Automatic Detection of Chinese Transcription of English Personal Names based on Regular Expression Matching of Consonant Clusters

Wenxin Xiong

National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China

xiongwenxin@bfsu.edu.cn

Keywords: English personal name; Chinese transcription; consonant cluster; Regular expression.

Abstract. Automatic detection of the transcription of personal names in different languages is a key component for cross-lingual information retrieval systems. A plenty of complex machine learning algorithms have been employed in machine transliteration tasks with considerable computational costs. We propose an efficient linguistics-based solution for recognizing its Chinese transcription for an English personal name with little computational burden. A phonological knowledge-base for mapping English phonetic representation to Chinese phonetic representation is constructed offline firstly, and then a lenient matching scheme based on consonant cluster skeleton implemented by fuzzy regular expression matching is carried out during online processing. An experiment achieving state-of-the-art performance shows its efficiency and effectiveness.

1. Introduction

Recognition of transliterated personal name and its original name is important for cross-lingual information retrieval tasks. Name transliteration has attracted much attention for decades. A series of shared tasks on transliteration have been conducted since 2009. Transliteration is defined as phonetic translation of names across languages. The representation in the target language (i) is phonemically equivalent to the source name, (ii) conforms to the phonology of the target language, and (iii) matches the user intuition of the equivalent of the source language name in the target language [1].

2. Related Work

Machine transliteration can be generally grouped into two major tasks: (1) extracting transliterated word pairs from comparable corpora; and (2) developing automatic transliteration systems.

Simon et al. [2] introduced a novel approach to automatically extract divergent transliterations of foreign named entities by bootstrapping co-occurrence statistics from tagged Chinese corpora. The transliteration variants can be detected by means of their distributions in parallel corpora statistically.

Based on the units to be transliterated, Oh et al. [3] differentiated four machine transliteration models: grapheme-based model, phoneme-based model, hybrid model, and correspondence-based model. Grapheme-based transliteration directly transforms source language graphemes into target language graphemes without any phonetic knowledge of the source language words. The others utilize source language phonemes as a pivot to target language graphemes from source language graphemes. Basically Transliteration is a phonetic process, rather than an orthographic one [4]. The phonetic information is important to detect transliteration. Amount of machine learning algorithms have been applied from joint Source-Channel Model, Hidden Markov Model, Conditional Random Field, and deep learning techniques, with good performances and computational complexities.

Halpern cleared up the differences between Transliteration, Transcription and Romanization, and proposed linguistic knowledge base to support accurate conversion of personal names between different languages [5]. The so-called transliteration of personal names from English to Chinese is indeed a phonetic-based transcription process of scripts in different writing systems. We tried to build mapping rule sets from English phonetic representation to Chinese one, and then made some matching operations based on linguistics-inspired consonant cluster based skeleton scheme.

3. Linguistically-inspired solution for Chinese transliteration

Transcription of personal names across languages heavily relies on the sound similarity between source and target pronunciations. As English and Chinese have their own different writing systems respectively, it is almost impossible to computing the similarities between words in their original writing forms. Though Chinese is represented by means of a string of ideograms, with little hints for its pronunciation, there is a national standard for spelling of Chinese word, namely Pinyin, which utilizes Latin alphabet. So the Romanization from Chinese character to Pinyin makes it feasible for comparing two strings of Chinese pronunciation and English spelling. In the other hand, there is no strict correspondence between English spelling and its pronunciation, a conversation process from English word to its sound representation should be employed. The overall workflow is depicted as Fig. 1. There are three steps for detecting the sound similarity between Chinese and English personal name, i.e. A) Romanization of Chinese characters by Pinyin, B) English phonetic transcription by pronunciation dictionary, C) a matching algorithm inspired by linguistic heuristic rules.

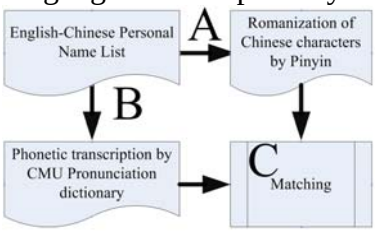

Fig. 1 Workflow for detection of English/Chinese personal names

3.1 Romanization

The transformation from Chinese characters to pinyin strings is carried out first. There are plenty of programs which may accomplish this task. A Perl module named Lingua::Han::Pinyin is employed in this conversation processing. Because Chinese is a tonal language, the program rendered a string with tonal mark ending in each word spelling form. Tonal marks are implemented in Arabic numerals (1, 2, 3, 4), which is not used in comparison of different personal names, therefore the tonal marks were dropped in the transformation process. The pronunciation of Chinese counterpart of an English personal name “Clinton” is “Ke4 Lin2 Dun4”, and is abbreviated as “Ke Lin Dun”.

