units in the documents.

7.1 Experimental Setup

The document models in the HMM-based CLIR model are obtained from transcriptions of spoken documents at corresponding scales (syllable bigrams, character bigrams, or words) based on maximum likelihood estimation. These probabilities are given by

$$p\left(q^m|D_i^m\right) = \frac{tf\left(q^m \in D_i^m\right)}{\sum_{\forall q_j^m \in D_i^m} tf\left(q_j^m \in D_i^m\right)} \tag{22}$$

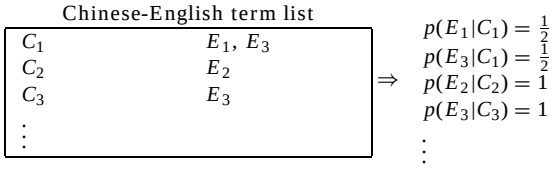

Fig. 9. Calculation of the word translation probability based on the translation dictionary.

where $tf(q^m)$ represents the term frequency of q^m, and q^m is the indexing term in either syllable bigram, character bigram or word scale.

Similarly, general language models are also obtained based on maximum likelihood estimation over the whole collection of spoken documents,

$$p_{glm}(q^m) = \frac{\sum_{\forall D_i^m} tf(q^m \in D_i^m)}{\sum_{\forall D_i^m} \left[\sum_{\forall q_j^m \in D_i^m} tf(q_j^m \in D_i^m) \right]} \tag{23}$$

7.1.1 *Estimation of Word Translation Probability.* The translation probability for an English query term q^e given the Chinese term q^m is obtained with the assumption that the translation of the terms is independent of other terms (see Section 3.3.3). All translation probabilities are calculated based on the number of English translations for the Chinese term in the translation dictionary[4] as shown in Figure 9. The equation is given by

$$p(q^e|q^m) = \frac{1}{\#\ of\ translation\ for\ q^m}.$$

7.1.2 *Estimation of Subword Translation Probability.* In subword scale CLIR experiments, estimation for the translation probability is achieved by decomposing the probability into two elements (see Eq. (17)): word translation probability and *subword-word* probability. Word translation probabilities are obtained as described previously. Subword-word probability is the probability of observing a particular word given a subword. It can be estimated using the list of Chinese words in the translation dictionary, as shown below,

$$p(q^m|q_s^m) = \frac{1}{\#\ of\ words\ containing\ subword\ q_s^m}$$

By estimating the subword-word probability this way, all translatable Chinese words are guaranteed to be covered. It should be noted that this is an approximation based on the assumption that all vocabulary words are equally probable.

Figure 10 illustrates the derivation process for the subword-word probability from a word list. Given a list of words, subword-word probabilities can be derived by rearranging the word-subword pairs to subword-word pairs, as shown.

[4]The translation dictionary used in this work is the *Chinese English Translation Assistance* (CETA).

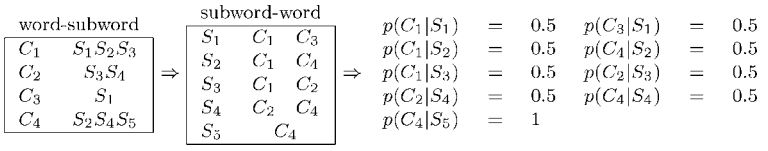

Fig. 10. Calculating the subword-word probability based on a given vocabulary of words.

The list of Chinese words in the translation dictionary (CETA) is adopted as our vocabulary for estimating subword-word probabilities. As long as the translation dictionary is not changed, word translation probabilities and the set of Chinese words are also invariant to queries and documents. Therefore, these probabilities can be precalculated and combined using Eq. (17). With these simplifications, retrieval at subword scales are implemented as

$$\arg\max_{D_i^m} p\left(Q^e|D_i^m\right) \approx \arg\max_{D_i^m} \prod_{q^e \in Q^e} \sum_{\forall q_s^m} \left[p\left(q^e|q_s^m\right) \cdot \left[\alpha \cdot p\left(q_s^m|D_i^m\right) + (1-\alpha) \cdot p_{glm}\left(q_s^m\right)\right]\right],$$

which is practically the same as the word scale CLIR formulation (Eq. (21)).

In practice, every document in the collection has a different distribution of words. This implies that subword-word probabilities are different across documents. The estimation above is based on information available in the vocabulary. Consequently, these estimated subword-word probabilities may introduce some degree of ambiguity in the translation process.

