Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval

Helen Meng,[1] Berlin Chen,[2] Sanjeev Khudanpur,[3] Gina-Anne Levow,[4] Wai-Kit Lo,[1] Douglas Oard,[4] Patrick Schone,[5] Karen Tang,[6] Hsin-Min Wang[2] and Jianqiang Wang[4]
The Chinese University of Hong Kong,[1] Academia Sinica,[2] Johns Hopkins University,[3] University of Maryland at College Park,[4] US Department of Defense,[5] Princeton University[6]
Phone: +852.2609.8327
[1]hmmeng@se.cuhk.edu.hk

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
Cross-language, spoken document retrieval, English-Chinese

1. INTRODUCTION
Mandarin-English Information (MEI) is a research project conducted in the Johns Hopkins University Summer Workshop 2000. We have developed 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. Massive quantities of audio and multimedia content are becoming increasingly available in the global information infrastructures – 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 (speechbot.research.compaq.com), 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,[1] we have selected this language pair in our investigation of cross-language spoken document retrieval techniques. As multimedia content continues to grow in the global information infrastructure, we need to develop technologies which 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, spoken document retrieval, automated routing of information, and automatically alerting the user when special events occur.

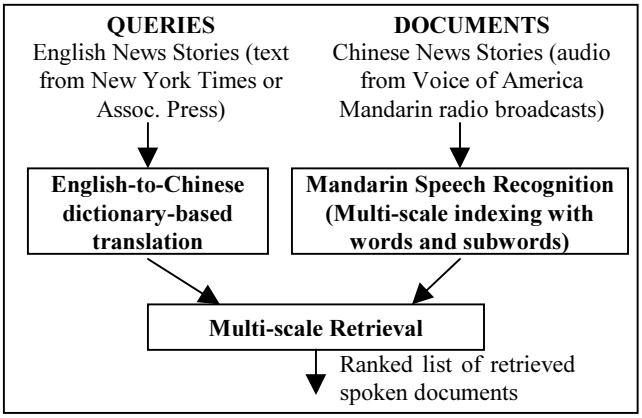

Figure 1. Overview of the our English-Chinese cross-lingual spoken document retrieval system. In this task, the query is formed from an entire English news story (text) from the New York Times or Associated Press. The spoken documents are Mandarin news stories (audio) from Voice of America news broadcasts.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

The MEI task involves the use of an entire English newswire story (text) as query, to retrieve relevant Mandarin Chinese[2] radio broadcast news stories (audio) in the document collection. Such a retrieval context is termed *query-by-example.* As illustrated in Figure 1, MEI integrates speech recognition, machine translation,

[1] Source: Global Reach, 2000.
[2] Mandarin is the official Chinese dialect.

and information retrieval technologies for English-Chinese CL-SDR.

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 Chinese. 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:

(i) *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 thirty translations, rights has about seven and together they form over two hundred translation alternatives for "human rights". The use of phonetic translation can help address the out-of-vocabulary problem in translation, e.g. Kosovo becomes /ke suo fu/ (科索沃), and its *subword translation* (pinyin transcription) can be utilized for SDR.

(ii) *Open vocabulary in recognition* – indexing spoken documents with word-based speech recognition is constrained to the recognizer's vocabulary. Out-of-vocabulary words (OOV) cannot be indexed by this method. Since Mandarin Chinese can be fully represented by about 400 base syllables or 6000 characters,[3] we can obtain full phonological / lexical coverage of the spoken documents using syllables / characters for indexing.

(iii) *Ambiguity in Chinese homophones* – each Chinese character is pronounced as a single syllable, and the 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 ("negative number"), 復數 ("complex number" or "plural"), and 覆述 ("repeat"). Homophones are often confused with one another during speech recognition, and the use of syllables for retrieval offers a solution to such recognition errors.

(iv) *Ambiguity in Chinese word tokenization* – the Chinese word contains one or more characters, with no 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 at least two 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.)

This problem can be addressed by retrieval based on overlapping character n-grams.

(v) *Speech recognition errors* – speech recognition output is imperfect. Errors may be caused by OOV words or acoustic confusions among in-vocabulary words (especially with respect to homophones). SDR based on syllables can improve robustness with respect to recognition errors in retrieval.

As can be seen, our multi-scale paradigm involves the use of variable-sized units. Query translation involves the translation of English *phrases* to reduce the translation ambiguity. Subsequent retrieval is based on the translated words. In addition, 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. 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 English-Chinese cross-language spoken document retrieval problem.

2. 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 furnished with word transcriptions produced by the Dragon automatic speech recognition system. 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. Table 1 describes the content in these collections.

