Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 18 (2004) 163–179
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Mandarin–English Information (MEI): investigating translingual speech retrieval

Helen M. Meng [a,*], Berlin Chen [b], Sanjeev Khudanpur [c], Gina-Anne Levow [d], Wai-Kit Lo [a], Douglas Oard [d], Patrick Schone [e], Karen Tang [f], Hsin-min Wang [b], Jianqiang Wang [d]

[a] *Department of Systems Engineering and Engineering Management, Human-Computer Communications Laboratory, The Chinese University of Hong Kong, Shatin, NT, Hong Kong*
[b] *Academia Sinica, Taiwan*
[c] *Johns Hopkins University, USA*
[d] *University of Maryland at College Park, USA*
[e] *Department of Defense, USA*
[f] *Princeton University, USA*

Received 30 May 2001; received in revised form 8 August 2003; accepted 15 September 2003

Abstract

This paper describes the Mandarin–English Information (MEI) project, where we investigated the problem of cross-language spoken document retrieval (CL-SDR), and developed one of the first English–Chinese CL-SDR systems. Our system accepts an entire English news story (text) as query, and retrieves relevant Chinese broadcast news stories (audio) from the document collection. Hence, this is a cross-language and cross-media retrieval task. We applied a multi-scale approach to our problem, which unifies the use of phrases, words and subwords in retrieval. The English queries are translated into Chinese by means of a dictionary-based approach, where we have integrated phrase-based translation with word-by-word translation. Untranslatable named entities are transliterated by a novel subword translation technique. The multi-scale approach can be divided into three subtasks – multi-scale query formulation, multi-scale audio indexing (by speech recognition) and multi-scale retrieval. Experimental results demonstrate that the use of phrase-based translation and subword translation gave performance gains, and multi-scale retrieval outperforms word-based retrieval.

Keywords: Multi-scale spoken document retrieval; English–Chinese cross-language spoken document retrieval

* Corresponding author. Tel.: +852-2609-8327; fax: +852-2603-5505.
E-mail address: hmmeng@se.cuhk.edu.hk (H.M. Meng).

0885-2308/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2003.09.003

1. Introduction

Mandarin–English Information (MEI) is one of the first English–Chinese cross-language spoken document retrieval (CL-SDR) systems. Our objective is to develop technologies for cross-language and cross-media information retrieval, and our system can use an English text query to search for related Chinese audio documents. Massive quantities of audio and multimedia content are becoming increasingly available in the global information infrastructure. For example, www.real.com in mid-March 2001 listed over 2500 Internet-accessible radio and television stations. Of these, over a third were broadcasting in languages other than English. Monolingual speech retrieval is now practical, as evidenced by services such as SpeechBot, [1] and it is clear that there is a potential demand for CL-SDR if effective techniques can be developed. Since English and Mandarin Chinese are projected to be the two predominant languages of the Internet user population, [2] we have selected this language pair in our investigation of cross-language spoken document retrieval techniques. Such techniques enable the user to retrieve personally relevant content on demand, and across the barriers of language and media. Possible applications of this work include audio and video browsing, automated routing of information, and automatically alerting the user when special events occur.

The MEI task involves the use of an entire English newswire story (text) as query, to retrieve relevant Mandarin Chinese [3] radio broadcast news stories (audio) in the document collection. Such a retrieval context is termed query-by-example. It may be noted that retrieving documents by documents (i.e., long queries) may be considered as a simpler task than retrieving documents by short queries. This is because long queries provide more context that may help better define the topic of interest. The MEI system is illustrated in Fig. 1.

Our work demonstrates the use of a multi-scale paradigm for English–Chinese CL-SDR. The paradigm leverages off of our knowledge about the linguistic and acoustic–phonetic properties related to English and Mandarin. We unify multi-scale units for retrieval, and these units include phrases, words as well as subwords (Chinese characters and syllables). Our multi-scale paradigm aims to alleviate problems related to English–Chinese CL-SDR, such as

Multiplicity in translation: dictionary-based term-by-term translation may produce multiple translation alternatives, or no translations, e.g., for proper names. The use of phrases can often resolve translation ambiguity, e.g., ''human rights'' as a phrase has one translation; but ''human'' has about 30 translations, ''rights'' has about seven and together they form over 200 translation alternatives for ''human rights''. The use of phonetic translation can help address the out-of-vocabulary (OOV) problem in translation, e.g., ''Kosovo'' becomes /ke suo fu/ (科索沃), and its subword translation (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 by this method. Since Mandarin Chinese can be fully represented by about 400 base syllables or 6000 characters, [4] we can obtain full phonological/lexical coverage of the spoken documents using syllables/characters for indexing.

[1] http://speechbot.research.compaq.com.
[2] Source: Global Reach, 2000.
[3] Mandarin is the official Chinese dialect.
[4] According to the GB-2312 character set.

