Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval

Helen M. Meng[1], Wai-Kit Lo[1], Berlin Chen[2] and Karen Tang[3]
The Chinese University of Hong Kong[1], National Taiwan University[2], Princeton University[3]
hmmeng@se.cuhk.edu.hk, wklo@ee.cuhk.edu.hk, berlin@iis.sinica.edu.tw, kpytang@princeton.edu

ABSTRACT

We have developed a technique for automatic transliteration of named entities for English-Chinese cross-language spoken document retrieval (CL-SDR). Our retrieval system integrates machine translation, speech recognition and information retrieval technologies. An English news story forms a textual query that is automatically translated into Chinese words, which are mapped into Mandarin syllables by pronunciation dictionary lookup. Mandarin radio news broadcasts form spoken documents that are indexed by word and syllable recognition. The information retrieval engine performs matching in both word and syllable scales. The English queries contain many named entities that tend to be out-of-vocabulary words for machine translation and speech recognition, and are omitted in retrieval. Names are often transliterated across languages and are generally important for retrieval. We present a technique that takes in a name spelling and automatically generates a *phonetic cognate* in terms of Chinese syllables to be used in retrieval. Experiments show consistent retrieval performance improvements by including the use of named entities in this way.

1. INTRODUCTION

We have developed an English-Chinese cross-language spoken document retrieval (CL-SDR) system, where English textual queries are used to retrieve Mandarin spoken documents, i.e. a cross-language and cross-media information retrieval task. With the growing multi-media and multi-lingual content in the global information infrastructure, CL-SDR technologies are potentially very powerful, as they enable the user to search for personally relevant audio content, (e.g. recordings of meetings, lectures or radio broadcasts), across the barriers of language and media.

Our system accepts an *entire* English textual story (from newspapers) as the input query, and automatically retrieves relevant Mandarin audio stories (from radio broadcasts). We refer to the English story as our *query exemplar*, and this retrieval context as *query-by-example*. Our task is illustrated in Figure 1. Mandarin is the key dialect of Chinese. English and Chinese are two predominant languages used by the global population. They are very different linguistically, hence English-Chinese CL-SDR presents unique research challenges.

A prevailing problem in our task is that the topically diverse news domain contains many named entities, and these are often out-of-vocabulary words (OOV) in recognition and translation. In word recognition for audio indexing, OOV[1] may be erroneously substituted by other in-vocabulary words. Our solution to this problem is to use syllable recognition, where the OOV is transcribed as its constitutent syllables. This is feasible because a compact inventory of approximately 400 base syllables can provide full phonological coverage for the Chinese language. Additionally, a syllable forms the pronunciation of a Chinese character with a many-to-many mapping. An inventory of approximately 6,000 characters provides full textual coverage in Chinese. However, the Chinese word may consist of one to multiple characters, hence character combinations can produce an unlimited number of Chinese words. There is no explicit word delimiter and the task of segmenting a character sequence into a word sequence contains much ambiguity. Consequently, we have augmented word-based retrieval with character- and syllable-based retrieval. We use overlapping character/syllable n-grams to circumvent the problem of tokenization ambiguity. Character/Syllable bigrams fare best among *n*-grams in retrieval performance, and character bigrams outperform words, based on our experiments with the Topic Detection and Tracking (TDT) Collection from the LDC.[2]

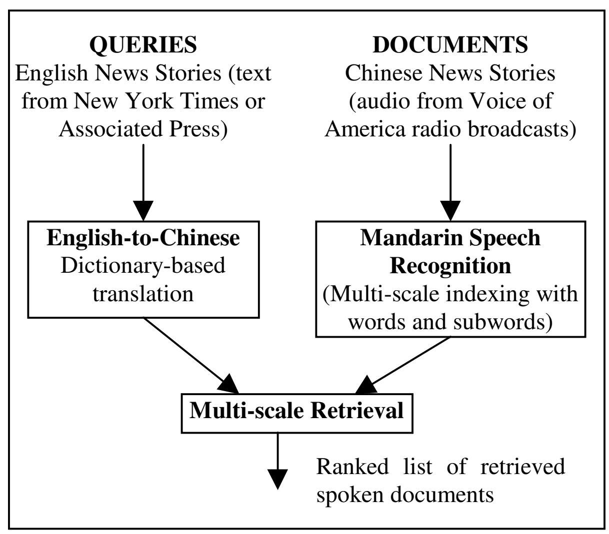

Figure 1. Overview of our English-Chinese Spoken Document Retrieval Task.