7.1.3 *Multistage Back-Off Phrase Translation.* In the extended HMM-based retrieval model, the translation component described in Meng et al. [2000], Levow and Oard [2000], and Resnik et al. [2001] is modified to cater to any potential mismatches between the marked English phrases and the English vocabulary in the translation dictionary. In case there is no English phrase matching in the dictionary, the phrase-based translation process is backed-off to word-based by breaking down the phrases into word sequences. The stemming and matching processes are repeated for all extracted words. The modified translation process is shown in Figure 11.

7.2 Results and Analysis

Figure 12 shows the retrieval results for the English-Mandarin CL-SDR task using the extended HMM-based retrieval model. From these results, it is found that retrieval at the character bigram scale achieves the best performance (mAP = 0.490) among other indexing scales. In order to obtain a reference performance level, monolingual Chinese SDR on the word scale has also been performed using the HMM-based retrieval model (Eq. (6)). Chinese news articles are used to retrieve Mandarin news recordings on the same topics (see Figure 8). This experiment achieves a mean average precision (mAP) of 0.566. Compared to this performance level, it can be seen that the HMM-based retrieval model

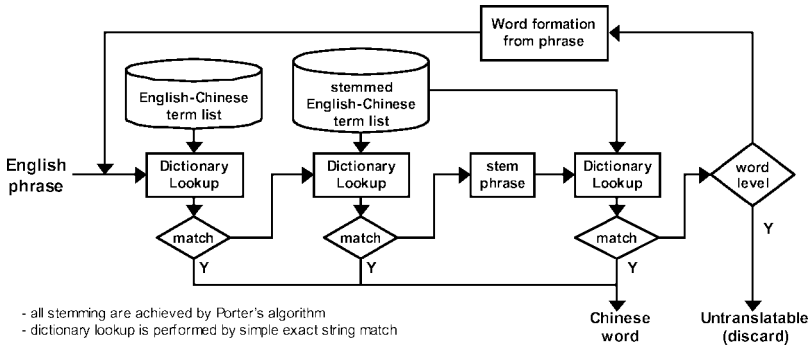

Fig. 11. Multistage back-off translation process employed in the cross-language HMM-based retrieval model. All unmatched phrases are broken down to word sequences. The stemming and matching processes are repeated for every extracted word.

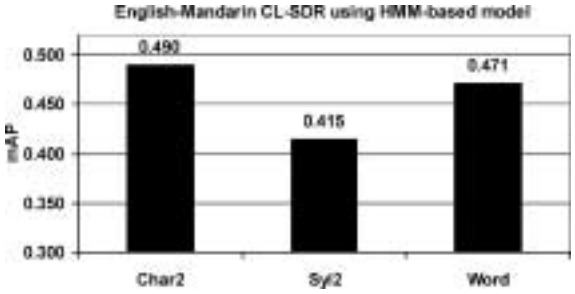

Fig. 12. Cross-language spoken document retrieval results using an HMM-based retrieval model in a query-by-example task that search for Mandarin spoken documents using English textual queries.

incorporated with a translation component achieves a performance level at 83% of that obtained in a monolingual SDR task. This shows that the extended HMM-based retrieval model is applicable to our CL-SDR task and that satisfactory performances are achievable.

7.2.1 *Character Bigrams Recover Loss Due to Vocabulary Mismatches.* Performance improvements in the CL-SDR by using character bigrams are due to the robustness of the subword units to vocabulary mismatches and partial recognition errors. In the extended HMM-based retrieval model for CL-SDR, when there is any mismatch among the following pairs of vocabularies, those words become untranslatable at the word scale. These pairs of vocabularies include

(1) the vocabulary of the English queries and the list of English words/phrases in the translation dictionary;
(2) the list of Chinese words in the translation dictionary and the vocabulary of the speech recognizer.

The first case is obvious, since it is impossible to translate a word when it is not found in the translation dictionary. The second case stems from the fact

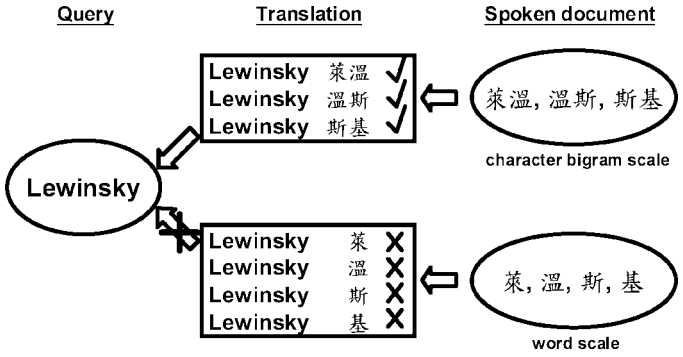

Fig. 13. The word and subword translations for the proper name "Lewinsky" are shown with the number of translations. Word scale CL-SDR fails when there is any character error in the recognized Chinese word or the word is recognized as separate characters due to OOV. At the character bigram scale, either one of the three character bigrams can recover the original English word by subword translation and contribute to the retrieval process.