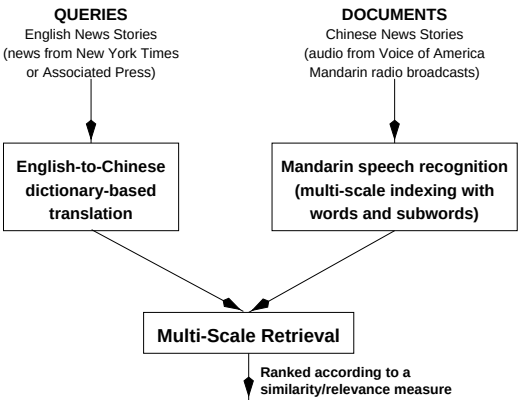

Fig. 1. Overview of the English–Chinese cross-lingual spoken document retrieval system. In this task, the query is formed from an entire English news story (text) from the New York Times or Associated Press. The spoken documents are Mandarin news stories (audio) from Voice of America news broadcasts. Multi-scale retrieval of the spoken documents is evaluated based on the relevance of the ranked list of spoken documents retrieved for each query.

Ambiguity in Chinese homophones: each Chinese character is pronounced as a single syllable, and the character/syllable mapping is many-to-many. Hence, there are many Chinese homophones, which can cause word-level confusions in SDR. For example, the bi-syllable word pronounced as /fu shu/ may be 富庶 (meaning "rich"), 負數 (meaning "negative number"), 複數 ("complex number"), or 覆述 ("repeat"). Homophones are often confused with one another during speech recognition, and such errors affect retrieval performance. The use of syllables provides a single representation for all homophones and thus help improve retrieval recall. However, syllables are phonological units and retrieval based on syllables cannot incorporate sufficient lexical information and this may lower retrieval precision. Referring to the example above, if the user is looking for documents that contain the word 複數, retrieval based on syllables alone cannot help but retrieve other documents containing 富庶, 負數 and 覆述 as well. Hence, our multi-scale paradigm maintains the use of words and characters in addition to syllables.

Ambiguity in Chinese word tokenization: the Chinese word contains one or more characters, with no explicit word delimiter. Word tokenization has much ambiguity, which can cause word-level mismatches between queries and documents in retrieval. Consider the following character string with several plausible word segmentations:

這一晚 會 如常 舉行
(Meaning: It will take place tonight as usual.)
這一 晚會 如常 舉行
(Meaning: The evening banquet will take place as usual.)
這一 晚會 如 常 舉行
(Meaning: If the evening banquet take place often.)

This problem of word tokenization ambiguity can be addressed by retrieving based on overlapping character *n*-grams. [5]

[5] The overlapping character bigrams of the above example is "這一 一晚 晚會 會如 如常 常舉 舉行".

Speech recognition errors: speech recognition output is imperfect. Errors may be caused by OOV words (Woodland et al., 2000) or acoustic confusions among in-vocabulary words (especially with respect to homophones). It is believed that SDR based on syllables can improve robustness against recognition errors in retrieval by improving the recall when documents contain recognition errors due to homophones.

As can be seen, our multi-scale paradigm involves the use of variable-sized units. English phrases (e.g. "human rights", "guiding principles", etc.) are translated intact during the query translation procedure in order to reduce translation ambiguity. The translated phrases to words in Chinese are used in subsequent retrieval. It should be noted that there is no clear distinction between phrases and words in Chinese. In addition to using words, we also use overlapping character *n*-grams, where the overlap aims to handle tokenization ambiguity, and the *n*-gram serves to capture some sequential (lexical) constraints. Since each character is pronounced as a syllable in Chinese, overlapping character *n*-grams can be converted to overlapping syllable *n*-grams for retrieval. [6] As mentioned above, the use of syllables can handle the OOV problem in recognition as well as ambiguity due to Chinese homophones. Characters and syllables are subword units for the Chinese language. Hence, our multi-scale approach unifies phrases, words and subwords for the English–Chinese cross-language SDR problem.

2. Previous work

CL-SDR is a relatively new research area, but is rapidly gaining momentum. One of the earliest works was reported on retrieval of German spoken news documents with French text (Sheridan and Ballerini, 1996), and the retrieval of Serbo-Croatian news broadcasts with English text (Hauptmann et al., 1998). CL-SDR integrates speech recognition, machine translation and information retrieval technologies to accomplish the task. It merges two lines of research which have evolved separately until recently. Hence, we can draw upon previous techniques used in (i) cross-lingual information retrieval (CLIR), which generally involves textual queries and documents, and closely couples machine translation with information retrieval technologies; and (ii) spoken document retrieval (SDR), which generally involves textual queries and spoken documents, and closely couples speech recognition with information retrieval technologies. Both lines of research have been driven by the TREC [7] evaluations in recent years.

We referenced techniques used in CLIR to cross the language barrier for retrieval. This includes query translation, document translation, interlingual techniques and cognate matching (comparing the similarities between the spellings and pronunciations in order to match untranslatable terms) (Oard and Diekema, 1998). Query translation is popular, because document translation may be impractical for large and remote collections (Carbonell et al., 1997). The main resource for query translation is the bilingual machine-readable dictionary (Oard and Diekema, 1998). Dictionary-based translation is hampered by out-of-vocabulary (OOV) words and ambiguities from multiple translation alternatives. The OOV problem may be alleviated by the use of cognates. For

[6] The overlapping syllable bigrams of the above example is "*zhe-yi yi-wan wan-hui hui-ru ru-chang chang-ju ju-xing*".
[7] http://trec.nist.gov.

example, Buckley et al. (1997) performed English–French CLIR simply by "spelling corrections", since English and French share many cognates that have similar spellings. Davis (1996) used syntactic information like part-of-speech to help resolve translation ambiguities. Ballesteros and Croft (1998) found that it is desirable to identify phrases by parsing before translation. This is because phrases may not have the compositional meaning of its constituent words, and phrase translation is often more precise than term-by-term translation.

