美

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

Helen Meng,[1] Sanjeev Khudanpur,[2] Gina Levow,[3] Douglas W. Oard,[3] Hsin-Min Wang[4]
[1]The Chinese University of Hong Kong, [2]Johns Hopkins University,
[3]University of Maryland and [4]Academia Sinica (Taiwan)
{hmmeng@se.cuhk.edu.hk, sanjeev@clsp.jhu.edu, gina@umiacs.umd.edu,
oard@glue.umd.edu, whm@iis.sinica.edu.tw}

Abstract

We describe a system which supports English text queries searching for Mandarin Chinese spoken documents. This is one of the first attempts to tightly couple speech recognition with machine translation technologies for cross-media and cross-language retrieval. The Mandarin Chinese news audio are indexed with word and subword units by speech recognition. Translation of these multi-scale units can effect cross-language information retrieval. The integrated technologies will be evaluated based on the performance of translingual speech retrieval.

1. Introduction

Massive quantities of audio and multimedia programs are becoming available. For example, in mid-February 2000, www.real.com listed 1432 radio stations, 381 Internet-only broadcasters, and 86 television stations with Internet-accessible content, with 529 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 translingual speech retrieval if effective techniques can be developed. The Mandarin-English Information (MEI) project represents one of the first efforts in that direction.

MEI is one of the four projects selected for the Johns Hopkins University (JHU) Summer Workshop 2000.[1] Our research focus is on the integration of speech recognition and embedded translation technologies in the context of *translingual speech retrieval.* 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.

At the time of this writing, most of the MEI team members have been identified. This paper provides an update beyond our first proposal [Meng et al., 2000]. We present some ongoing work of our current team members, as well as our ideas on an evolving plan for the upcoming JHU Summer Workshop 2000. We believe the input from the research community will benefit us greatly in formulating our *final* plan.

2. Background

2.1 Translingual Information Retrieval

The earliest work on large-vocabulary cross-language information retrieval from free-text (i.e., without manual topic indexing) was reported in 1990 [Landauer and Littman, 1990], and the topic has received increasing attention over the last five years [Oard and Diekema, 1998]. Work on large-vocabulary retrieval from recorded speech is more recent, with some initial work reported in 1995 using subword indexing [Wechsler and Schauble, 1995], followed by the first TREC[2] Spoken Document Retrieval (SDR)

[1] http://www.clsp.jhu.edu/ws2000/

[2] Text REtrieval Conference, http://trec.nist.gov

evaluation [Garofolo et al., 2000]. The Topic Detection and Tracking (TDT) evaluations, which started in 1998, fall within our definition of speech retrieval for this purpose, differing from other evaluations principally in the nature of the criteria that human assessors use when assessing the relevance of a news story to an information need. In TDT, stories are assessed for relevance to an event, while in TREC stories are assessed for relevance to an explicitly stated information need that is often subject- rather than event-oriented.

The TDT-3[3] evaluation marked the first case of translingual speech retrieval – the task of finding information in a collection of recorded speech based on evidence of the information need that might be expressed (at least partially) in a different language. Translingual speech retrieval thus merges two lines of research that have developed separately until now. In the TDT-3 topic tracking evaluation, recognizer transcripts which have recognition errors were available, and it appears that every team made use of them. This provides a valuable point of reference for investigation of techniques that more tightly couple speech recognition with translingual retrieval. We plan to explore one way of doing this in the Mandarin-English Information (MEI) project.

2.2 The Chinese Language

In order to retrieve Mandarin audio documents, we should consider a number of linguistic characteristics of the Chinese language:

The Chinese language has many dialects. Different dialects are characterized by their differences in the phonetics, vocabularies and syntax. Mandarin, also known as Putonghua ("the common language"), is the most widely used dialect. Another major dialect is Cantonese, predominant in Hong Kong, Macau, South China and many overseas Chinese communities.