	TDT-2 (Dev set)	**TDT-3 (Eval set)**
English news (New York Times or Associated Press)	17 topics, variable # of exemplars	56 topics, variable # of exemplars
Mandarin audio news (Voice of America)	2265 stories, 46.0 hours	3371 stories, 98.4 hours

Table 1. Statistics of TDT-2 and TDT-3: our development and evaluation data sets. (The Mandarin audio documents are accompanied by recognized words from the Dragon system).

3. THE MULTI-SCALE PARADIGM

This section describes our multi-scale paradigm in detail. It is divided into several sub-tasks – query formulation, audio indexing and retrieval. As described earlier, we make use of phrases,

[3] According to the GB-2312 character set.

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., 2000) indicated that bigrams are most effective (among the different n-grams) for retrieval.

3.1 Multi-scale Query Formulation

3.1.1 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 terms from the query exemplar.[4] First we excluded all stopwords, based on the English default stopword list used by the InQuery retrieval engine (Callan et al. 1992). 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 one thousand 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. The pure χ^2 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.

3.1.2 Query Translation

Named entities have been tagged by the BBN Identifinder (Bikel et al. 1997) system in our English query exemplars. Examples of named entities include "U.N. Security Council," and "partners of Goldman, Sachs and Co." Additional multi-word expressions (e.g. “human rights”, “guiding principles”, and “best interests”) are identified in our bilingual term list (BTL), which we formed by combining LDC’s English-Chinese bilingual term list with translated extracted from the CETA (Chinese-English Translation Assistance) dictionaries. Our BTL has nearly 200,000 total English terms corresponding to 400,000 translation pairs. The multi-word expressions (from Identifinder or our BTL) are treated as a “single term” in our 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).

3.1.3 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 produces the following translations:

科索沃 /ke-suo-wo/, 科索佛 /ke-suo-fo/,

科索夫 /ke-suo-fu/,科索伏 /ke-suo-fu/, or

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 represent 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 which are otherwise untranslatable and cannot be used for retrieval. Details of this technique are described in (Meng et al., 2001). We provide a succint description in the following.

Figure 2 presents an overview of the named entity transliteration process. We examine units in our query exemplar that are tagged by the BBN Identifnder system, and those absent from our translation dictionary (the BTL) are processed by our transliteration system. As shown in Figure 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, as we will see later. We attempt to look up the English pronunciation of the non-Chinese names.[5] Failing that, we generate the English pronunciation automatically from the spelling, by the application of letter-to-sound rules acquired by the transformation-based error-driving learning technique (Brill, 1994). For example, we can generate the English phoneme pronunciation /kk rr ih ss tt aa ff er/ from the spelling “Christopher” (see Figure 2).

[4] These may be multi-word units which are tagged or found in our term list, as will be explained later.

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

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 Figure 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.

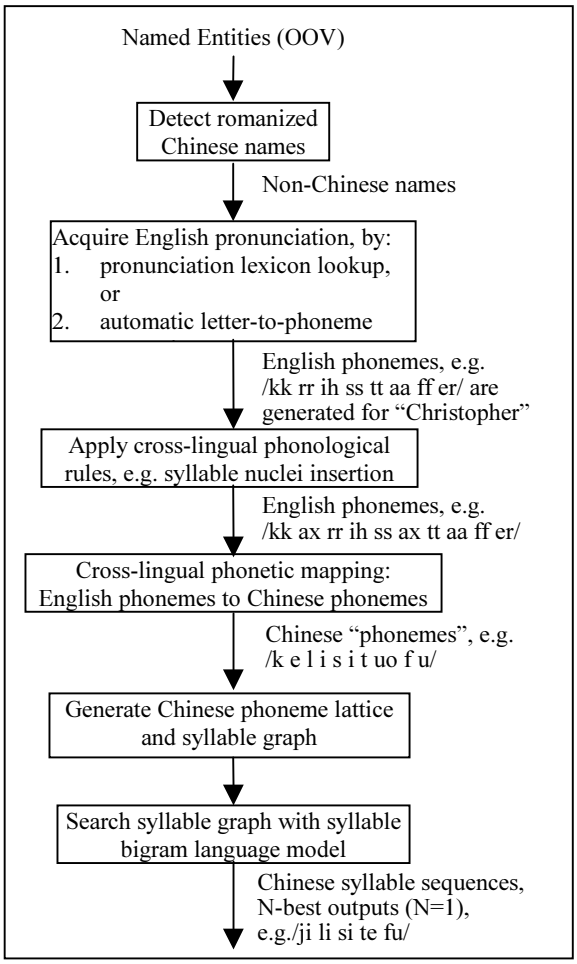

Figure 2. Overview of our subword translation process for handling untranslatable named entities in the query exemplars.