As regards spoken document retrieval, a popular approach (Garofolo et al., 2000) is to couple word transcription using large-vocabulary speech recognition (LVCSR) with information retrieval techniques. However, word transcription inevitably encounters the open vocabulary (OOV) problem in recognition (Woodland et al., 2000), especially when new words or proper names appear in spoken audio. Subword transcription has been used to handle this problem, where spoken documents may be indexed with phoneme *n*-grams (Ng, 2000; Wechsler and Schauble, 1995) or phone lattices (Brown et al., 1996). Subwords can enhance recall by providing complete phonological coverage of the spoken audio, thus circumventing the OOV problem.

Our current work on English–Mandarin CL-SDR marks an initial effort in English–Chinese CL-SDR. English newswire stories are translated into Chinese to query for topically related Mandarin broadcast news stories. We used query translation for crossing the language barrier and augmented word-based indexing with subword-based indexing for spoken document retrieval.

3. The TDT collection

We used the Topic Detection and Tracking (TDT) collection for this work. 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, because we are performing retrospective retrieval, which permits knowledge of the statistics for the entire document collection. Nevertheless, the TDT collection serves as a valuable resource for our work. The TDT multi-lingual collection includes English and Mandarin news text as well as (audio) broadcast news. Most of the Mandarin audio data are transcribed by the Dragon automatic speech recognition system (Zhan et al., 1999). All news stories are exhaustively tagged with event-based topic labels, which serves as the relevance judgements for performance evaluation of our CL-SDR work. We used the TDT-2 corpus as our development test set, and TDT-3 as our evaluation test set. Table 1 describes the content in these collections.

Table 1
Statistics of TDT-2 and TDT-3 – our development and evaluation datasets

	TDT-2 (Dev. set)	TDT-3 (Eval. set)
English news (New York Times or Associated Press)	17 topics, variable no. of exemplars (average length is 270 phrases)	56 topics, variable no. of exemplars (average length is 543 phrases)
Mandarin audio news (Voice of America)	2265 stories, 46.0 h	3371 stories, 98.4 h

The Mandarin audio documents have been transcribed by the Dragon automatic speech recognition system.

4. The multi-scale paradigm

This section describes our multi-scale paradigm in detail. It is divided into several subtasks – query formulation, audio indexing and retrieval. As described earlier, we make use of phrases, words, overlapping character *n*-grams and overlapping syllable *n*-grams in retrieval. We mainly use subword bigrams since previous work (Kwok and Grunfeld, 1996; Wang, 2000; Meng et al., 2000a) indicated that bigrams are most effective (among the different *n*-grams) for retrieval.

4.1. Multi-scale query formulation

4.1.1. Query term selection

In the MEI task, the query consists of an entire English news story. Such queries tend to be long, and not all query terms are important for retrieval. The first step in query formulation is to select English terms from the query exemplar. [8] First we excluded all stopwords, based on the English default stopword list used by the InQuery retrieval engine (Callan et al., 1992; Broglio et al., 1994).

Then we ranked all of the terms in the exemplar and all the single word components of multi-word units according to how well they distinguish the exemplar from a background model. This model is formed from the terms of approximately 1000 temporally earlier documents in the English collection from which the exemplars were drawn. We used a χ^2 test in a manner similar to that used in Schuetze et al. (1995) to select these terms. Specifically, we apply a χ^2 test to a table containing the number of relevant and non-relevant documents in which each term appears and the number of relevant and non-relevant documents in which the term does not occur. We compute the measure according to the equation:

$$\chi^2 = \frac{N \cdot (N_{\mathrm{r},+} \cdot N_{\mathrm{n},-} - N_{\mathrm{r},-} \cdot N_{\mathrm{n},+})^2}{(N_{\mathrm{r},+} + N_{\mathrm{r},-})(N_{\mathrm{n},+} + N_{\mathrm{n},-})(N_{\mathrm{r},+} + N_{nr,+})(N_{\mathrm{r},-} + N_{\mathrm{n},-})}, \tag{1}$$

where N refers to the number of documents as characterized by the subscripts with the following conditions: r is relevant, n is non-relevant, + means the current terms appears in the document, and − indicates that it does not. This value increases when terms are more strongly associated (or disassociated) with a topic in contrast to those that appear with similar frequency in both relevant and non-relevant documents. This statistic is symmetric, assigning equal value to terms that help to recognize known relevant stories and those that help to reject the other contemporaneous stories. We limited our choice to terms that were positively associated with the known relevant training stories.

4.1.2. Query translation

Named entities in our English query exemplars have been tagged by the BBN Identifinder (Bikel et al., 1997) system. Examples of named entities include ''U.N. Security Council'', and ''partners of Goldman, Sachs and Co.'' Additional multi-word expressions (e.g. ''human rights'',

[8] These may be multi-word units that are tagged with reference to our term list, as will be explained later.

"guiding principles", and "best interests") are identified by referring to our bilingual term list (BTL). This list is formed by combining LDC's English–Chinese bilingual term list [9] with translations extracted from the CETA (Chinese-English Translation Assistance) [10] dictionaries. Our BTL covers 200,000 distinct English terms. Among these English terms, some have multiple translations and there is a total of approximately 400,000 English-to-Chinese translation pairs. A multi-word expression (from Identifinder or our BTL) is treated as a "single term" in English term selection and query formulation procedures.