that any word in the spoken documents not included in the recognizer vocabulary may be returned as separate characters. In a Chinese speech recognition system, the recognizer vocabulary usually contains a list of words together with all Chinese characters. This is done as a backup for OOV words. When there is an OOV word encountered during recognition, it will not be possible to return the OOV word as a word unit in the recognition output. In this case, the best possible alternative output during recognition is to have all of the composite characters in the OOV word be returned as separate characters—a sequence of single-character word units. This is made possible by having all Chinese characters included in the recognizer vocabulary. However, returning a word in separate characters will make translation at the word level fail. If character bigrams are used, these bigrams can then be recovered and they can contribute to the retrieval process by translation at the subword level. These OOV words are common for proper names in spoken documents. An example extracted from our experiment is shown in Figure 13.

From the experiment, it is found that "Lewinsky" is included in the translation dictionary, but the Chinese word is not included in the recognizer vocabulary. As a result, this word is returned as separate characters in the transcription of the spoken document. When a query is given with the word "Lewinsky," even if the document actually contains this word as separate characters, no correct translation can be found at word level. If the retrieval is carried out at the character bigram level, the correct English word translation can be obtained and contribute to the retrieval process. However, at the syllable bigram scale, there are many irrelevant translations included due to the homophone effect. As shown in Table II, even though the translation is recovered, the large number of spurious translations causes much confusion in the retrieval, and therefore the performance is degraded.

7.2.2 *Character Bigrams Recover Minor Recognition Errors.* In practice, transcriptions of spoken documents are obtained using automatic speech

Table II

Chinese	# translations		English examples
	word scale		
萊溫斯基	1		Lewinsky
	character bigram	syllable bigram	
萊溫	2	9	Lewinsky, Malaysian (the language), ...
溫斯	6	18	Lewinsky, Winston, Winslow ...
斯基	28	107	Lewinsky, Eskimo, Tchalkovsky, ...

This table shows the number of translations for the units at word and subword scales. It can be seen that many irrelevant translations are introduced at the syllable bigram scale due to the homophone effect. These ambiguities inevitably lower the retrieval performance.

English word in dictionary	*Chinese word in dictionary*	*Recognition result*	*# matching words*	*Matching subwords*	*# matching subwords*
Lewinsky	萊溫斯基	萊 文 斯 基	0	斯基	1
Nuclear nonproliferation	核不擴散	科 不 擴散	0	不擴 擴散	2

Fig. 14. Examples of transcriptions extracted from the Mandarin spoken document collection, showing that there are recognition errors (underlined) that make translation at word level fail. These errors can be partially recovered by performing retrieval at the subword scale (e.g., character bigram).

recognition. Since there are errors in the speech recognition outputs, there are erroneous words in transcriptions such as substitutions by homophones or partial recognition errors. Partial recognition errors refer to the case when the recognition output of a multicharacter word has some wrong characters. Figure 14 shows extractions of retrieval results in our CL-SDR experiments that illustrate the effect of partial recognition errors. When whole words are not correctly recognized, correct translations cannot be found in the translation dictionary. These single character errors make translation at word level fail totally. By using character bigrams in CL-SDR, those character bigrams formed from the correctly recognized characters can contribute to the translation at subword scale. Therefore, correct translations can be obtained from these subword units for the partially misrecognized words and contribute to the CL-SDR process.

7.2.3 *Multi-Scale Fusion Between Word and Subword.* Figure 15 shows the multi-scale retrieval results from fusion for the word and subword indexing scales. The scale weightings (w_k) in Eq. (19) are empirically tuned using the same collection of documents. While constraining the sum of scale weightings to one ($\sum w_k = 1$), w_k is changed from 0.1 to 1 in steps of 0.1 to obtain the best fusion result. It can be observed that when multi-scale retrieval is applied, the best performance (mAP = 0.510) is obtained from fusing character bigrams with words.

The results also show that fusion of syllable bigrams and words can improve retrieval performance over the composite scales. However, multi-scale fusion between the character bigrams and syllable bigrams does not show any improvement. Table III shows the relative improvements with multi-scale fusion.