Chinese is a syllable-based language, where each syllable carries a lexical tone. Mandarin has about 400 base syllables and four lexical tones, plus a "light" tone for reduced syllables. There are about 1,200 distinct, tonal syllables for Mandarin. Certain syllable-tone combinations are non-existent in the language. The acoustic correlates of the lexical tone include the syllable's fundamental frequency (pitch contour) and duration. However, these acoustic features are also highly dependent on prosodic variations of spoken utterances.

The structure of Mandarin (base) syllables is (CG)V(X), where (CG) the syllable onset – C the initial consonant, G is the optional medial glide, V is the nuclear vowel, and X is the coda (which may be a glide, alveolar nasal or velar nasal). Syllable onsets and codas are optional. Generally C is known as the *syllable initial*, and the rest (GVX) *syllable final.*[4] Mandarin has approximately 21 initials and 39 finals.[5]

In its written form, Chinese is a sequence of characters. A word may contain one or more characters. Each character is pronounced as a tonal syllable. The character-syllable mapping is degenerate. On one hand, a given character may have multiple syllable pronunciations – for example, the character 行 may be pronounced as /hang2/,[6] /hang4/, or /xing2/. On the other hand, a given tonal syllable may correspond to multiple characters. Consider the two-syllable pronunciation /fu4 shu4/, which corresponds to a two-character word. Possible homophones include 富庶, (meaning "rich"), 負數, ("negative number"), 複數, ("complex number" or "plural"), 覆述 ("repeat").[7]

Aside from homographs and homophones, another source of ambiguity in the Chinese language is the definition of a Chinese word. The word has no delimiters, and the distinction between a word and a phrase is often vague. The lexical structure of the Chinese word is very different compared to English. Inflectional forms are minimal, while morphology and word derivations abide by a different set of rules. A word may inherit the syntax and semantics of (some of) its compositional characters, for

[3] http://morph.ldc.upenn.edu/Projects/TDT3/

[4] http://morph.ldc.upenn.edu/Projects/Chinese/intro.html

[5] The corresponding linguistic characteristics of Cantonese are very similar.

[6] These are Mandarin pinyin, the number encodes the tone of the syllable.

[7] Example drawn from [Leung, 1999].

example,[8] 紅 means *red* (a noun or an adjective), 色 means *color* (a noun), and 紅色 together means "the color red"(a noun) or simply "red" (an adjective). Alternatively, a word may take on totally different characteristics of its own, e.g. 東 means *east* (a noun or an adjective), 西 means *west* (a noun or an adjective), and 東西 together means *thing* (a noun). Yet another case is where the compositional characters of a word do not form independent lexical entries in isolation, e.g. 彷彿 means *fancy* (a verb), but its characters do not occur individually. Possible ways of deriving new words from characters are legion. The problem of identifying the words string in a character sequence is known as the *segmentation / tokenization* problem. Consider the syllable string:

/zhe4 yi1 wan3 hui4 ru2 chang2 ju3 xing2/

The corresponding character string has three possible segmentations – all are correct, but each involves a distinct set of words:

這一晚 會 如常 舉行

(Meaning: It will be 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 this evening banquet takes place frequently...)

The above considerations lead to a number of techniques we plan to use for our task. We concentrate on three equally critical problems related to our theme of translingual speech retrieval: (i) indexing Mandarin Chinese audio with word and subword units, (ii) translating variable-size units for cross-language information retrieval, and (iii) devising effective retrieval strategies for English text queries and Mandarin Chinese news audio.

3. Multiscale Audio Indexing

A popular approach to spoken document retrieval is to apply Large-Vocabulary Continuous Speech Recognition (LVCSR)[9] for audio indexing, followed by text retrieval techniques. Mandarin Chinese presents a challenge for word-level indexing by LVCSR, because of the ambiguity in tokenizing a sentence into words (as mentioned earlier). Furthermore, LVCSR with a static vocabulary is hampered by the out-of-vocabulary (OOV) problem, especially when searching sources with topical coverage as diverse as that found in broadcast news.

By virtue of the monosyllabic nature of the Chinese language and its dialects, the syllable inventory can provide a *complete phonological coverage* for spoken documents, and circumvent the OOV problem in news audio indexing, offering the potential for greater recall in subsequent retrieval. The approach thus supports searches for previously unknown query terms in the indexed audio.