We traverse the tagged English text exemplar and, for each identified term, if it is on the list of selected terms, we translate it. This approach preserves term frequency information in the query. Translation proceeds on the phrasal scale, word scale, as well as the subword scale. For tagged named entities, we first attempt to translate the entity as a single unit by lookup in our BTL. If the named entity is not found, we translate the individual words one by one. For example, "security council" is present in the bilingual term list and can be translated directly. "First Bank of Siam", however, is not present and is translated word by word. All other terms are translated directly by searching the bilingual term list. We also incorporated a stemming backoff translation procedure to maximize matching with the translation dictionary (Oard et al., 2001).

4.1.3. Named entity transliteration

Despite the use of an extensive BTL for phrasal and word-based translation, there will inevitably be untranslatable terms. These are often named entities (names of people, places, locations and organizations), since we are dealing with a topically diverse domain. These untranslatable named entities need to be salvaged since they tend to be important for retrieval. Chinese translations of foreign names often strive to attain phonetic similarity, though the mapping may be inconsistent. For example, consider the translation of "Kosovo" – sampling Chinese newspapers in China, Taiwan, Hong Kong and Singapore produced the following translations:

(1) 科索沃 /ke-suo-wo/,
(2) 科索佛 /ke-suo-fo/,
(3) 科索夫 /ke-suo-fu/,
(4) 科索伏 /ke-suo-fu/, or
(5) 柯索佛 /ke-suo-fo/.

To this end, we have developed a technique for subword translation. This is another research contribution in the MEI project. In designing the subword translation procedure, we applied our knowledge in acoustic–phonetics and phonology related to both English and Chinese, we also applied machine learning techniques and other techniques used in speech recognition. The aim of subword translation is to transliterate named entities in the queries and represents them in the phonetic space, and if the document collection is also indexed in the phonetic space, we can perform matching in the phonetic space for retrieval. In this way, we salvage the use of named entities that are otherwise untranslatable and cannot be used for retrieval. Details of this technique are described in Meng et al. (2001). We provide a succinct description in the following.

[9] http://www.ldc.upenn.edu.
[10] Licensed as Optilex by MRM Inc. 3910 Knowles Avenue, Kensington, MD 20895, USA.

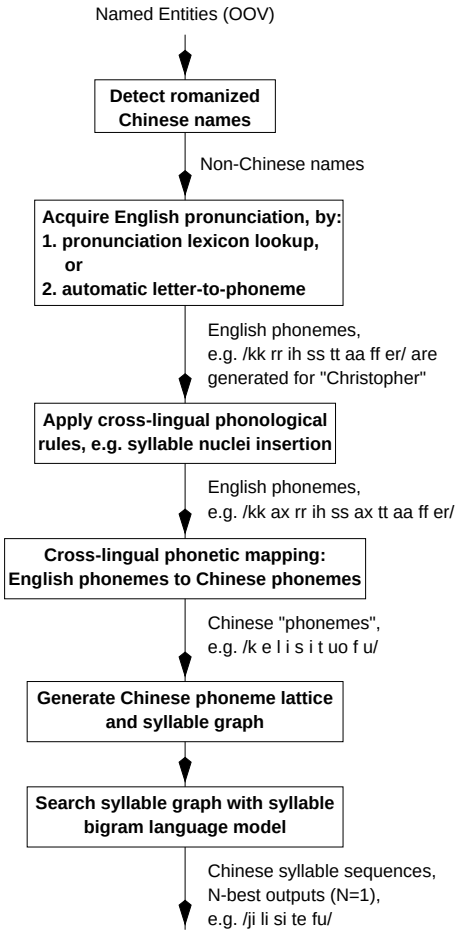

Fig. 2. Overview of our subword translation process for handling untranslatable named entities in the query exemplars.

Fig. 2 presents an overview of the named entity transliteration process. We examine units in our query exemplars that are tagged by the BBN Identifinder system, and those absent from our Bilingual Term List (BTL, i.e., translation dictionary) are processed by our transliteration system. As shown in Fig. 2, subword translation begins by discriminating between Chinese names and non-Chinese names. Chinese names are often represented in English by means of their syllable pinyin transcription, e.g., ''Diaoyutai'' consists of the three syllables /diao/, /yu/ and /tai/. As mentioned, there is a finite set of pinyin syllables, so identification of Chinese names is accomplished by string matching, with reference to the syllable inventory. Non-Chinese names are modeled as a single category, which is an over-generalization for the sake of simplicity. We attempt to look up the English pronunciation of the non-Chinese names. [11] Failing that, we generate the English pronunciation automatically from the spelling, using letter-to-sound rules

[11] We used the pronunciation lexicon PRONLEX provided by the LDC.

acquired by the transformation-based error-driving learning technique (Brill, 1995). For example, we can generate the English phoneme pronunciation /kk rr ih ss tt aa ff er/ from the spelling "Christopher" (see Fig. 2).