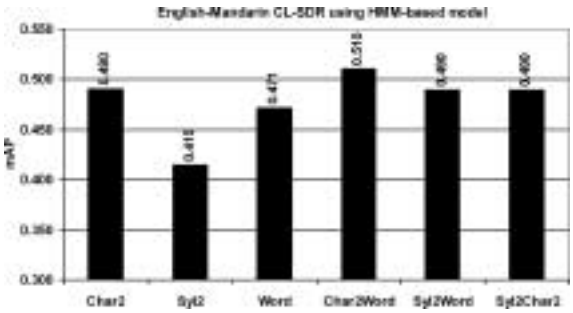

Fig. 15. Multi-scale cross-language spoken document retrieval results using an HMM-based retrieval model in a query-by-example task that searches for Mandarin spoken documents using English textual queries.

Table III

Composite scales	Char2	Word	Syl2	Word	Syl2	Char2
Composite performance	0.490	0.471	0.415	0.471	0.415	0.490
Fused performance	0.510		0.490		0.490	
% gain relative to the better of the composite scales	4.11% (w.r.t. Char2)		3.96% (w.r.t. Word)		−0.12% (w.r.t. Char2)	

The gain from multi-scale fusion (using long queries) in the retrieval performance relative to the better of the composite scales.

Table IV

Composite scales	Char2	Word	Syl2	Word	Syl2	Char2
Composite performance	0.428	0.344	0.317	0.344	0.317	0.428
Fused performance	0.444		0.381		0.424	
% gain relative to the better of the composite scales	3.74% (w.r.t. Char2)		10.8% (w.r.t. Word)		−0.93% (w.r.t. Char2)	

The gain from multi-scale fusion (using short queries) in the retrieval performance relative to the better of the composite scales.

In order to show the effect of multi-scale fusion in the CL-SDR task using short queries, we also performed the CL-SDR experiments using the topic labels in the TDT-2 collection as queries.[5] The retrieval performances are shown in Table IV.

The analyses show that when multi-scale fusion is performed for retrieval results from word and subword scales, there are additional improvements in retrieval performance relative to that of the composite scales (around 4% for long queries and up to 10% for short queries, with respect to the better of the composite scales). This confirms the hypothesis that word and subword indexing scales are complementary. Therefore, it is possible to improve retrieval performance by fusing of the complementary indexing scales.

Furthermore, there is only a marginal improvement in retrieval performance when the three indexing scales are fused (mAP = 0.511 for the long queries and mAP = 0.447 for the short queries). Since multi-scale retrieval using character

[5]The labels range from 3 to 13 words long with an average of 5.1 words per label.

bigram and syllable bigram scales does not achieve any improvement, by implication these scales are not complementary. Therefore, the addition of syllable bigrams to character bigrams and words cannot further improve retrieval performance.

8. CONCLUSIONS

In this work we have extended the HMM-based retrieval model to incorporate the cross-language translation component for CLIR. After the introduction of intermediate queries and application of some simplifications, the HMM-based retrieval model was successfully reformulated for cross-language retrieval. Furthermore, this extended HMM-based retrieval model was also modified to facilitate retrieval at the subword scales. These extensions enable CLIR using the HMM-based retrieval model to be carried out at multiple indexing scales.

Experiments on an English-Mandarin CL-SDR task were performed using the extended HMM-based retrieval models at the character bigram, syllable bigram, and word scales. Due to the incompleteness and inconsistency between the translation dictionary and the recognition vocabulary, some words were not translatable. Retrieval at the subword bigram scale recovered terms at subword scales from some untranslatable words. At the character bigram scale, better retrieval was achieved because there were recognition out-of-vocabulary (OOV) words and partial recognition errors. By applying multi-scale retrieval to CL-SDR, the combined specificity of words and robustness of subwords can achieve further improvement in retrieval performance over the composite scales.

REFERENCES

BAI, B. R., CHEN, B., AND WANG, H. M. 2000. Syllable-based Chinese text/spoken document retrieval using text/speech queries. *J. Pattern Recogn. Artif. Intell. 4*, 603–616.

BALLESTEROS, L. AND CROFT, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retireval. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 84–91.

BARTELL, B. T., COTTRELL, G. W., AND BELEW, R. K. 1994. Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 173–181.

BBN. 2000. Identifinder(TM). http://www.bbn.com/speech/identifinder.html.

BELKIN, N. J., KANTOR, P., FOX, E. A., AND SHAW, J. A. 1995. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage. 31*, 431–448.

