MULTI-SCALE AUDIO INDEXING FOR TRANSLINGUAL SPOKEN DOCUMENT RETRIEVAL

Hsin-min Wang[1], Helen Meng[2], Patrick Schone[3], Berlin Chen[1], and Wai-Kit Lo[4]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2] Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong
[3] Dept. of Defense, Ft. Meade, MD 20755, USA
[4] Dept. of Electronic Engineering, The Chinese University of Hong Kong

ABSTRACT

MEI (Mandarin-English Information) is an English-Chinese crosslingual spoken document retrieval (CL-SDR) system developed during the Johns Hopkins University Summer Workshop 2000. We integrate speech recognition, machine translation, and information retrieval technologies to perform CL-SDR. MEI advocates a *multi-scale paradigm*, where *both* Chinese words and subwords (characters and syllables) are used in retrieval. The use of subword units can complement the word unit in handling the problems of Chinese word tokenization ambiguity, Chinese homophone ambiguity, and out-of-vocabulary words in audio indexing. This paper focuses on multi-scale audio indexing in MEI. Experiments are based on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3), where we indexed Voice of America Mandarin news broadcasts by speech recognition on both the word and subword scales. In this paper, we discuss the development of the MEI syllable recognizer, the representations of spoken documents using overlapping subword n-grams and lattice structures. Results show that augmenting words with subwords is beneficial to CL-SDR performance.

1. INTRODUCTION

The global information infrastructure is continually enriched with multilingual and multimedia content (text, graphics, audio, video). Hence there is increasing demand for *cross-lingual* and *cross-media* information retrieval technologies, to enable the user search for personally-relevant information efficiently. MEI (Mandarin-English Information) is an English-Chinese crosslingual spoken document retrieval (CL-SDR) system developed during the Johns Hopkins University Summer Workshop 2000 [1-3]. We integrate speech recognition, machine translation, and information retrieval technologies to perform CL-SDR. English and Chinese are two of the predominant languages used by the global population, and their differences present novel research challenges for our CL-SDR task.

MEI advocates a *multi-scale paradigm* for English-Chinese CL-SDR, where both words and subwords are used for the task, as illustrated in Figure 1. This paper focuses on *multi-scale audio indexing* of Mandarin Chinese spoken documents. The document collection is indexed on the word scale, as well as the subword scale (with Chinese characters and syllables) via speech recognition. Multi-scale audio indexing provides the following advantages:

(i) Lexical knowledge – this is important for retrieval precision and is provided by the Chinese words.

(ii) Robustness against Chinese word tokenization ambiguity – written Chinese consists of character sequences with no word delimiters. A given character sequence may be tokenized in multiple ways, and the different word sequences may give rise to different meanings. With *crosslingual retrieval*, the translated Chinese queries are tokenized with reference to the translation dictionary, while the indexed Chinese audio documents are tokenized with reference to the recognizer's language model. Hence tokenization mismatch is possible, and will affect retrieval performance. This problem may be handled by the use of overlapping character n-grams in indexing and retrieval.

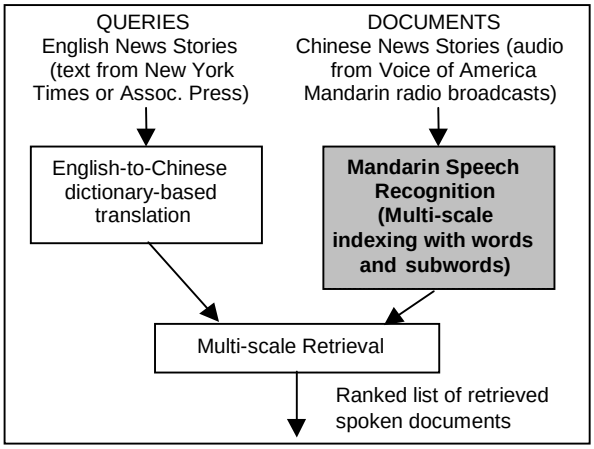


Figure 1. *Overview of the MEI system. In this English-Chinese spoken document retrieval task, the query is formed from an entire English news story (text) from the New York Times or Associated Press. The spoken documents are Mandarin news stories (audio) from the Voice of America news broadcasts. Multi-scale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

(iii) Robustness against Chinese homophone ambiguity – the use of Chinese syllables for indexing and retrieval may alleviate this problem. 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"), 復數 , ("complex number" or "plural"), 覆述 ("repeat") [4]. With *cross-media retrieval*, a word may occur in the textual query, but the recognizer may misrecognize the homophone in place of the word from the audio documents. Retrieval based on words or characters will suffer under this situation, but retrieval based on syllables will not.