Since Chinese is a monosyllabic language, transliteration of names to Chinese syllables should abide by a set of phonological rules. We have hand-designed a set of cross-lingual phonological rules that partially transforms an English pronunciation into a Chinese pronunciation. The transformation involves such processes as syllable nuclei insertion to separate consonant clusters. This is followed by an automatic mapping of English phonemes to Chinese "phonemes", a procedure we termed cross-lingual phonetic mapping (CLPM). This is also an automatic procedure in which we have applied the transformation-based error-driven learning technique. By this time, our process has transformed the English phonemes into Chinese phonemes, e.g., /kk rr ih ss tt aa ff er/ is transformed into /k e l i s i t uo f u/ (see Fig. 2). This is essentially a phoneme translation procedure. The technique of subword translation based on pronunciation lexicons has previously been applied to English/Japanese and English/Arabic translation (Knight and Graehl, 1997; Stalls and Knight, 1998). Ours is one of the first attempts in phoneme translation for English and Chinese and incorporating the automatic letter/phoneme generation technique.

In order to obtain name transliteration alternatives from this Chinese phoneme sequence, we borrow ideas from lexical access in speech recognition. By expanding each Chinese phoneme into its list of acoustically confusable counterparts, we obtain a Chinese phoneme lattice. Applying the Chinese syllable constraints to the Chinese phoneme lattice produces a Chinese syllable lattice, and searching the syllable lattice with a syllable bigram language model can produce N-best hypotheses of Chinese syllable sequences. [12] These form the output of our names transliteration procedure, e.g., /ji li si te fu/ (see Fig. 2). It is interesting to note that when we use character bigrams in place of syllable bigrams, our transliteration algorithm can produce subword translations in terms of character sequences, e.g., 基里斯特弗. In other words, the character bigrams can import lexical knowledge while searching through the syllable lattice and decode the lattice by generating output character sequences. This means that English proper names can be automatically transliterated into Chinese. The transliterated names can be used for word-based and character-based retrieval.

Fig. 3 shows some examples of the generated outputs of our cross-lingual phonetic mapping procedure. As can be seen in Fig. 3, we are able to salvage some of the syllable bigrams which partially index the named entity concerned.

For each bilingual name pair, the hypothesized syllable sequence is generated from the English name spelling by our transliteration procedure. The reference syllable sequence is obtained by pronunciation lexicon lookup based on the Chinese name characters.

4.1.4. Multi-scale query construction

The input to our query construction process is a bag of English query terms. Multi-scale query construction integrates the translated phrases, named entities, individual translated words as well as translated syllables. Hence, the output of our query construction process is a representation

[12] We set $N = 1$ for simplicity. We believe it will be worthwhile to investigate the incorporation of alternate machine transliteration hypotheses as a future step.

CHARLES DAVENPORT
查爾斯達文波特
Ref: /cha er si da wen bo te/
Hyp: /zha e er si di wei wu n bo e te/

CYRIL SUK
西里爾蘇克
Ref: /xi li er su ke/
Hyp: /xi li er su ke/

DON FRANCIS
唐弗朗西斯
Ref: /tang fu lang xi si/
Hyp: /deng fu neng si/

HARVEY GRANT
哈維格蘭特
Ref: /ha wei ge lan te/
Hyp: /ha e fu ge lan te/

Fig. 3. Examples of generated outputs from our named entity transliteration process. This is an excerpt of a held-out set of English names and their Chinese translations that are obtained from the web.

which includes Chinese words, subwords, or a mixture of both. Subwords refer to character *n*-grams (to capture sequential constraints) or syllable *n*-grams. This process is depicted in Fig. 4.

4.2. Multi-scale audio indexing

The Dragon large-vocabulary continuous speech recognizer (Zhan et al., 1999) provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3). Based on these word transcriptions, we can use the same procedures as in query formulation to obtain overlapping character bigrams and overlapping syllable bigrams from the word transcriptions. Hence, we can index our audio on the word, character and syllable scales. To assess the performance level of the recognizer, we spot-checked a fraction of the TDT-2 test set (about 23 h) by comparing the Dragon recognition hypotheses with the anchor scripts (treated as ground truth), [13] and obtained error rates of 18.0% (word), 12.1% (character), and 7.9% (syllable). Spot-checking approximately 27 hours of the TDT-3 test set gave error rates of 19.1% (word); 13.0% (character) and 8.6% (syllable). We feel that the Dragon recognizer has a respectable performance level (Wang et al., 2001).

4.3. Multi-scale retrieval

We use the InQuery retrieval engine developed by the University of Massachusetts (Callan et al., 1992). [14] InQuery uses a probabilistic belief network as the main data structure behind its query language.

[13] We treat the anchor script as ground truth for evaluating speech recognition performance. It should be noted that the anchor scripts may not be equivalent to a verbatim transcription of the audio documents (which is unavailable). However, we have tried to make the best use of the available resources in the TDT corpora.
[14] We used InQuery with a trivial modification to handle two-byte characters.

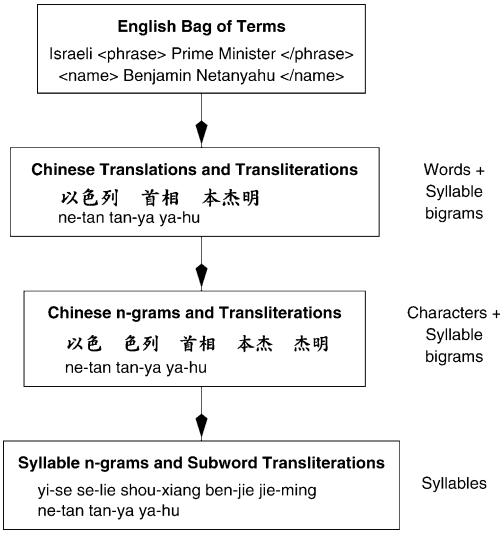

Fig. 4. The process of multi-scale query construction in our system. The query representations at various stages of processing may be used. The representations seek to integrate information from phrase-based translation, word-based translation, subword translation and overlapping character/syllable *n*-grams which alleviates the problem of word tokenization ambiguities. Transformation from characters to syllables references a Chinese pronunciation lexicon (the LDC CALLHOME lexicon).