BERGER, A. AND LAFFERTY, J. 1999a. Information retrieval as statistical translation. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 222–229.

BERGER, A. AND LAFFERTY, J. 1999b. The Weaver system for document retrieval. In *Proceedings of the 8th Text REtrieval Conference*. NIST, 163–174.

BRASCHLER, M., KRAUSE, J., PETERS, C., AND SCHAUBLE, P. 1998. Cross-language information retrieval (CLIR) track overview. In *Proceedings of the 7th Text REtrieval Conference*. NIST.

CHEN, A. 2000. Phrasal translation for English-Chinese cross language information retrieval. In *Proceedings of Workshop on English-Chinese Cross Language Information Retrieval at the 2000 International Conference on Chinese Language Computing*. 195–202.

CHEN, A. 2001. Berkeley at NTCIR-2: Chinese, Japanese, and English IR experiments. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*. 32–39.

CHEN, A., JIANG, H., AND GEY, F. 2000. English-Chinese cross-language IR using bilingual dictionaries. In *Proceedings of the 9th Text REtrieval Conference*. NIST.

CHEN, B., WANG, H. M., AND LEE, L. S. 2000. Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. 1771–1774.

CHEN, B., WANG, H. M., AND LEE, L. S. 2001. An HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval. In *Proceedings of the 7th European Conference on Speech Communication and Technology*. Vol. 2. 1045–1048.

FOX, E. A. AND SHAW, J. 1993. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference*. NIST, 243–252.

GAO, J., NIE, J. Y., XUN, E., ZHANG, J., ZHOU, M., AND HUANG, C. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 96–104.

GAO, J., NIE, J. Y., ZHANG, J., AND XUN, E. 2000. TREC-9 CLIR experiments at MSRCN. In *Proceedings of the 9th Text REtrieval Conference*. NIST, 343–353.

GAROFOLO, J. S., AUZANNE, C. G. P., AND VOORHEES, E. M. 1999. The TREC spoken document retrieval track: A success story. In *Proceedings of the 8th Text REtrieval Conference*. NIST, 107–129.

GAROFOLO, J. S., VOORHEES, E. M., AUZANNE, C. G. P., STANFORD, V. M., AND LUND, B. A. 1998. 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of the 7th Text REtrieval Conference*. NIST, 79–89.

GAROFOLO, J. S., VOORHEES, E. M., STANFORD, V. M., AND JONES, K. S. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*. NIST, 83–91.

GREFENSTETTE, G. 1998. *Cross-Language Information Retrieval*. Kluwer Academic, Boston MA.

HAUPTMANN, A. G., SCHEYTT, P., WACTLAR, H. D., AND KENNEDY, P. E. 1998. Multi-lingual Informedia: A demonstration of speech recognition and information retrieval across multiple languages. In *Proceedings of 1998 Broadcast News Transcription and Understanding Workshop*.

HIEMSTRA, D. 2000. Using language models for information retrieval. Ph.D. thesis, Centre for Telematics and Information Technology, University of Twente,.

HUANG, S., BIAN, X., WU, G., AND MCLEMORE, C. 1997. CALLHOME Mandarin Chinese lexicon. Tech. Rep., Linguistic Data Consortium, [online] http://www.ldc.upenn.edu/Catalog/LDC96L15.html.

HULL, D. A. AND GREFENSTETTE, G. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 49–57.

KWOK, K. L. 1997. Comparing representations in Chinese information retrieval. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 34–41.

KWOK, K. L. 1999. English-Chinese cross-language retrieval based on a translation package. In *Proceedings of the Workshop of Machine Translation for Cross Language Information Retrieval, Machines Translation Summit VII*.

LDC. 2000. Project topic detection and tracking phase two (TDT-2). Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/TDT2.

LEVOW, G. A. AND OARD, D. W. 2000. Translingual topic tracking: Applying lessons from the MEI project. In *Proceedings of the 2000 Topic Detection and Tracking Workshop*.

LO, W. K., SCHONE, P., AND MENG, H. M. 2001. Multi-scale retrieval in MEI: an English-Chinese translingual speech retrieval system. In *Proceedings of the 7th European Conference on Speech Communication and Technology*. Vol. 2. 1303–1306.

MAKHOUL, J., KUBALA, F., LEEK, T., LIU, D., NGUYEN, L., SCHWARTZ, R., AND SRIVASTAVA, A. 2000. Speech and language technologies for audio indexing and retrieval. *Proc. IEEE 88*, 1338–1353.