(iv) Robustness against the open vocabulary problem – written Chinese can be fully represented by ~6000 characters, and spoken Chinese by ~400 syllables. Hence indexing with characters / syllables provides full lexical / phonological coverage of Mandarin Chinese broadcast news.

(v) Robustness against speech recognition errors – each Chinese character is pronounced as a syllable, but the mapping is many-to-many, producing a large number of homophones. Homophones, acoustically confusable words and out-of-vocabulary (OOV) words are main causes of recognition errors. As mentioned above, subword-based indexing can alleviate the homophone / OOV problems. We can also utilize the recognition N-best hypotheses to provide near-miss alternatives during audio indexing, to enhance robustness against recognition errors.

This paper presents our approach towards multi-scale audio indexing for Mandarin Chinese spoken documents. The use of subwords can handle the problems presented above, and subword N-grams capture sequential constraints [5] to partially compensate for the loss of lexical knowledge which is important for precision. We will compare multi-scale indexing with word-based indexing for English-Chinese CL-SDR.

2. EXPERIMENTAL CORPORA

We use two Topic Detection and Tracking (TDT) collections for our work.[1] TDT-2 is our development set, and TDT-3 our evaluation set. The English news stories (text) from New York Times or Associated Press are used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts are used as our spoken documents. All are manually labeled by topic. Specifics of our corpora are summarized in Table 1. The Mandarin audio documents are furnished with recognized words from the Dragon system [6].

	TDT-2 (Dev set)	**TDT-3 (Eval set)**
English text queries (NYT or AP newswire)	17 topics, each with variable # of exemplars	56 topics, each with variable # of exemplars
Mandarin audio documents (VOA news broadcasts)	2,265 manually segmented stories, 46.03hrs of audio	3,371 manually segmented stories, 98.4hrs of audio

Table 1*. Statistics of TDT-2 and TDT-3 – our development and evaluation data sets.*

3. THE DRAGON BENCHMARK

We tried to assess the performance of Dragon's recognizer on the TDT corpora, whose Mandarin news audio collections have not been manually transcribed, and hence a "gold standard" does not exist. However, we spot checked several news audio files against their *anchor scripts*, and found that these can serve as an approximation of a gold standard for evaluation.

3.1 Word Error Rates

Dragon is a large-vocabulary continuous speech recognizer (LVCSR), and hence its output contains word boundaries (tokenizations) resulting from its language model and vocabulary definitions. The anchor scripts are running text with no word boundaries. In order to compute word error rates, we need to have consistent tokenizations for both the recognition hypotheses (a single word sequence for each audio news story) and the anchor scripts. Hence we discarded the Dragon's word boundaries and re-tokenized both the Dragon's recognition outputs and anchor scripts with the MEI Mandarin lexicon (with 48K entries)[2] using a greedy algorithm [7].[3] We then aligned the two word sequences (hypothesis with reference) and computed word error rates. For simplicity, we only include the *subset* of audio documents whose anchor scripts have less than 500 characters / syllables. Statistics of this subset and their word error rates are shown in Table 2.

3.2 Character / Syllable Error Rates

The character error rate is a simpler evaluation metric than word error rate, as it does not require word tokenization. We simply align the hypothesized and reference character strings for computation. It should be noted that the character error rate is the only evaluation metric that gives a perfect evaluation, and hence it is most popular for Chinese.

To obtain the syllable error rate, both the Dragon's recognition outputs and the anchor scripts need to be converted into syllable strings by pronunciation lookup. We referenced the word pronunciations in the MEI lexicon.

The character and syllable error rates are both shown in Table 2. These results agree well with those reported in [6].