A key feature that we have employed is the "balanced query" mechanism (Leek et al., 2000; Levow and Oard, 2000). Suppose we had a query given by $E_1, E_2, \ldots, E_n$, where E_i represent the English query terms, and that E_1 has three possible Chinese translations, C_{11}, C_{12} and C_{13}. With balanced translation, the belief value for E_1 in the Chinese document will be computed as the mean of the belief values for C_{11}, C_{12} and C_{13} in that document. Repeating the same process for additional terms produces a set of belief values for each English query term with respect to every Chinese document. The InQuery #sum operator implements this computation, so a balanced translation of the query would be represented as $\#\mathrm{sum}(\#\mathrm{sum}(C_{11}, C_{12}, C_{13})\#\mathrm{sum}(C_{21}, C_{22}) \cdots \#\mathrm{sum}(C_{n1}, C_{n2}, C_{n3}))$ in InQuery, with the outer #sum operator being the typical way of combining belief values across query terms and the inner #sum operators implementing balanced translation. Balanced translation prevents a query term that has a disproportionate number of translations from dominating the computing of the scores by which the ranked list of documents are sorted.

Our main strategy for multi-scale retrieval is as follows: retrieval proceeds for each scale (word, characters and syllables) individually, and each scale produces its own retrieved list of documents, ranked in decreasing order of retrieval status values. We can then combine these ranked lists into a single ranked list by a linear combination of their respective scores. The weights used in linear combination are selected based on empirical trials within the range of 0–1 at 0.1 intervals, and running retrieval experiments with the development set (TDT-2). Such linear combination is term *loose coupling*. An alternative strategy, *tight coupling*, integrates different unit types into a hybrid query/document representation, and then produces a single ranked list in retrieval. In our experiments, tight coupling gave significantly lower performance than loose coupling. This is partly

due to the constraint of the retrieval engine which does not permit easy reweighting of components of the document vectors. Hence, we are unable to tune for a better configuration for tight coupling. Comparison between loose and tight coupling is further investigated in Lo (2002).

5. Experiments

5.1. Evaluation criterion

In order to evaluate our retrieval performance, we use a variant of the non-interpolated mean average precision as our evaluation metric.

We compute the non-interpolated mean average precision for a 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 values is the average precision for that particular query. An average is then made across all queries in the batches for each of the topics. Taking another average over all queries produce a single value as our evaluation metric. Eq. (2) summarizes the process

$$\text{mean average precision} = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{M_i}\sum_{j=1}^{M_i}\left\{\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{\text{rank}_{ijk}}\right\}\right\}, \tag{2}$$

where mean average precision is across our batches of queries, L is the number of topics; M_i is a sample of the exemplars for topic i; N_i is number of relevant documents for topic i; and rank_{ijk} is the rank of the kth relevant document retrieved for exemplar j of topic i. In order to minimize potential random effect among different topics, we used 17 topics (i.e., $L = 17$) for evaluation. In addition, we also used up to 12 exemplars (i.e., $M_i = 12$) for each of the 17 topics whenever available.

5.2. Tuning with the development test set

The TDT-2 collection was our development test set, which forms our basis for tuning free parameters, e.g., the number of query terms to include, the number of translation alternative to use, the linear combination weights used in our multi-scale retrieval strategy, etc. In addition, the TDT-2 audio collection was also used in training the MEI recognizer to optimize its recognition performance.

For our multi-scale query construction, we translated query terms (after stopword removal) that are positively associated with the known relevant training documents. Translated terms are then combined with a #sum operator to achieve balanced queries for retrieval. Based on a histogram of translations, it suggested that almost all terms had 50 or fewer translations. In order to eliminate the extremity of outliers, we limited the maximum number of translations to 50. In applying our subword translation technique, we took all tagged named entities from the TDT-2 collection and translated them at the subword level. We took the 200 most frequent translated named entities to augment the queries in both the TDT-2 and TDT-3 runs. Hence, the development test set should have greater leverage based on subword translation.

Table 2
Effect of phrase-based translation in CL-SDR retrieval performance

Query translation method	Retrieval performance (mAP)	Relative improvement
Word-by-word translation	0.350	–
Augmented with phrase-based translation	0.392	12% (statistically significant)

Results are based on TDT-2 only.

Table 3
English–Chinese CL-SDR results for word-based retrieval, in comparison with retrieval based on overlapping character bigrams

	Word-based retrieval (mAP)	Character bigrams (mAP)	Syllable bigrams (mAP)
TDT-2 (dev. test)	0.471	0.522	0.468
TDT-3 (eval. test)	0.462	0.477	0.422

5.3. Experimental results

In the MEI project, we have investigated a variety of issues related to English–Chinese CL-SDR. This paper focuses on the use of phrases in query translation, the merits of multi-scale retrieval in comparison with word-based retrieval, and the use of subword translation to salvage untranslatable named entities. Key results are presented in the following.