The OOV problem is also present in English text query translation – query terms absent from our translation dictionary implies that they will not be translated into the Chinese query for

[1] These are words unknown to the speech recognizer.

[2] Linguistic Data Consortium, http://www.ldc.upenn.edu/

subsequent retrieval. Very often these are named entities, i.e. names of people, organizaiton, location, etc., which are in fact important for retrieval. If we reference contemporaneous English and Chinese news corpora, we will find that named entites are often transliterated from the source to the target language. Transliteration involves generation of a phonetic cognate, i.e. the transliteration of a name into the target language aims to achieve a pronounciation similar to that in the source language of origin. For example, "Ireland" is commonly transliterated as 愛爾蘭, which is pronounced as /ai-er-lan/ in pinyin transcription for the Chinese syllable. However, there are no hard-and-fast rules in the generation of phonetic cognates, and the mapping may have variations. For example, consider the translation of "Kosovo" (pronounced /k ow so ax v ow/[1]) – sampling Chinese newspapers in China, Taiwan and Hong Kong 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 incorporate named entities into retrieval, we have developed an automatic names transliteration procedure that involves *cross-lingual phonetic mapping* (CLPM) to generate phonetic cognates. A similar idea has previously been applied to English/Japanese (Katakana) and English/Arabic translation (Knight and Graehl, 1997), (Stalls and Knight, 1998). Ours is one of the first attempts for English/Chinese transliteration which also incorporates automatic English spelling-to-pronunciation generation followed a mapping of English phones into Chinese syllables.

2. NAMED ENTITY TRANSLITERATION

Figure 2 presents an overview of the named entities transliteration process. Our English query exemplars have been tagged by the BBN Identifinder (Bikel et al., 1997) system for named entities. The tagged units which are not found in our translation dictionary will be processed by our transliteration system. In the following, we provide a description for every module in Figure 2.

2.1 Detect Chinese Names

The first step in our process is to detect romanized Chinese names. These may be in the (commonly used) Wade Giles or pinyin conventions.[2] We have extracted the two syllable inventories from the Internet, as well as the mapping from Wade Giles to pinyin. Detection of romanized Chinese names is achieved by a left-to-right maximum-matching (greedy) segmentation algorithm. The two syllable lists are used in turn for segmentation, since only one convention will be used at a time. If we can successfully segment the input named entity into a sequence of Chinese syllables, our procedure returns the corresponding pinyin syllable sequence, which can be used for query formulation in retrieval. Otherwise we proceed to the next step.

[1] This English pronunciation is transcribed with ARPABET symbols.