	TDT-2	**TDT-3**
Evaluated subset	1,954 documents 22.98 hrs of audio	2,430 documents 27.52 hrs of audio
Word error rates	17.96 (3.54/2.65/11.77)	19.12 (4.15/2.87/12.10)
Character error rates	12.06 (1.93/0.77/9.36)	12.98 (2.61/0.83/9.54)
Syllable error rates	7.91 (1.94/0.79/5.18)	8.60 (2.61/0.83/5.16)

Table 2. *Error rates (Insertion/Deletion/Substitution) (%) of Dragon's recognized outputs with respect to the anchor scripts on a subset of TDT-2 and TDT-3 Mandarin news audio collections.*

4. MULTI-SCALE AUDIO INDEXING

As mentioned in the introductory section, MEI seeks to investigate the use of multi-scale audio indexing for English-Chinese CL-SDR. In addition to using the word sequence outputs from LVCSR, we also index with Chinese subwords, i.e. overlapping character and syllable n-grams.[4] We will also index with a lattice representation, in which we provide an alternate syllable recognition hypothesis to augment Dragon's hypothesis. This is because speech recognition errors degrade SDR performance, and including alternate (correct) hypotheses may offer partial performance recovery. We also begin by producing a syllable lattice, because the syllable is most basic – syllable-level mismatches already imply word / character mismatches during retrieval.

Hence, multi-scale audio indexing in MEI includes:

(i) the single best word sequence provided by Dragon for each audio document;
(ii) the overlapping character bigrams derived from (i);
(iii) the overlapping syllable bigrams derived from (i) with word pronunciation lookup;
(iv) the alternate syllable hypothesis provided by the MEI syllable recognizer to augment Dragon's syllable

[1] The TDT corpora are pre-released for our work by the Linguistic Data Consortium http://www.ldc.upenn.edu

[2] The MEI Mandarin lexicon is a union of the LDC CALLHOME Mandarin lexicon and all Mandarin single character words.

[3] Using Dragon's own lexicon would have been more desirable, but this resource is not available to us.

[4] Our previous experiments indicate the bigrams are most effective [8,9].

hypothesis. This is captured in a lattice representation for the audio documents

The first three items above can be obtained directly from Dragon's output. In the following we focus on (iv), the incorporation of alternate recognition hypothesis in the hopes of complementing Dragon's output.

4.1 Imperfect Recognition and Syllable-based Indexing

Audio indexing with overlapping subword n-grams is affected by recognition errors such as insertions, deletions and substitutions. Among these, substitutions are the most problematic. A single substitution error will delete two correct bigrams for indexing, and insert two erroneous bigrams instead. Hence it is desirable to include an alternate syllable hypothesis to increase the likelihood of obtaining the correct indexing bigrams.

Our previous work [8] in monolingual Chinese SDR has shown that by incorporating an alternate syllable hypothesis (which complements the top-level syllable hypothesis) in a lattice enhanced robustness in retrieval against imperfect recognition. A noteworthy point is that the alternate syllable hypothesis may be correct or incorrect. The former condition salvages the correct indexing bigrams for retrieval. The latter introduces incorrect indexing bigrams which degrade retrieval. In [8], the syllable recognition error rate was about 30-40%, and we applied a *syllable verification technique* to downweigh the possibly incorrect indexing bigrams by their low recognition scores.

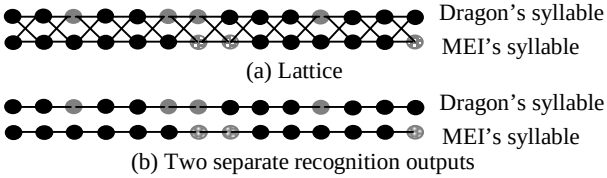
However, in MEI, we have another consideration because Dragon already offers a good recognition performance. Dragon's substitution error rates are around 5% (see Table 2). This implies that if we were to augment every Dragon's syllable with an alternate (and different) hypothesis from the MEI syllable recognizer, then no more than 5% of the included MEI's syllables are correct. Hence we devised a special lattice as illustrated in Figure 2a, in which we inserted the top-scoring syllable hypothesis from the MEI recognizer, and allowed it to be the same as Dragon's hypothesis. We also considered the parallel configuration in Figure 2b instead of the lattice (Figure 2a). The parallel configuration does not allow criss-crossing in between the two strands, and aims to reduce the number of (possibly incorrect) indexing bigrams generated. However, preliminary experiments indicated that the Figure 2a configuration was better.

4.2 The MEI Syllable Recognizer

The acoustic processing in MEI syllable recognizer (MEI_rec) is the same as what was reported in [8]. 112 context-dependent initials and 38 context-independent finals are used in acoustic modeling. Each Chinese syllable has an initial and final. The initial model is a 3-state HMM, and the final model is a 4-state HMM. These were trained on 11 hours of Voice of America from the Hub4 Mandarin corpus from LDC. In addition, the silence model is a single-state HMM trained on non-speech segments. Cepstral mean subtraction was applied to all the spoken documents prior to indexing.