In order to obtain names 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.[6] These form the output of our names transliteration procedure, e.g. /ji li si te fu/ (see Figure 2). It is interesting to note that when we use a character bigram in place of a syllable bigram, our transliteration algorithm can produce subword translations in terms of character sequences, e.g. 基里斯特弗.

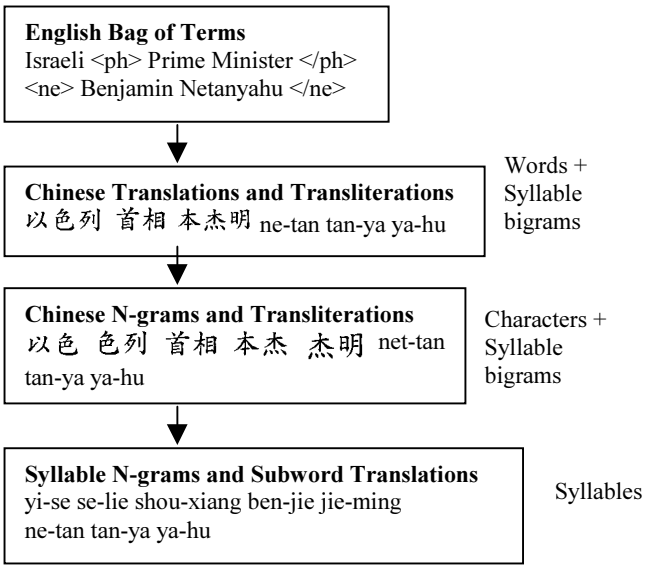

Figure 3. The process of multi-scale query construction in our system. The query representations at various stages of processing may be used. The representations seek to integrate information from phrase-based translation, word-based translation, subword translation and overlapping character/syllable n-grams which alleviates the problem of word tokenization ambiguities. Transformation from characters to syllables references a Chinese pronunciation lexicon.[7]

3.1.4 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 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 Figure 3.

[6] We set N=1 for simplicity.
[7] This is the LDC CALLHOME lexicon.

3.2 Multi-Scale Audio Indexing

The Dragon large-vocabulary continuous speech recognizer (Zhan et al., 1999) provided Chinese word transcriptions for our Mandarin audio collections (TDT2 and TDT3). 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 (~23 hours) by comparing the Dragon recognition hypotheses with the anchor scripts (treated as ground truth), 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.

We have also developed our own recognizer (the MEI recognizer) to provide a syllable hypothesis as an alternative to Dragon's. Both recognition outputs have been combined in retrieval, in an attempt to achieve robustness against speech recognition errors. However, using two speech recognizers instead of one did not bring obvious gains in CL-SDR performance. Possible reasons may be that Dragon's performance is quite good to begin with, and we have to further investigate methods to effectively combine multiple recognizer outputs for audio indexing. Details regarding to this investigation is reported in (Wang et al., 2001).

3.3 Multi-Scale Retrieval

We use InQuery as our retrieval engine, developed by the University of Massachusetts (Callan et al., 1992).[8] InQuery uses a probabilistic belief network as the main data structure behind its query language.

A key feature that we have employed is the "balanced query" mechanism (Leek et al., 2000) (Levow and Oard, 2000). Suppose that we had a query given by E_1, E_2, … E_n, where E_i represent the English query terms, and that E_1 has three possible Chinese translations, C_{11}, C_{12}, 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}, 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 #sum(#sum(C_{11}, C_{12}, C_{13})#sum(C_{21}, C_{22})…#sum(C_{n1}, C_{n2}, C_{n3})) in InQuery, with the outer #sum operators being the typical way of combining belief values across query terms in Inquery and the inner #sum operators implementing balanced translation. Balanced translation prevents query term that have 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 scores. 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 obtained by optimization experiments based on training data. This is termed *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.

[8] We used InQuery with a trivial modification to handle two-byte characters.

4. EXPERIMENTS

4.1 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. Equation (1) summarizes the process:

$$metric = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{M_i}\sum_{j=1}^{M_i}\left\{\frac{1}{N_i}\sum_{k=1}^{N_i}P_{ijk}\right\}\right\} \quad \ldots \quad (1)$$

where *metric* is the non-interpolated mean average precision, 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* P_{ijk} is the precision after the *kth* relevant document is retrieved for exemplar j of topic i.

In order to achieve statistical significance, we used up to twelve exemplars (i.e., $M_i = 12$) for each of the 17 topics whenever available.

4.2 Tuning with the Development Test Set

The TDT2 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 TDT2 audio collection was also used in training the MEI recognizer to optimize its recognition performance.