5.3.1. Phrase-based translation

Our investigation of phrase-based translation took place in an early phase on our project. Phrases are first identified from the English queries by a left-to-right greedy algorithm (maximum matching). This ensures that compound words will have the proper translation whenever identified. For example, "Prime Minister" will be translated incorrectly if it is not considered as a phrase. Purely word-based translation gave a retrieval performance of mean average precision (mAP) = 0.35 using the TDT-2 development set. The addition of phrase-based translation raised it to 0.392. The 12% relative improvement was statistically significant, based on a paired two-tailed t test on the means across exemplars of each topic, with $p < 0.05$. These results are tabulated in Table 2. Thereafter, we have always included phrase-based translation in our experiments.

5.3.2. Multi-scale retrieval

Overlapping character bigrams gave the best retrieval performance overall, and even outperformed words. The trend is consistent across our development and evaluation test sets. Results are shown in Table 3.

The relative difference of 3.2% (w.r.t. TDT-3) is also statistically significant, based on a paired two-tailed t test with $p < 0.05$. This suggests that character bigrams may be effective in ameliorating the problem of word tokenization ambiguities. We also tried to loosely couple the retrieval lists based on words and character bigrams, using weights optimized from TDT-2, and tested on TDT-3.

This gave a performance of mAP = 0.482 on TDT-3, which is better than retrieval on each scale alone and the improvement is statistically significant ($p < 0.05$). Overlapping syllable

Table 4
Investigation into the use of subword translation to salvage untranslatable named entities for CL-SDR

	TDT-2 performance (mAP)	TDT-3 performance (mAP)
Words only	0.464	0.462
Words with subword translation	0.471	0.462
Character bigrams only	0.514	0.475
Character bigrams with subword translation	0.522	0.477

bigrams performed below words – TDT-2 and TDT-3 results were at 0.468 and 0.422, respectively. [15]

5.3.3. Subword translation

Subword translation improved retrieval performance across multiple unit types. We reference the named entities that were tagged by Identifinder but cannot be translated with our BTL, and we extracted the 200 most frequent ones to be processed by subword translation. Results based on the words and character bigrams (the two units giving the highest retrieval performance) are shown in Table 4.

While such improvements were not statistically significant, they were consistent across the units. We expect that the benefits of subword translation will be greater if the technique is used for a greater number of (untranslatable) terms, or if we need to retrieve collections for which our bilingual term list has lower coverage.

6. Conclusions

In this paper, we have described the Mandarin-English Information (MEI) project (Meng et al., 2000b), where we developed one of the first English–Chinese cross-language spoken document retrieval systems. Our system accepts an entire English news story (text) as query, and retrieves relevant Chinese broadcast news stories (audio) from the document collection. Hence, this is a cross-language and cross-media retrieval task. This task uses long queries (entire documents) from a vast collection. This may be considered as a simpler task than retrieval using short queries because long queries provide more context that help better define the topic of interest. Hence, special processing steps in query formulation aim to extract pertinent information for retrieval. We applied a multi-scale approach to our problem, which unifies the use of phrases, words as well as subword in retrieval. The English queries are translated into Chinese by means of a dictionary-based approach, where we have integrated phrase-based translation with word-by-word translation. Untranslatable named entities are transliterated by a novel subword translation technique. This can automatically generate a Chinese pinyin representation that sounds similar to the name's original pronunciation. The multi-scale approach can be divided into three subtasks – multi-scale

[15] Syllables present a compact inventory of units (400 in total) with full phonological coverage for Mandarin. Homophones have identical indices in terms of syllables. The use of syllables tends to improve recall-robustness but degrade precision-robustness in retrieval performance.

query formulation, multi-scale audio indexing and multi-scale retrieval. We experimented with the TDT collections, which have English newswire from New York Times and Associated Press, and Mandarin Chinese radio news broadcasts from Voice of America. The radio news is transcribed by Dragon's large-vocabulary continuous speech recognizer.

Experimental results show that augmenting word-by-word query translation with phrase-based translation brought statistically significant improvements in retrieval performance. Overlapping character bigrams gave the best retrieval results overall, and outperformed words. Words, in turn, performed better than overlapping syllable bigrams. Using both words and character bigrams together (by loose coupling) gave better retrieval performance than each alone. In addition, both word-based retrieval and character-based retrieval benefit from the use of subword translation to salvage untranslatable named entities. These results suggest that our multi-scale approach is promising and applicable to the English–Chinese CL-SDR task. It should also be possible to leverage off of our experience in a translingual setting, which involves SDR across other language pairs.

Acknowledgements

The MEI project was conducted during the Johns Hopkins University Summer Workshop 2000 (an NSF Workshop). [16] We thank Erika Grams from Advanced Analytic Tools for her active participation. This work is supported by the NSF Grant No. IIS-00712125, Gina's work was supported by the DARPA cooperative agreement N660010028910, and Berlin's participation was supported by Academia Sinica (Taiwan), as well as the research grant (88-S-0128) from Professor Lin-Shan Lee of National Taiwan University. Continuation work beyond year 2000 was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. 4223/01E). We thank the Linguistic Data Consortium for providing the TDT Corpora. We also thank Charles Wayne, George Doddington, James Allan, John Garofolo, Hsin-Hsi Chen, Richard Schwartz and Ralph Weischedel for their help. We are grateful to Fred Jelinek and his staff at CLSP for organizing the workshop.