The syllable bigram language model is trained on 40 million Chinese characters derived from the 1998 XinHua newswire text corpus. This newswire text was selected because it is almost contemporaneous with the TDT-2 and TDT-3 broadcast news collections. The training text for our language model was word tokenized by using the MEI lexicon and a greedy algorithm, and transformed into syllable sequences by pronunciation lookup.

A three-stage search procedure was used to reduce recognition time due to the tight Workshop schedule: Stage one – free syllable decoding (FSD) at the sentence level, where a segment of silence longer than 0.2 sec (according to the time-aligned Dragon's outputs) is treated as a sentence boundary. Stage two – a Viterbi search to output multiple syllable candidates for each syllable segment; in order to form an initial lattice. Stage three – an A* search to generate a new best syllable sequence from the initial syllable lattice, where the syllable bigram language model is applied in the forward search and syllable trigram language model applied in the backward search.



(a) Lattice

(b) Two separate recognition outputs

Figure 2*. Bigram terms extracted from different combinations of Dragon's syllables and MEI's syllables.*

	TDT-2 subset (22.98 hrs of audio)	**TDT-3 subset (27.52 hrs of audio)**
FSD	37.09	36.13
FSD+LM	25.88	25.53
FSD + MAP	12.28	12.90
FSD + MAP + LM	10.05	10.82
WLFA + MAP	15.98	16.32
WLFA + MAP + LM	14.27	15.80

Table 3*. Error rates (%) of different approaches of the MEI syllable recognizer.*

We also implemented a MAP speaker adaptation procedure, based on Dragon's high performance output, in an attempt to refine the acoustic models in the MEI_rec.

Recognition experiments were first conducted on TDT-2. Results are summarized in Table 3. Baseline free syllable decoding (FSD) has an error rate of 37.09%. This drops to 25.88% with the syllable bigram language model (FSD+LM); and further to 10.05% by adding MAP speaker adaptation (FSD+MAP+LM). This is the lowest error rate for MEI_rec, which is about 2% more than Dragon. Running MEI_rec on a given audio sentence may not produce the syllable sequence with the same length as Dragon. But for lattice construction, we need to obtain syllable hypotheses in locked step with each other (see Figure 2). To achieve this, we performed a word-level forced alignment (WLFA) on the audio with Dragon's output, and used MEI_rec for syllable recognition in between adjacent word boundaries. Error rates for this procedure with (WLFA+MAP+LM) and without (WLFA+MAP) the language model are 14.27% and 15.98% respectively. Recognition between fixed sentence boundaries has lower error rate than between fixed word boundaries, because the latter is more restrictive, but is necessary for lattice construction. Results from TDT-3 display the same trends as TDT-2.

5. RETRIEVAL EXPERIMENTS

Experiments were run with the setup in Figure 1. Each experiment is run with a batch of queries. A batch contains one story exemplar per topic. Each query retrieves the top 1000 audio stories and their relevance is evaluated in terms of non-interpolated average precision (nAP). For each topic, we average the nAP attained by its queries, and thereafter average across all topics to obtain the mean average precision (mAP). Referring to Figure 1, selected terms in the English query exemplar were

translated into Chinese using a machine-readable dictionary. Based on the translated Chinese query, we produced query representations in terms of words and overlapping character / syllable bigrams. These match with the document representations from audio indexing during retrieval. We used INQUERY, which is developed at the University of Massachusetts [10], as our retrieval engine. Highlights of our extensive experiments are presented in the following. Details are described in [3].

5.1 Retrieval Robustness against Recognition Errors

Our first set of experiments investigated how the substitution errors affected retrieval. Our queries include all terms in the English exemplar, and one translation per term. Table 4 shows retrieval results (mAP) on syllable bigram indexing. Dragon's syllable bigrams gave 0.271, which improved to 0.279 if we replaced all substitution errors with their correct syllables (a reference upper bound). This rose further to 0.286 if we deleted the erroneous indexing bigrams due to substitution errors, which suggests that audio segments of less importance for retrieval tend to be spoken less clearly and misrecognized. This may imply that term selection from the document collection may help SDR, and is worthy of further study.