The pros and cons of subword indexing for an English spoken document retrieval task was studied in [Ng, 2000]. Ng pointed out that the exclusion of lexical knowledge when subword indexing is performed in isolation may adversely impact discrimination power for retrieval, but that some of that impact can be mitigated by modeling sequential constraints among subword units. We plan to investigate the efficacy of using *both* word and subword units for Mandarin audio indexing [Meng et al., 2000]. Although Ng found that such an approach produced little gain over words alone for English, the structure of Mandarin Chinese may produce more useful subword features.

3.1 Modeling Syllable Sequence Constraints

We have thus far used overlapping syllable *N*-grams for spoken document retrieval for two Chinese dialects – Mandarin and Cantonese. Results on a known-item retrieval task with over 1,800 error-free news transcripts [Meng et al., 1999] indicate that constraints from overlapping bigrams can yield significant improvements in retrieval performance over syllable unigrams, producing retrieval performance competitive

[8] Examples drawn from [Meng and Ip, 1999].

[9] The lexicon size of a typical large-vocabulary continuous speech recognizer can range from 10,000 to 100,000 word forms.

with that obtained using automatically tokenized Chinese words.

The study in [Chen, Wang and Lee, 2000] also used syllable pairs with skipped syllables in between. This is because many Chinese abbreviations are derived from skipping characters, e.g. 國家科學委員會 National Science Council" can be abbreviated as 國科會 (including only the first, third and the last characters). Moreover, synonyms often differ by one or two characters, e.g. both 中華文化 and 中國文化 mean "Chinese culture". Inclusion of these "skipped syllable pairs" also contributed to retrieval performance.

When modeling sequential syllable constraints, lexical constraints on recognized words may be helpful. We thus plan to explore the potential for integrated sequential modelling of *both* words and syllables [Meng et al., 2000].

4. Multiscale Embedded Translation

Figures 1 and 2 illustrate two translingual retrieval strategies. In query translation, English text queries are transformed into Mandarin and then used to retrieve Mandarin documents. For document translation, Mandarin documents are translated into English before they are indexed and then matched with English queries. McCarley has reported improved effectiveness from techniques that couple the two techniques [McCarley, 1999], but time constraints may limit us to exploring only the query translation strategy during the six-week Workshop.

4.1 Word Translation

While we make use of sub-word transcription to smooth out-of-vocabulary(OOV) problems in speech recognition as described above, and to alleviate the OOV problem for translation as we discuss in the next section, accurate translation generally relies on the additional information available at the word and phrase levels. Since the "bag of words" information retrieval techniques do not incorporate any meaningful degree of language understanding to assess similarity between queries and documents, a word-for-word (or, more generally, term-for-term) embedded translation approach can achieve a useful level of effectiveness for many translingual retrieval applications [Oard and Diekema, 1998].

We have developed such a technique for the TDT-3 topic tracking evaluation [Levow and Oard, 2000]. For that work we extracted an enriched bilingual Mandarin-English term list by combining two term lists: (i) A list assembled by the Linguistic Data Consortium from freely available on-line resources; and (ii) entries from the CETA file (sometimes referred to as "Optilex"). This is a Chinese to English translation resource that was manually compiled by a team of linguists from more than 250 text sources, including special and general-purpose print dictionaries, and other text sources such as newspapers. The CETA file contains over 250,000 entries, but for our lexical work we extracted a subset of those entries drawn from contemporary general-purpose sources. We also excluded definitions such as "particle indicating a yes/no question." Our resulting Chinese to English merged bilingual term list contains translations for almost 200,000 Chinese terms, with average of almost two translation alternatives per term. We have also used the same resources to construct an initial English to Chinese bilingual term list that we plan to refine before the Workshop.

Three significant challenges faced by term-to-term translation systems are term selection in the source language, the source language coverage of the bilingual term list, and translation selection in the target language when more than one alternative translation is known. Word segmentation is a natural by-product of large vocabulary Mandarin speech recognition, and white space provides word boundaries for the English queries. We thus plan to choose words as our basic term set, perhaps augmenting this with the multiword expressions found in the bilingual term list.

