美

MANDARIN-ENGLISH INFORMATION (MEI)

Helen Meng,[1] *Sanjeev Khudanpur,*[2] *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,
oard@glue.umd.edu, whm@iis.sinica.edu.tw}

ABSTRACT

Mandarin-English Information (MEI) is one of the four projects selected for the Johns Hopkins University Summer Workshop 2000. We plan to develop technologies for using written queries to search spoken documents (cross-media) between English and Mandarin Chinese (cross-language). Our research focus is on the integration of speech recognition and machine translation technologies in the context of translingual speech retrieval. We plan to work on the problems of: (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.

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. Our research focus is on the integration of speech recognition and machine 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.

It was suggested[1] that the TDT-3 corpora are valuable resources for our work. In addition to the presence of Mandarin news audio, the availability of topic descriptions, relevance judgements, baseline recognizer transcripts, and machine translation results are well suited to support our investigation.

At the time of this writing, more than half of the MEI team members have been identified. Since the TDT-3 Workshop precedes the first meeting of the MEI team, this paper describes some ongoing work of our current team members, as well as our ideas and *preliminary* plan for the upcoming JHU Summer Workshop 2000. We believe the input from the TDT community will benefit us greatly in formulating our *final* plan.

2 BACKGROUND

2.1 Previous Developments in 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 Spoken Document Retrieval (SDR) evaluation [Garofolo et al., 1997]. 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.

The TDT-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, baseline recognizer transcripts 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 at the Johns Hopkins Workshop this summer.

2.2 The Chinese Language

In order to tackle the problem of indexing Mandarin audio, we should consider the linguistic characteristics of the Chinese language when developing our technologies.

[1] Idea attributed to Charles Wayne and George Doddington at the JHU Worshop 2000 Plannng Meeting, December 1999.

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*.[2] Mandarin has approximately 21 initials and 39 finals.[3]

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/,[4] /hang4/, /heng2/ 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 which can be found in LDC's CALLHOME Mandarin Lexicon include 富庶, (meaning "rich"), 負數, ("negative number"), 復數, ("complex number" or "plural"), 覆述 ("repeat").[5]

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 example,[6] 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 questions related to indexing (by speech recognition) and retrieving Chinese spoken documents, such as:

Which unit should we use to index and retrieve Chinese spoken audio with high efficacy ?

To what extent is the lexical tone important for the task ?[7]

How should we maintain retrieval robustness, in lieu ambiguities due to homophones, homographs and tokenization?

How should we maintain retrieval robustness, in lieu of recognition errors which affect indexing ?

What retrieval strategies can best effect cross-language and cross media retrieval?

3 THE MEI PROJECT

The Johns Hopkins Summer Workshop will run for six weeks in July and August of this year. The central theme of the MEI project is *translingual speech retrieval*, and we seek to investigate the integration of speech recognition and machine translation technologies for this purpose. We concentrate on three equally critical problems related to our theme: (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.

3.1 Unit Selection for Indexing Mandarin News Audio

A popular approach for spoken document retrieval is to apply large-vocabulary continuous speech recognition (LVCSR) for audio indexing, followed by text retrieval techniques.

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

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

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

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

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

[7] There are 400 base syllables and 1200 tonal syllables available for indexing.

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.

This advantage was pointed out by Ng in [Ng, 2000], which is a thorough study on subword indexing based on the TREC-8 spoken document retrieval evaluation. The subword approach is also more generalizable to other languages in translingual speech retrieval. However, Ng also cautioned that the subword inventory may lose discrimination power between relevant and irrelevant documents when compared to word indexing, due to the exclusion of lexical knowledge. It is important to mitigate the loss by modeling the sequential constraints of subword units. Monolingual English experiments were conducted, and he demonstrated that the overlapping phoneme trigrams were the best subword unit to index. The resultant retrieval performance is comparable to that of word-based indexing when error-free recognition is simulated.

We plan to investigate the efficacy of syllables as subword units for Mandarin audio indexing. First, the syllables need to be recognized accurately from the audio. Second, we need to model the syllable sequential constraints effectively for audio indexing. Third, we will investigate the use of a “syllable/word hybrid”.

3.1.1 Syllable Recognition

Recognition performance is critically dependent on acoustic modeling, especially since broadcast news recordings present challenging acoustic conditions and speaking styles. Acoustic models for syllable recognition may be based on syllable models, or sub-syllable models, e.g. initials / finals (finals with or without tone) [Lin, Lee and Ting, 1993], premes / core-finals (without tone), premes / tonemes (with tones) [Liu et al., 1996], or phones. Figure 1 shows a spectrogram (frequency-time plot) of sub-syllable units within the syllable /jiang/.