[2] http://lcweb.loc.gov/catdir/pinyin/romcover.html/

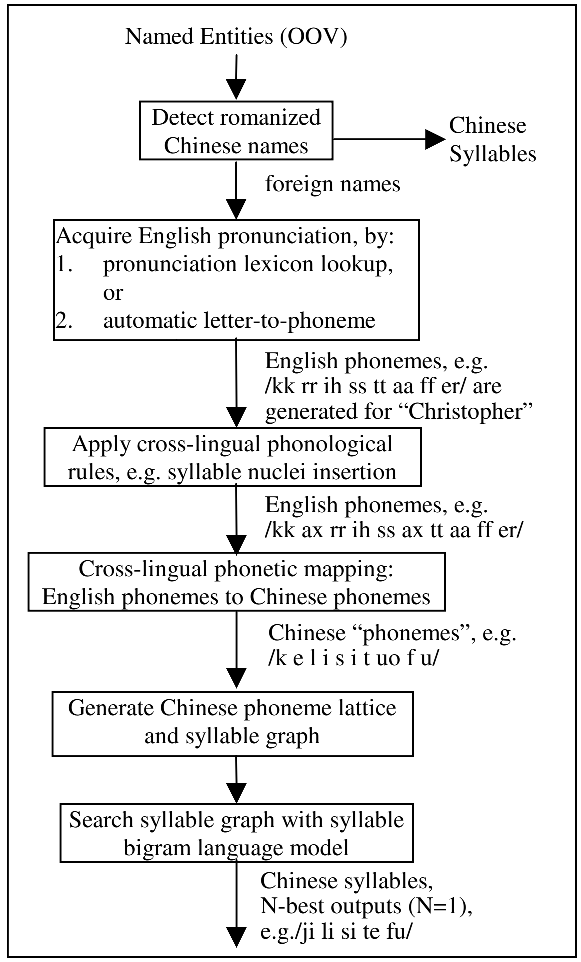

Figure 2. Overview of our named entity transliteration process.

2.2 Generate English Pronunciations

If the input is not a romanized Chinese name, we attempt to automatically acquire a pronunciation for the foreign name in terms of English phonemes. We begin by looking up the pronunciation lexicon PRONLEX provided by LDC. If the name is found, this procedure outputs an English phoneme sequence. Otherwise the spelling of the name is passed to our automatic letter-to-phoneme generation process.

Our letter-to-phoneme generator applies a set of rules to generate an English pronunciation from the input spelling. This set of letter-to-phoneme rules has been automatically inferred from data by the following process: We used the entire PRONLEX lexicon which contains 90,000 words for training. For each word, we aligned the spelling with the pronunciation in a Viterbi-style to achieve a one-to-one letter-to-phoneme mapping, e.g. “appraise” is aligned with /ax pp null rr ey null zz null/. A /null/ phoneme is inserted when we encounter geminate letters, or in cases where more than one letters map into a single phoneme. We then apply the transformation-based error-driven learning (TEL) approach (Brill 1995) to these alignments to obtain a set of transformation rules for spelling-to-pronunciation generation. Referring to Figure 2, these rules were able to generate the pronunciation /kk rr ih ss tt aa ff er/[1] for the input spelling “Christopher”.

2.3 Apply Cross-Lingual Phonological Rules

Chinese is monosyllabic in nature, but English is not. Therefore we observe some phonological differences between the two languages. For example, the name *Bush* is pronounced as a single syllable /bb uh sh/ in English, but transliterated as two syllables in Chinese – /bu shu/. Another example, e.g. *Clinton* /kk ll ih nn tt ih nn/ contains a consonant cluster (/kk ll/), but its Chinese transliteration inserts a syllable nucleus in between the consonants, and is pronounced as /ke lin dun/.

We have written a set of phonological rules to transform the English pronunciation, in an attempt to bridge some of the discrepancies mentioned above. This serves to ease the subsequent process of cross-lingual phonetic mapping (CLPM). Examples of rules include:

- Insert a reduced syllable nuclei (the ‘schwa’ /ax/) between clustered consonants. This takes care of pronunciations as in the example *Clinton* mentioned earlier.
- Duplicate the nasals /mm/, /nn/ and /nx/ (syllabic nasal) whenever they are surrounded by vowels. For example, *Diana*, pronounced as /dd ay ae nn ax/ in English, is often transliterated as /dai an na/ in Chinese, where the nasal /nn/ forms part of the syllable final in the second syllable, as well as the onset of the third syllable.
- For all consonant endings, except /ll/, append a syllable nuclei (/ax/) to it. For example, *Bennett*, pronounced as /bb eh nn ih tt/ in English, is often transliterated as /bei nei te/ in Chinese. If the syllable ends with /ll/, it is treated differently – consider the example *Bell*, pronounced as /bb eh ll/ in English, and often transliterated as /bei er/ in Chinese.

2.5 Cross-lingual Phonetic Mapping (CLPM)

This procedure aims to map the English phonemes into Chinese “phonemes” (derived from syllable initials and finals) by applying a set of transformation rules. Again, these rules are learnt automatically from data by the technique of transformation-based error-driven learning (TEL). The process is as follows:

We collected a bilingual proper name list which contain English proper names with their Chinese transliterations. Our list is derived from LDC’s English-Chinese bilingual term list with CETA (Chinese-English Translation Assistance), a list from the National Taiwan University,[2] and some name pairs harvested from the Internet. We randomly allocated training and test sets, with 2233 and 1541 names respectively. Each name pair contains the English name and corresponding Chinese translation / transliteration. We looked up the English name pronunciation from PRONLEX, and the Chinese pronunciation from LDC’s Mandarin CALLHOME lexicon.