References

Ballesteros, L., Croft, B., 1998. Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71.

Bikel, D., Miller, S., Schwartz, R., Weischedel, R., 1997. Nymble: a high-performance learning name-finder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194–201.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. Computational Linguistics 21, 543–565.

Broglio, J., Callan, J., Croft, B., 1994. Inquery system overview. In: Proceedings of the TIPSTER Text Program (Phase I), pp. 47–67.

Brown, M., Foote, J., Jones, G., Young, S., 1996. Open vocabulary speech indexing for voice and video mail retrieval. In: Proceedings of the ACM Multimedia Conference, pp. 307–316.

[16] http://www.clsp.jhu.edu/ws2000/groups/mei/welcome.html.

Buckley, C., Mitra, M., Walz, J., Cardie, C., 1997. Using clustering and superconcepts within SMART: TREC 6. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), pp. 107–124.

Callan, J.P., Croft, W.B., Harding, S.M., 1992. The INQUERY retrieval system. In: Proceedings of the 3rd International Conference on Database and Expert Systems Applications, pp. 78–83.

Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R., Geng, Y., Lee, D., 1997. Translingual information retrieval: A comparative evaluation. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 708–714.

Davis, M., 1996. New experiments in cross-language text retrieval at NMSU's computer research lab. In: Proceedings of the Fifth Text Retrieval Conference (TREC-5), pp. 447–454.

Garofolo, J., Auzanne, G.P., Voorhees, E.M., 2000. The TREC spoken document retrieval task: a success story. In: Proceedings of the Recherche d'Informations Assistee par Ordinateur: Content-Based Multimedia Information Access Conference. Available from <www.nist.gov/speech/tests/sdr/sdr2000/>.

Hauptmann, A.G., Scheytt, P., Wactlar, H.D., Kennedy, P. E., 1998. Multi-lingual informedia: a demonstration of speech recognition and information retrieval across multiple languages. In: Proceedings of DARPA Workshop on Broadcast News Transcription and Understanding Workshop. Available from <www.cs.cmu.edu/People/alex/>.

Knight, K., Graehl, J., 1997. Machine transliteration. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 128–135.

Kwok, K.L., Grunfeld, L., 1996. TREC-5 English and Chinese Experiments using PIRCS. In: Proceedings of the Fifth Text Retrieval Conference (TREC-5), pp. 133–142.

Leek, T., Jin, H., Sista, S., Schwartz, R., 2000. The BBN crosslingual topic detection and tracking system. In: Proceedings of the 1999 Topic Detection and Tracking Workshop. Available from <www.nist.gov/speech/tests/tdt>.

Levow, G., Oard, D., 2000. Translingual topic tracking with PRISE. In: Proceedings of the 1999 Topic Detection and Tracking Workshop. Available from </speech/tests/tdt>.

Lo, W.K., 2002. Information fusion for monolingual and cross-language spoken document retrieval. Doctor of Philosophy Thesis, Department of Electronic Engineering, The Chinese University of Hong Kong.

Meng, H., Lo, W.K., Li, Y.C., Ching, P.C., 2000a. Multi-scale audio indexing for Chinese spoken document retrieval. In: Proceedings of the Sixth International Conference on Spoken Language Processing IV, pp. 101–104.

Meng, H., Chen, B., Gram, E., Khudanpur, S., Lo, W.K., Levow, G., Oard, D., Schone, P., Tang, K., Wang, H.M., Wang, J.Q., 2000b. Mandarin–English Information (MEI): Investigating Translingual Retrieval. Final Report for Johns Hopkins University Summer Workshop 2000.

Meng, H., Chen, B., Lo, W.K., Tang, K., 2001. Automatic named entity transliteration for English–Chinese cross-language spoken document retrieval. In: Proceedings of the 2001 Workshop on Automatic Speech Recognition and Understanding.

Ng, K., 2000. Subword-based approaches for spoken document retrieval. Speech Communication 32, 157–186.

Oard, D., Diekema, A., 1998. Cross-language information retrieval. Annual Review of Information Science and Technology 33, 223–256.

Oard, D., Levow, G., Cabezas, C., 2001. CLEF experiemnts at Maryland: statistical stemming and backoff translation. Lecture Notes in Computer Science (forthcoming).

Schuetze, H., Hull, D., Pedersen, J.O., 1995. A comparison of classifiers and document representations for the routing problem. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 229–237.

Sheridan, P., Ballerini, J.P., 1996. Experiments in multilingual information retrieval using the SPIDER system. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 58–65.

Stalls, B., Knight, K., 1998. Translating names and technical terms in Arabic text. In: Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages. Available from <www.isi.edu/natural-language/people/knight.html>.

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

Wang, H.M., Meng, H., Schone, P., Chen, B., Lo, W.K., 2001. Multi-scale audio indexing for translingual spoken document retrieval. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, CD-ROM proceedings.

Wechsler, M., Schauble, P., 1995. Speech retrieval based on automatic indexing. In: Proceedings of the Final Workshop on Multimedia Information Retrieval.

Woodland, P.C., Johnson, S.E., Jourlin, P., Jones, K. Sparck, 2000. Effect of out of vocabulary words in spoken document retrieval. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 372–374.

Zhan, P., Wegmann, S., Gillick, L., 1999. Dragon systems' 1998 broadcast news transcription system for Mandarin. In: Proceedings of the 1999 DARPA Broadcast News Workshop.