We expected that lattice usage has only a narrow margin of improvement since the substitution error rate is low (~5%, see Table 2). We formed the *ideal* lattice by augmenting Dragon's syllables with the *correct* syllables, which gave an improved retrieval performance at 0.303. If instead we formed the lattice with the MEI syllable hypotheses (WLFA condition), retrieval performance degraded to only 0.268. This should likely due to MEI_rec's errors and their associated erroneous indexing bigrams. We are considering methods to downweigh potentially erroneous indexing bigrams, e.g. by incorporating speech recognition scores into retrieval.

5.2 Subwords Bring Benefits

Our second set of experiments investigated the benefits of augmenting words with subwords in multi-scale audio indexing. Here, each query in the batch includes 50 statistically selected terms from the exemplars, and all translation alternatives are used. Results are shown in Table 5. Character bigrams outperformed words, and syllable bigrams produced competitive performance. The lattice performed poorly, possibly due to the erroneous indexing bigrams generated, but the issue warrants further investigation. Trends are mostly consistent between TDT-2 and TDT-3. Results on TDT-2 (our development set) are better because various optimizations were performed on it. It is, however, surprising that the MEI syllable bigram indexing showed reverse trends.

In general, we observe that augmenting words with subwords in multi-scale indexing brings benefits to CL-SDR. Fusion of words and subwords in retrieval gives further improvements. Fusion strategies are described in [11].

6. CONCLUSIONS

In the MEI project, we have designed and developed one of the first English-Chinese CL-SDR systems, and advocated a novel multi-scale paradigm for the task. In this paper, we discuss the development of the MEI syllable recognizer, the representations of spoken documents in terms of overlapping n-grams or lattice structures. Results show that the multi-scale approach is beneficial to the performance in CL-SDR.

Dragon's syllable (DragSyl)	0.271
DragSyl, replace substitutions with desired syllables	0.279
DragSyl, erroneous bigrams deleted	0.286
DragSyl + desired syllable (Best possible lattice)	0.303
DragSyl + MEISyl (Real lattice)	0.268

Table 4*. Retrieval based on syllable bigram indexing, tested on the TDT-2 subset. Mean average precisions are shown.*

	TDT-2	**TDT-3**
Words	0.464	0.462
Character bigrams	0.514	0.475
Dragon syllable bigrams	0.468	0.422
MEI syllable bigrams	0.446	0.499
Lattice (Dragon + MEI syl)	0.318	0.300

Table 5*. Retrieval performance on the full sets of TDT corpora. MEI syllable bigrams are based on (FSD+MAP+LM), and the lattice is based on (WLFA+MAP+LM).*

7. ACKNOWLEDGMENTS

We wish to acknowledge our fellow MEI team members for their contributions: Erika Grams, Sanjeev Khudanpur, Gina Levow, Douglas Oard, Karen Tang and Jianqiang Wang. The project is conducted during the Johns Hopkins University Summer Workshop 2000 (an NSF Workshop). 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 and Richard Schwartz for their help. We are grateful to Fred Jelinek and his staff at CLSP for organizing the workshop.

8. REFERENCES

1 H. Meng et al., "*Mandarin-English Information (MEI),*" Proc. of Topic Detection and Tracking Workshop, 2000.

2 H. Meng et al., "*Mandarin-English Information (MEI) - Investigating Translingual Speech Retrieval,*" Proc. of NAACL Workshop on Embedded Machine Translation 2000.

3 H. Meng et al., "*Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval,*" Technical Report for the Johns Hopkins Univ. Summer Workshop 2000.

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

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

6 P. Zhan, S. Wegmann, and L. Gillick, "*Dragon Systems' 1998 Broadcast News Transcription System for Mandarin,*" Proc. of the DARPA Broadcast News Workshop 99.

7 K.J. Chen and S.H. Liu, "*Word Identification for Mandarin Chinese Sentences,*" Proc. of COLING-1992, pp. 101-107.

8 H.M. Wang, "*Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese,*" Speech Communication, 32 (1-2), pp. 49-60, 2000.

9 H. Meng et al., "*Multi-scale Audio Indexing for Chinese Spoken Document Retrieval,*" Proc. of ICSLP-2000.

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

11 H. Meng et al., "*A Multi-Scale Paradigm for English-Chinese Cross-Language Spoken Document Retrieval,*" submitted to SIGIR-2001.