Achieving adequate source language coverage is challenging in news retrieval applications of the type modelled by TDT, because proper names and technical terms that may not be present in general-purpose lexical resources often provide important retrieval cues. Parallel (translation equivalent) corpora have proven to be a useful source of translation

equivalent terms, but obtaining appropriate domain-specific parallel corpora in electronic form may not be practical in some applications. We therefore plan to investigate the use of comparable corpora to learn translation equivalents, based on techniques in [Fung, 1998]. Subword translation, described below, provides a complementary way of handling terms for which translation equivalents cannot be reliably extracted from the available comparable corpora.

One way of dealing with multiple translations is to weight the alternative translations using either a statistical translation model trained on parallel or comparable corpora to estimate translation probability conditioned on the source language term. When such resources are not sufficiently informative, it is generally possible to back off to an unconditioned preference statistic based on usage frequency of each possible translation in a representative monolingual corpus in the target language. In retrospective retrieval applications the collection being searched can be used for this purpose. We have applied simple versions of this approach with good results [Levow and Oard, 2000].

We have recently observed that a simpler technique introduced by [Pirkola, 1998] can produce excellent results. The key idea is to use the structure of the lexicon, in which several target language terms can represent a single source language term, to induce structure in the translated query that the retrieval system can automatically exploit. In essence, the translated query becomes a bag of bags of terms, where each smaller bag corresponds to the set of possible translations for one source-language term. We plan to implement this structured query translation approach using the Inquery [Callan, 1992] "synonym" operator in the same manner as [Pirkola, 1998], and to the potential to extend the technique to accommodate alternative recognition hypothesis and subword units as well.

4.2 Subword Translation

Since Mandarin spoken documents can be indexed with both words and subwords, the translation (or "phonetic transliteration") of subword units is of particular interest. We plan to make use of *cross-language phonetic mappings* derived from English and Mandarin pronunciation rules for this purpose. This should be especially useful for handling named entities in the queries, e.g. names of people, places and organizations, etc. which are generally important for retrieval, but may not be easily translated. Chinese translations of English proper nouns may involve semantic as well as phonetic mappings. For example, "Northern Ireland" is translated as 北愛爾蘭 — where the first character 北 means 'north', and the remaining characters 愛爾蘭 are pronounced as /ai4-er3-lan2/. Hence the translation is both *semantic* and *phonetic*. When Chinese translations strive to attain phonetic similarity, the mapping may be inconsistent. For example, consider the translation of "Kosovo" – sampling Chinese newspapers in China, Taiwan and Hong Kong produces the following translations:

科索沃 /ke1-suo3-wo4/, 科索佛 /ke1-suo3-fo2/, 科索夫 /ke1-suo3-fu1/,科索伏 /ke1-suo3-fu2/, or 柯索佛 /ke1-suo3-fo2/.

As can be seen, there is no systematic mapping to the Chinese character sequences, but the translated Chinese pronunciations bear some resemblance to the English pronunciation (/k ow s ax v ow/). In order to support retrieval under these circumstances, the approach should involve approximate matches between the English pronunciation and the Chinese pronunciation. The matching algorithm should also accommodate phonological variations. Pronunciation dictionaries, or pronunciation generation tools for both English words and Chinese words / characters will be useful for the matching algorithm. We can probably leverage off of ideas in the development of universal speech recognizers [Cohen et al., 1997].

5. Multiscale Retrieval

5.1 Coupling Words and Subwords

We intend to use both words and subwords for retrieval. *Loose coupling* would involve separate retrieval runs using words and subwords, producing two ranked lists, followed by list merging using techniques such as those explored by [Voorhees, 1995]. *Tight coupling*, by

contrast, would require creation of a unified index containing both word and subword units, resulting in a single ranked list. We hope to explore both techniques during the Workshop.