MATEEV, B., MUNTEANU, E., SHERIDAN, P., WECHSLER, M., AND SCHAUBLE, P. 1997. ETH TREC-6: Routing, Chinese, cross-language and spoken document retrieval. In *Proceedings of the 6th Text REtrieval Conference*. NIST, 623–636.

MENG, H. M., CHEN, B., GRAMS, E., KHUDANPUR, S., LO, W. K., LEVOW, G. A., OARD, D., SCHONE, P., TANG, K., WANG, H. M., AND WANG, J. Q. 2000. Mandarin-English information (MEI): Investigating

translingual speech retrieval. Tech. Rep., Johns Hopkins Univ., Baltimore, MD. Final report: [online] http//www.clsp.jhu.edu/ws2000/final_reports/mei.

MENG, H. M., CHEN, B., KHUDANPUR, S., LEVOW, G. A., LO, W. K., OARD, D., SCHONE, P., TANG, K., WANG, H. M., AND WANG, J. Q. 2001. Mandarin-English information (MEI): Investigating translingual speech retrieval. In *Proceedings of the 2001 Human Language Technology Conference*.

MENG, H. M., LO, W. K., LI, Y. C., AND CHING, P. C. 2000. Multi-scale audio indexing for Chinese spoken document retrieval. In *Proceedings of the 6th International Conference on Spoken Language Processing*. Vol. IV. 101–104.

MILLER, D. R. H., LEEK, T., AND SCHWARTZ, R. M. 1998. BBN at TREC7: Using hidden Markov models for information retrieval. In *Proceedings of the 7th Text REtrieval Conference*. NIST, 133–142.

NG, K. 2000. Subword-based approaches for spoken document retrieval. *Speech Commun. 32*, 157–186.

NIE, J. Y. AND REN, F. 1999. Chinese information retrieval: using characters or words? *Inf. Process. Manage. 35*, 443–462.

NIE, J. Y., SIMARD, M., ISABELLE, P., AND DURAND, R. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts. In *Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 74–81.

PIRKOLA, A., HEDLUND, T., AND KESKUSTALO, H. 2001. Dictionary-based cross-language information retrieval: problems, methods and research findings. *Inf. Retrieval 4*, 209–230.

RESNIK, P., OARD, D. W., AND LEVOW, G. A. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the 2001 Human Language Technology Conference*.

SALTON, G. AND MCGILL, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SCHAUBLE, P. AND SHERIDAN, P. 1997. Cross-language information retrieval (CLIR) track overview. In *Proceedings of the 6th Text REtrieval Conference*. NIST, 31–44.

SHERIDAN, P. AND BALLERINI, J. P. 1996. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 58–65.

SHERIDAN, P., BRASCHLER, M., AND SCHAUBLE, P. 1997. Cross-language information retrieval in a multi-lingual legal domain. In *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries*. 253–268.

SHERIDAN, P., WECHSLER, M., AND SCHAUBLE, P. 1997. Cross language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 99–108.

SONG, F. AND CROFT, W. B. 1999a. A general language model for information retrieval. In *Proceedings of the 8th International Conference on Information and Knowledge Management*. 316–321.

SONG, F. AND CROFT, W. B. 1999b. A general language model for information retrieval. In *Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 279–280.

VOGT, C. C. AND COTTRELL, G. W. 1999. Fusion via a linear combination of scores. *Inf. Retrieval 1*, 151–173.

WANG, H. M. 2000. Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese. *Speech Commun. 32*, 49–60.

WANG, H. M. AND CHEN, B. 2001. Comparison of word and subword indexing techniques for Mandarin Chinese spoken document retrieval. In *Proceedings of the 2nd Pacific-Rim Conference on Multimedia*.

WANG, H. M., MENG, H. M., SCHONE, P., CHEN, B., AND LO, W. K. 2001. Multi-scale audio indexing for translingual spoken document retrieval. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 605–608.

WOODLAND, P. C., JOHNSON, S. E., JOURLIN, P., AND JONES, K. S. 2000. Effect of out of vocabulary words in spoken document retrieval. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 372–374.

XU, J. AND WEISCHEDEL, R. 2000. TREC-9 cross-lingual retrieval at BBN. In *Proceedings of the 9th Text REtrieval Conference*. NIST, 106–115.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 334–342.

ZHAN, P. 1999. Dragon systems' 1998 broadcast news transcription system for Mandarin. In *Proceedings of the DARPA Broadcast News Workshop '99*.

Received September 2002; revised August 2003; accepted September 2003