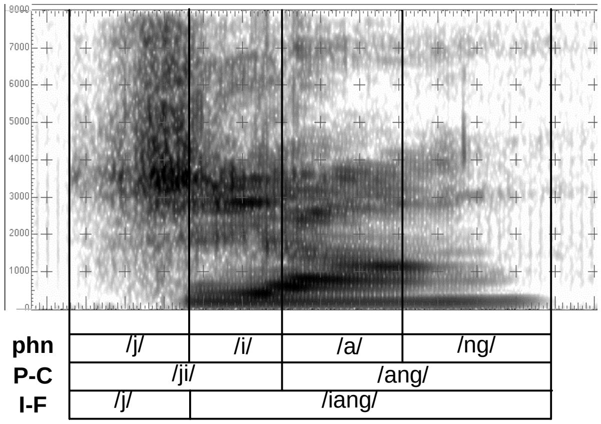

Figure 1. Spectrogram (frequency-time plot) illustrating the sub-syllable structures within the syllable /jiang/. phn: phones, P-C: premes/core-finals, I-F: initials/finals.

A previous study [Choy, 1999] compared the use of different sub-syllabic units for acoustic modeling, and showed that premes / tonemes gave the highest syllable recognition performance on the LDC Mandarin CALLHOME corpus. Use of tone information in premes / tonemes provided a slight advantage over premes / core-finals. We plan to run a similar comparison on the TDT-3 Mandarin corpus, in an attempt to optimize performance on audio transcription based on syllable recognition.[8]

In addition, a high performance Mandarin Chinese speech recognizer has been developed by Wang et al., which uses initial/final models for syllable recognition. The recognizer has been used in monolingual Chinese retrieval experiments [Wang, 2000], [Chien et al., 2000], and can produce filtered syllable lattices to achieve robust retrieval based on imperfect recognized transcripts.

3.1.2 Modeling Constraints in Syllable Sequences for Retrieval

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] suggest that constraints from overlapping bigrams brought about significant improvements in retrieval performance over syllable unigrams, and the retrieval performance is competitive with that of automatically tokenized Chinese words.

The study in [Chen, Wang and Lee, 2000] also used syllable pairs with *M* 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.

3.1.3 The Syllable/Word Hybrid

In modeling syllable sequential constraints, it is conceivable that the lexical constraints of the in-vocabulary words should be most important. Out-of-vocabulary words can still be indexed by overlapping syllables. Our experiments based on 1,800 news stories and a vocabulary size of 47,000 words showed that this strategy can reduce the number of index terms by factor of seven, suffering a slight decline in retrieval effectiveness [Meng et al., 1999]. The use of *both* word and subword units for spoken document retrieval is one aspect of *information fusion*, described in [Ng, 2000]. The results suggested that the word/subword combination might outperform both the "word only" and "subword only" approaches. Some results from TREC-6 [Wilkinson, 1997] on Chinese document retrieval also suggested potential advantages from using word-character combinations (a character corresponds to a syllable).

[8] Robustness of speech recognition in different acoustic conditions is an important problem, and may be covered by a another team at the JHU Workshop.

We plan to explore the potential advantages of using *both* words and syllables for indexing Chinese broadcast news.

3.2 Mandarin/English Translingual Retrieval

In TDT-3, baseline recognizer transcripts were provided for the Mandarin Chinese audio. Use of those transcripts was required in the topic tracking evaluation, and we are not aware of contrastive runs using the Mandarin Chinese audio in any other way. Baseline translations of the transcribed Mandarin Chinese into English were also provided, but their use was not required. Some teams explored the potential for more closely coupling translation with retrieval, with contrastive runs showing better retrieval effectiveness from the closely coupled techniques than with baseline translations. We plan to investigate whether similar gains can be achieved by more closely coupling translation and retrieval technologies.

English queries are easily translated into Mandarin, and Mandarin words from speech recognition are easily translated into English using techniques that we developed for TDT-3 {Levow and Oard, 2000. We have developed translation lexicons for both directions, merged bilingual term lists provided by the Linguistic Data Consortium with similar lists extracted from an electronic version of the Chinese-English Translation Assistance (CETA) dictionary.

Query translation will require that we perform retrieval in the Chinese subword unit space, but the most sophisticated available retrieval systems are designed to work with single-byte characters. We plan to accommodate this by converting both Chinese words and subword units to space-delimited ASCII representations. We plan to use the Inquery text retrieval system because the Inquery synonym operator provides a natural way of accommodating translation ambiguity [Pirkola, 1998]. The syllable lattice contains more information than the top-scoring syllable recognition hypothesis, and [Chen, Wang and Lee, 2000] reports a *significant* improvement in the retrieval performance when it is used as a basis for retrieval. We plan to apply Pirkola's method to explore the potential of $n$-best syllable sequence recognition. We are also experimenting with a query expansion strategy, in which the syllable transcription of the textual query is expanded to include possibly confusable syllable sequences based on a syllable confusion matrix derived from recognition errors [Meng et al., 1999], and we may incorporate that technique.