We obtained a one-to-one phoneme-to-phoneme alignment between the English name pronunciation and the Chinese name pronunciation by means of a finite-state transducer (FST) (Mohri et al., 1998). The FST was initialized with some obvious English-phoneme-to-Chinese-phoneme correspondences, and trained iteratively on a set of phoneme pairs until convergence is reached. The converged FST is used to align our training words, and then we applied TEL to derive a set of transformation rules to map English phonemes into Chinese phonemes. Given a testing English phoneme sequence, application of our transformation rules will generate a single Chinese phoneme sequence.

2.6 Generate a Chinese Phoneme Lattice

Based on an English phoneme sequence, CLPM generates a single Chinese phoneme sequence as output. We need to apply Chinese syllabic constraints to this phoneme sequence to produce a syllable sequence (in pinyin). However, this Chinese phoneme sequence may contain errors. In order to include phoneme alternatives prior to syllabification, we try to capture common confusions in CLPM. To do this, we applied our transformation rules to each English pronunciation in the training set, and compared the generated Chinese phoneme sequence with the reference sequence to produce a confusion matrix. The matrix stores the frequency of confusion for each reference-phoneme/output-phoneme pair.

Upon testing, the confusion matrix is used to generate a phoneme lattice prior to syllabification. A phoneme lattice is illustrated in Figure 3. Given an English name (Cecil Taylor) and its English pronunciation (note that this is an over-generalization because not all names are of English origin, but we treat them as such for the sake of simplicity in letter-to-phoneme generation), we applied CLPM to give a corresponding Chinese phoneme string /s a x e er t ai l e/ (first row of nodes). For each Chinese phoneme in this string, we expand with all its confusable alternatives by referencing the confusion matrix. For example, the first Chinese phoneme /s/ has been confused with /a/ and /k/, and these are inserted to form a lattice. Similarly, the second phoneme /a/ has been confused with /ai/ which gets inserted as well. The inserted nodes in the lattice are also weighted by their probability of confusion, derived from the statistics in the confusion matrix. The expanded nodes serve to provide alternative phonemes for syllabification.

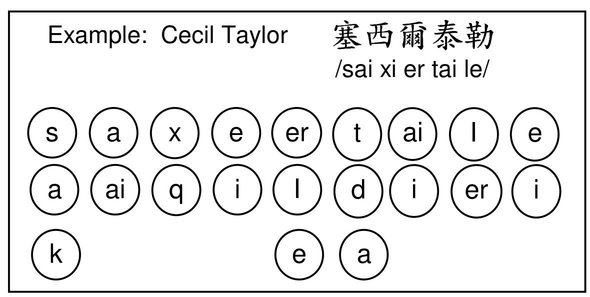

Figure 3. Example of a phoneme lattice generated from the output of CLPM.

2.7 Search Syllable Graph with a Syllable Bigram Language Model

We search our phoneme lattice exhaustively for Chinese phoneme sequences which can constitute legitimate syllables, to create a syllable graph (see Figure 4). We then traverse the graph by A* search to find the *N* most probable syllable sequence. Probabilities derived from the confusion matrix, as well as those from a syllable bigrams language model are considered. The syllable bigram language model is trained from

[1] The /null/ phoneme has been discarded in the generated output.
[2] This list is provided by H. H. Chen from National Taiwan University.

a list of 3,628 Chinese names harvested from the Internet. This configuration is capable of hypothesizing *N*-best syllable sequences – we currently set $N=1$ for the sake of simplicity. The idea behind this step and the previous one is inspired by lexical access in speech recognition, which produces word hypotheses from a lattice of recognized phones. Indeed if we use a character bigram instead of the syllable bigram during A* search, we can *potentially* generate and *N*-best list of character sequences, e.g. generating 基里斯特弗for Christopher. The pronunciation of the character sequence is /ji li si te fu/. Based on our test set of 1541 names, this procedure gave a transliterated syllable accuracy of about 47.5%.