5.2 Imperfect Indexing and Translation

It should be noted that speech recognition exacerbates uncertainty when indexing audio, and that translation or transliteration exacerbates uncertainty when translating queries and/or documents. To achieve robustness for retrieval, we have tried three techniques that we have found useful: (i) Syllable lattices were used in [Wang, 1999] and [Chien et al., 2000] for monolingual Chinese retrieval experiments. The lattices were pruned to constrain the search space, but were able to achieve robust retrieval based on imperfect recognized transcripts. (ii) Query expansion, in which syllable transcription were expanded to include possibly confusable syllable sequences based on a syllable confusion matrix derived from recognition errors, was used in [Meng et al., 1999]. (iii) We have expanded the document representation using terms extracted from similar documents in a comparable collection [Levow and Oard, 2000], and similar techniques are known to work well in the case of query translation (Ballesteros and Croft, 1997). We hope to add to this set of techniques by exploring the potential for query expansion based on cross-language phonetic mapping.

6. Using the TDT-3 Collection

We plan to use the TDT-2 collection for development testing and the TDT-3 collection for evaluation. Both collections provide documents from two English newswire sources, six English broadcast news audio sources, two Mandarin Chinese newswire sources, and one Mandarin broadcast news source (Voice of America). Manually established story boundaries are available for all audio collections, and we plan to exploit that information to simplify our experiment design. The TDT-2 collection includes complete relevance assessments for 20 topics, and the TDT-3 collection provides the same for 60 additional topics, 56 of which have at least one relevant audio story. For each topic, at least four English stories and four Chinese stories are known.

We plan to automatically derive text queries based on one or more English stories that are presented as exemplars, and to use those queries to search the Mandarin audio collection. Manually constructed queries will provide a contrastive condition. Unlike the TDT "topic tracking" task in which stories must be declared relevant or not relevant in the order of their arrival, we plan to perform retrospective retrieval experiments in which all documents are known when the query is issued. By relaxing the temporal ordering of the TDT topic tracking task, we can meaningfully search for Mandarin Chinese stories that may have arrived before the exemplar story or stories. We thus plan to report ranked retrieval measures of effectiveness such as average precision in addition to the detection statistics (miss and false alarm) typically reported in TDT.

7. Summary

This paper presents our current ideas and evolving plan for the MEI project, to take place at the Johns Hopkins University Summer Workshop 2000. Translingual speech retrieval is a long-term research direction, and our team looks forward to jointly taking an initial step to tackle the problem. The authors welcome all comments and suggestions, as we strive to better define the problem in preparation for the six-week Workshop.

Acknowledgments

The authors wish to thank Patrick Schone, Erika Grams, Fred Jelinek, Charles Wayne, Kenney Ng, John Garofolo, and the participants in the December 1999 WS2000 planning meeting and the TDT-3 workshop for their many helpful suggestions. The Hopkins Summer Workshop series is supported by grants from the National Science Foundation. Our results reported in this paper reference thesis work in progress of Wai-Kit Lo (Ph.D. candidate, The Chinese Unversity of Hong Kong) and Berlin Chen (Ph.D. candidate, National Taiwan University).

References

Ballesteros and W. B. Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," Proceedings of ACM SIGIR, 1997.

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

Carbonnell, J., Y. Yang, R. Frederking and R.D. Brown, "Translingual Information Retrieval: A Comparative Evaluation," Proceedings of IJCAI, 1997.

Chen, B., H.M. Wang, and L.S. Lee, "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan using Syllable-Level Statistical Characteristics," Proceedings of ICASSP, 2000.

Chien, L. F., H. M. Wang, B. R. Bai, and S. C. Lin, "A Spoken-Access Approach for Chinese Text and Speech Information Retrieval," Journal of the American Society for Information Science, 51(4), pp. 313-323, 2000.

Choy, C. Y., "Acoustic Units for Mandarin Chinese Speech Recognition," M.Phil. Thesis, The Chinese University of Hong Kong, Hong Kong SAR, China, 1999.