Dictionary-based query translation suffers from limited coverage of topic-specific terminology, particularly proper names. Three techniques have been suggested for overcoming this limitation: cross-language phonetic mapping [Knight and Graehl, 1997], identification of translation-equivalent terms in parallel corpora [Carbonnell, 1997], and identification of terms with similar usage in comparable corpora [Sheridan and Ballerini, 1996]. We plan to explore this space, seeking methods to translate words *and* subword units, and then to integrate those methods with techniques that *jointly* use word and subword translations as a basis for retrieval.

3.2.1 Cross-Language Phonetic Mapping

Cross-language phonetic mapping is of particular interest to us because it amounts to the translation of subword units. Newswire text is populated with proper nouns (names of people, places, organizations, etc.) that are generally important for retrieval but may not be present in bilingual dictionaries. Chinese translations of English proper nouns may involve semantic as well as phonemic 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 pronunced as /ai4-er3-lan2/" When Chinese translations strive to attain phonemic 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-sou3-wo4/, 科索佛 /ki1-sou3-fo2/, 科索夫 /ke1-sou3-fu1/, 科索伏 /ke1-sou3-fu2/, or 柯索佛 /ke1-sou3-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 ao r 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].

4 USE OF THE TDT-3 COLLECTION

We plan to base our evaluation on the 121 hours of Mandarin Chinese audio materials in the TDT-3 evaluation collection. In order to limit the complexity of our experiments, we plan to run the story-boundary-known condition, in which 4,624 Mandarin Chinese stories are known. Two sets of 60 English queries will be formed from the TDT-3 evaluation topic descriptions. Short queries will be formed manually using 1-4 words in a manner similar to that which might be used for queries posed to a Web search engine. Long queries will be formed automatically, and will consist of approximately 50 words that are observed to be highly selective (by the inverse document frequency measure) in the English newswire portion of the TDT-3 training collection. If time allows, we may also explore the use of entire documents or even sets of documents as a basis for forming queries. To support our development effort, we plan to use the approximately 56 hours (2,934 stories) in the TDT-3 development test collection and to form the two sets of 20 English queries for that collection as described above.

5 SUMMARY

This paper presents our current ideas and preliminary plan for the MEI project, to take place at the JHU 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 Fred Jelinek, Charles Wayne, and the other participants at the December 1999 Summer Workshop planning meeting for their many helpful suggestions. The Hopkins Summer Workshop 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

1. Carbonnell, J., Y. Yang, R. Frederking and R.D. Brown, "Translingual Information Retrieval: A Comparative Evaluation," Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence, 1997.
2. 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.
3. 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.
4. Choy, C. Y., “Acoustic Units for Mandarin Chinese Speech Recognition,” M.Phil. Thesis, The Chinese University of Hong Kong, Hong Kong SAR, China, 1999.
5. 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.
6. Garofolo, J., E. Voorhees, V. Stanford and K. Sparck Jones, “TREC-6 1997 Spoken Document Retrieval Track Overview and Results,” Proceedings of TREC-6, 1997.
7. Knight, K. and J. Graehl, “Machine Transliteration,” Proceedings of the 7th International Conference of the Association for Computational Linguistics, 1997.
8. 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 and Text Research, pp31-38, 1990.
9. Leung, R., "Lexical Access for Large Vocabulary Chinese Speech Recognition,” M. Phil. Thesis, The Chinese University of Hong Kong, Hong Kong SAR, China 1999.
10. Levow, G. and D.W. Oard, “Translingual Topic Tracking with PRISE,” Working Notes of the Third Topic Detection and Tracking Workshop, 2000.
11. 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.
12. 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.
13. 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.
14. 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.
15. Ng, K., "Subword-based Approaches for Spoken Document Retrieval," Ph.D. Thesis, Massachusetts Institute of Technology, February 2000.
16. Oard, D. W. and A.R. Diekema, “Cross-Language Information Retrieval,” Annual Review of Information Science and Technology, vol.33, 1998.
17. Sheridan P. and J. P. Ballerini, "Experiments in Multilingual Information Retrieval using the SPIDER System," Proceedings of ACM SIGIR-96, 1996.
18. 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.
19. Wechsler, M. and P. Schauble, “Speech Retrieval Based on Automatic Indexing,” Proceedings of MIRO-1995.
20. Wilkinson, R., “Chinese Document Retrieval at TREC-6,” Proceedings of the TREC-6, 1997.
21. Pirkola, A., “The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval,” Proceedings of ACM SIGIR98, 1998.