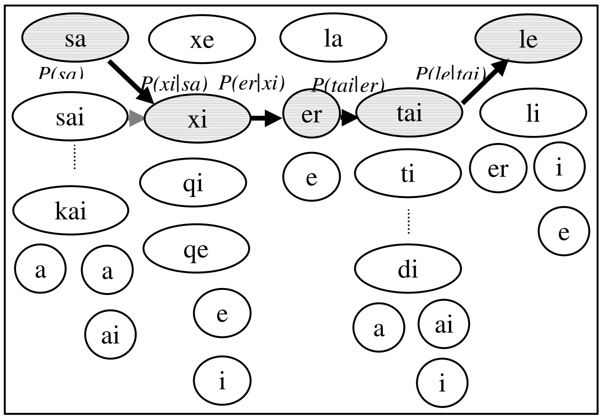

Figure 4. Syllable graph of the phoneme lattice in the previous example (in Figure 3).

3. IMPACT ON ENGLISH-CHINESE CL-SDR PERFORMANCE

We have incorporated the automatic names transliteration procedure into our task of English-Chinese CL-SDR. The experiment was based on the TDT Collection. Query exemplars were drawn from English news text (from the New York Times and Associated Press). Audio documents were drawn from Voice of America news broadcasts in Mandarin. The TDT collection has manual, exhaustive topic annotations that serve as relevance judgements for retrieval. There are 17 topics in total in the collection, and we included up to 12 query exemplars for each topic in our retrieval experiments.

Retrieval performance is measured by non-interpolated mean average precision. (mAP). As mentioned earlier, we used both words and character bigrams for retrieval, and the latter outperforms the former, as shown in the TDT-2 results in Table 2. We extracted the 200 most common named entities that have been tagged in our query exemplars (by the BBN Identifinder). These are processed by our named entity transliteration procedure and the output syllable sequences are used to augment the translated Chinese query. From Table 2 we see that named entity transliteration brought about small but consistent improvements to both word-based and character-based retrieval. The improvement is not statistically significant, though we believe this is due to the limited number of names have been transliterated. This is an ongoing research effort, and we plan to further investigate ways to enhance retrieval performance by handling OOV via transliteration.

	Baseline	With Translit.
Words	0.464	0.471
Character bigrams	0.514	0.522

Table 2. Performance evaluation for English-Chinese CL-SDR (mAP). The named entity transliteration procedure brought improvements to both word-based and subword-based (character bigrams) retrieval.

4. CONCLUSIONS

In this paper, we have presented a named entity transliteration technique for English-Chinese cross-lingual spoken document retrieval. In our retrieval task, the English queries often contain named entities that are absent from our translation dictionary, As a consequence, these names cannot be utilized for retrieval. To address this problem, the named entity transliteration procedure automatically generates a Chinese syllable sequence (in pinyin), based on the English spelling of the named entity. This syllable sequence is incorporated during query formulation, and used in retrieval by matching with the documents in syllable space.

We have adopted a data-driven approach for named-entity transliteration. The process involves automatic English spelling-to-pronunciation generation followed by application of cross-lingual phonetic mapping to transform the English pronunciation into its Chinese phonetic cognate(s). Transliterated syllable accuracy is about 47.5%. We ran retrieval experiments based on the Topic Detection and Tracking collection from the LDC. With named entity transliteration, word-based retrieval was improved from 0.464 to 0.471. Character-based retrieval was improved from 0.514 to 0.522. Our results suggest that the named entity transliteration procedure shows promise in salvaging untranslatable names to improve English-Chinese CL-SDR performance.

5. ACKNOWLEDGMENTS

This work is part of the Mandarin-English Information (MEI) project conducted at the Johns Hopkins University Center for Language and Speech Processing Summer Workshop 2000. We acknowledge contributions of the MEI team and those who supported the MEI project.

6. REFERENCES

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

Brill, E.. 1995. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4): pp 1-37.

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

Mohri, M., Pereira , F. and Riley, M., 1998. A Rational Design for a Weighted Finite-State Transducer Library. *Lecture Notes in Computer Science*, 1436.

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