Cohen, P., S. Dharanipragada, J. Gros, M. Mondowski, C. Neti, S. Roukos and T. Ward, "Towards a Universal Speech Recognizer for Multiple Languages," Proceedings of ASRU, 1997.

Fung, P., "A Statistical View on Bilingual Lexicon Extraction: From parallel corpora to non-parallel corpora," Proceedings of AMTA, 1998.

Garofolo, J.S., Auzanne, G.P., Voorhees, E.M., "The TREC Spoken Document Retrieval Track: A Success Story," Proceedings of the Recherche d'Informations Assistée par Ordinateur: Content-Based Multimedia Information Access Conference , April 12-14, 2000,to be published.

Knight, K. and J. Graehl, "Machine Transliteration," Proceedings of ACL, 1997.

Landauer, T. K. and M.L. Littman, "Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing," Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary, 1990.

Leung, R., "Lexical Access for Large Vocabulary Chinese Speech Recognition," M. Phil. Thesis, The Chinese University of Hong Kong, Hong Kong SAR, China 1999.

Levow, G. and D.W. Oard, "Translingual Topic Tracking with PRISE," Working notes of the DARPA TDT-3 Workshop, 2000.

Lin, C. H., L. S. Lee, and P. Y. Ting, "A New Framework for Recognition of Mandarin Syllables with Tones using Sub-Syllabic Units," Proceedings of ICASSP, 1993.

Liu, F. H., M. Picheny, P. Srinivasa, M. Monkowski and J. Chen, "Speech Recognition on Mandarin Call Home: A Large-Vocabulary, Conversational, and Telephone Speech Corpus," Proceedings of ICASSP, 1996.

McCarley, S., "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval," Proceedings of ACL, 1999.

Meng, H. and C. W. Ip, "An Analytical Study of Transformational Tagging of Chinese Text," Proceedings of the Research On Computational Lingustics (ROCLING) Conference, 1999.

Meng, H., W. K. Lo, Y. C. Li and P. C. Ching, "A Study on the Use of Syllables for Chinese Spoken Document Retrieval," Technical Report SEEM1999-11, The Chinese University of Hong Kong, 1999.

Meng, H., Khudanpur, S., Oard, D. W. and Wang, H. M., "Mandarin-English Information (MEI)," Working notes of the DARPA TDT-3 Workshop, 2000.

Ng, K., "Subword-based Approaches for Spoken Document Retrieval," Ph.D. Thesis, MIT, February 2000.

Oard, D. W. and A.R. Diekema, "Cross-Language Information Retrieval," Annual Review of Information Science and Technology, vol.33, 1998.

Pirkola, A., "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval," Proceedings of ACM SIGIR, 1998.

Sheridan P. and J. P. Ballerini, "Experiments in Multilingual Information Retrieval using the

SPIDER System," Proceedings of ACM SIGIR, 1996.

Voorhees, E., "Learning Collection Fusion Strategies," Proceedings of SIGIR, 1995.

Wang, H. M., "Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching," Proceedings of the Fourth International Workshop on Information Retrieval in Asian Languages, 1999.

Wechsler, M. and P. Schaüble, "Speech Retrieval Based on Automatic Indexing," Proceedings of MIRO-1995.