We found that in query term selection, it is beneficial to include all query terms (after stopword removal), and translate them. We also tuned the experimental configuration based on the number of translation alternatives to use, and results suggest that we include up to fifty translation alternatives, and combine them with a #sum operator for balanced queries. In applying our subword translation technique, we took the 200 most frequent names (tagged named entities) from the TDT2 collection and translated them at the subword level. This is used to augment the queries in both the TDT2 and TDT3 runs. Hence the development test set should have greater leverage based on subword translation.

4.3 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. We provide the key results in this section.

4.3.1 Phrase-based Translation

Our investigation of phrase-based translation took place in an early phase on our project. At the time, word-based translation gave a performance of mean average precision (mAP)=0.35. 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.

Query Translation Method	Retrieval Performance (mAP)	Relative Improve-ment
Word-by-Word Translation	0.350	--
Augmented with Phrase-based Translation	0.392	12% (statistically significant)

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

4.3.2 Multi-scale Retrieval

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

	Word-based Retrieval (mAP)	Character Bigrams (mAP)
TDT2 (dev test)	0.471	0.522
TDT3 (eval test)	0.462	0.477

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

The relative difference of 3.2% (w.r.t. TDT3) is also statistically significant, based on a paired two-tailed *t*-test with $p<0.05$. This suggests that the character bigrams may be effective in ameliorating the problem of word tokenization ambiguities. We also tried loosely coupling of the retrieval lists based on words and character bigrams, using weights optimized from TDT2, and tested on TDT3. This gave a performance of mAP=0.482 on TDT3, which is better than retrieval on each scale alone. Overlapping syllable bigrams performed below words. TDT2 and TDT3 results were at 0.468 and 0.422 respectively.

4.3.3 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.

	TDT2 Performance (mAP)	TDT3 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

Table 4. Investigation into the use of subword translation to salvage untranslatable named entities for CL-SDR. The procedure brought some performance gains.

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.

5. CONCLUSIONS

In this paper, we have described the Mandarin-English Information (MEI) project, 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. 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 of multi-scale 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, which, 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 any language pair.

6. ACKNOWLEDGMENTS

We acknowledge the contributions of all MEI team members. The MEI project is conducted during the Johns Hopkins University Summer Workshop 2000 (an NSF Workshop). www.clsp.jhu.edu/ws2000/groups/mei/welcome.html. 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. We thank the Linguistic Data Consortium for providing the TDT Corpora. We also thank Charles Wayne, George Doddington, James Allan, John Garafolo, 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.

7. REFERENCES

D. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a High-Performance Learning Name-finder," *Proceedings of the Fifth Conference on Applied Natural Language Processing,* pp. 194-201 (1997).

E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," *Computational Linguistics*, December 1995.

J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY Retrieval System," *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78-83, 1992.

K. Knight and J. Graehl, "Machine Transliteration," *Proceedings of the Conference of the Association for Computation Linguistics* (ACL), 1997.

K. L. Kwok and L. Grunfeld, "TREC-5 English and Chinese Experiments using PIRCS, " *Proceedings of the Fifth Annual Text Retrieval Conference (TREC-5),* 1996.

T. Leek, H. Jin, S. Sista and R. Schwartz, "The BBN Crosslingual Topic Detection and Tracking System," *Proceedings of the 1999 Topic Detection and Tracking Workshop*, 2000.

G. Levow and D. Oard, "Translingual Topic Tracking with PRISE," *Proceedings of the 1999 Topic Detection and Tracking Workshop*, 2000.

H. Meng, B. Chen, W. K. Lo and K. Tang, "Automatic Named Entity Transliteration for English-Chinese Cross-Language Spoken Document Retrieval," working paper, 2001.

H. Meng, W. K. Lo, Y. C. Li, and P. C. Ching, "Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval," *Proceedings of ICSLP2000,* Vol. IV, pp. 101-4, 2000.

D. Oard, G. Levow and C. Cabezas, "CLEF Experiemnts at Maryland: Statistical Stemming and Backoff Translation," *Lecture Notes in Computer Science*, forthcoming (2001).

H. Schuetze, D. Hull and J. O. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 229-237.

B. Stalls and K. Knight, "Translating Names and Technical Terms in Arabic Text," *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, 1998.

H. M. Wang, "Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese," *Speech Communication*, Vol. 32, pp. 49-60, 2000.

H. M. Wang, H. Meng, P. Schone, B. Chen and W. K. Lo, "Multi-scale Audio Indexing for Translingual Spoken Document Retrieval," *Proceedings of the IEEE Conference on Acoustics Speech and Signal Processing* (ICASSP), 2001.

P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 Broadcast News Transcription System for Mandarin," *Proceedings of the DARPA Broadcast News Workshop,* 1999.