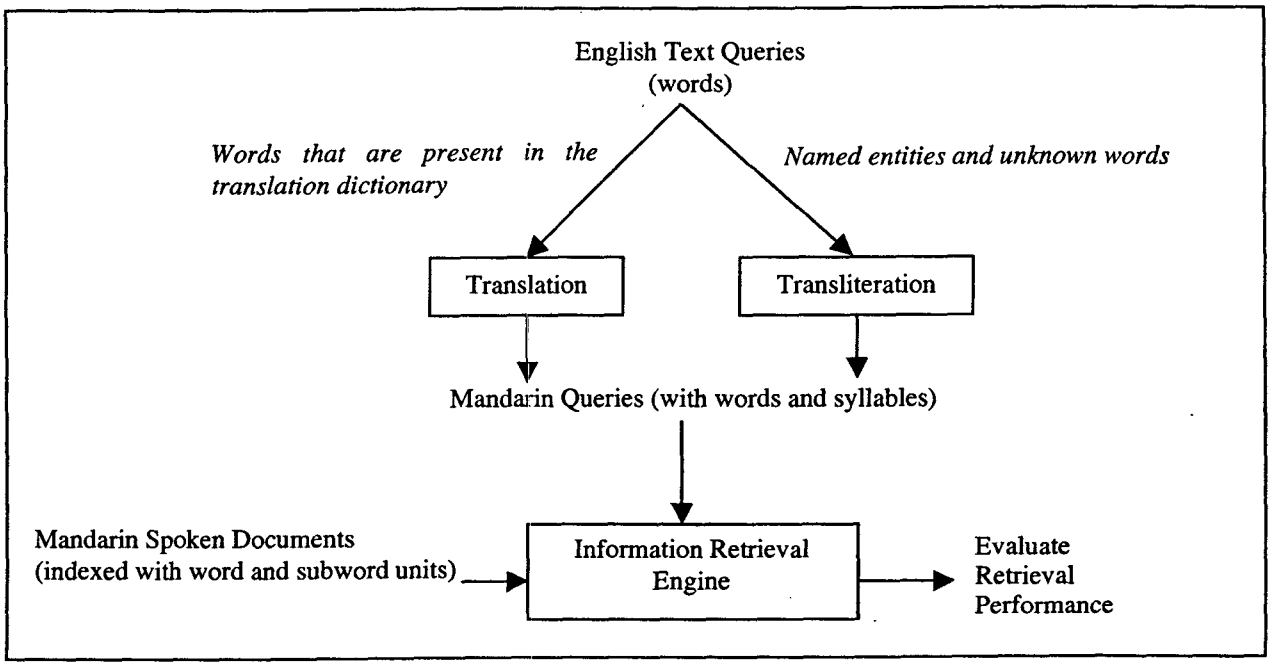

Figure 1. Query translation strategy.

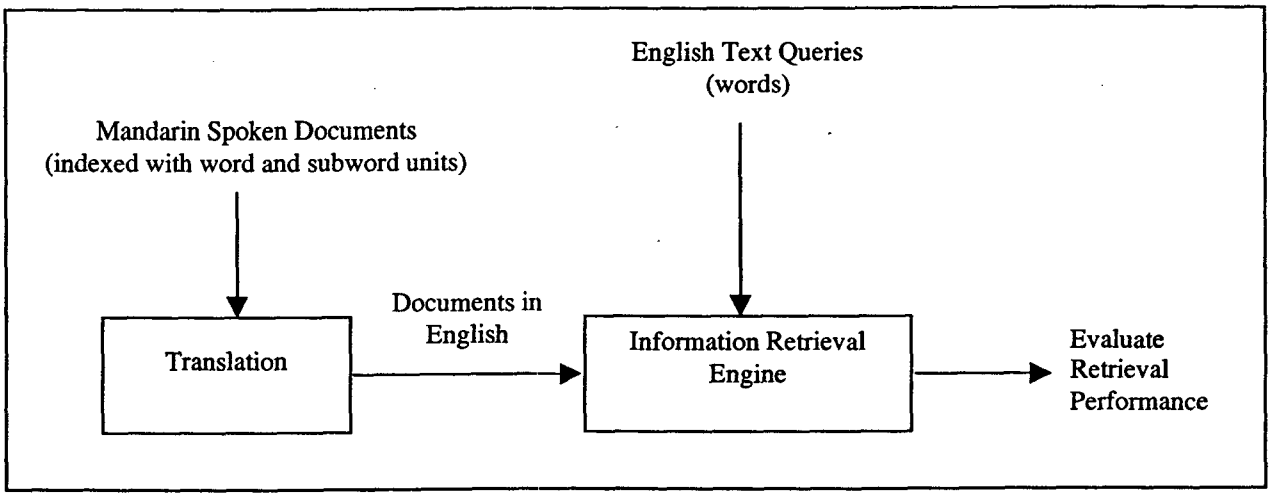

Figure 2. Document translation strategy.