3.2 Transcription

Cmudict 0.7 is an open-source machine-readable pronunciation dictionary for North American English that contains over 134000 words and their pronunciations. It was developed by School of Computer Science at Carnegie Mellon University. Its entries are particularly useful for speech recognition and synthesis, as it has mappings from words to their pronunciations in the ARPAbet phoneme set, a standard for English pronunciation. The current phoneme set contains 39 phonemes; vowels carry a lexical stress marker. Because ARPAbet can be used to represent each phoneme of general American English with a distinct sequence of ASCII characters, rather than International Phonetic Alphabet (IPA) always encoded in Unicode, it is easy for computational operation.

Category	Representation	Category	Representation
word	Clinton	**Chinese characters**	克林顿
ARPAbet	K L IH N T AH N	**Pinyin**	Ke Lin Dun
IPA	ˈklɪntən	**IPA**	kʰɤ lin tun

Fig. 2 Phonetic transcription from its original writing system

An output of phonetic transcription from English and Chinese words to their pronunciations is depicted as Fig. 2. The transliterated candidates are formatted in ARPAbet and Pinyin, both in Latin letters, for further comparison.

There exist many-to-many correspondences for a letter between English writing and oral systems. For example, the English letter *c* could be recorded as /k/, /s/ or others, but could only be pronounced as /k/ in a consonant cluster “cl”. It means that the pronunciation of an individual letter in an English word may be affected by its surrounding letters. An English word was segmented into syllable chunks according to phonological rules. The actual sound of a letter is bounded in a syllable. When an English word out of the pronunciation dictionary was met, a heuristic mapping rule was applied.

3.3 Matching algorithm

As the transformation process has been accomplished, a matching algorithm is triggered for comparing the similarity between Chinese and English words. Because the two languages have their own respective phonological systems, a proper mapping strategy between them should be developed.

A. Phonological parallelism

In Chinese, there is no distinction between voiced and unvoiced sound, i.e. /g/ and /k/, /d/ and /t/, therefore the different letter pairs should be discarded in consideration of their transcription of English counterparts. It means that both /k/ and /g/ in Chinese could match English consonant /g/ and /k/ interchangeably. Meanwhile Englishman does not differentiate unaspirated and aspirated initials, although the difference does exist in Chinese, such as /b/ and /p/. While evaluating phonological parallelism between English name "Porter" and its Chinese equivalent "Bo Te", phoneme /b/ and /p/ could be deemed as identical. A knowledge-base aiming at finding sound correspondence between Chinese and English phonology is established, and utilized in a further matching strategy.

B. Consonant-skeleton comparison

There is a balance between precision and recall in language processing projects. If exact matching strategy is applied, precision would be high while recall downgrades. On the other hand, if recall is given more priority, the lenient strategy works well. As for our task in automatic detection of personal name pairs in different languages, the problem can be illustrated as a binary classification task: given two candidate strings represented by Chinese pinyin and English phonetic form, a decision is made whether they are identical with each other, according to the likelihood of surface forms. The matching criteria could be in strict or lenient mode, dependent on its granularity and flexibility.

Soundex is a notable phonetic algorithm for indexing names by sound, as pronounced in English. Its goal is to make homophones to be encoded to the same notation so that they can be matched despite minor differences in spelling. It first drops all vowel letters, classifies consonant letters which share same characteristics in pronunciation into one category, and assigns it with a numeral digit. A Soundex code consists of a letter followed by there numeral digits. So English name *Robert* can be represented as *R163*, where *R* is the first letter in *Robert*, and 1 for the second consonant letter *s*, 3 for the third consonant letter *r*, and 3 for the fourth *t*. *Robert* and *Rupert* can be deemed as identical by same codes. A consonant-prominent comparison strategy works well for the normalization of different written forms with same pronunciation. Phonetic Soundex algorithm has been extended to other languages with success. The drawback existed in its limit units to be compared (with only four units) and overgeneralization (some different consonants are encoded in the same form).

Inspired by the consonant salience scheme, we proposed a lenient matching method concentrating on the comparison of consonant letter cluster of a name. We applied the correspondence mappings of English and Chinese consonants described in the former sections, to evaluate their similarity between two names. Our method differs from Soundex in that (1) no numeral digit is assigned, with the exact consonant letter in comparison; (2) all consonants are kept in comparison; and (3) the phonetic similarity metrics of consonants are based on contrastive analysis between two languages. A comparison of two transcriptions is simplified as comparison of two consonant sequences with some reasonable phonological variations. It is so-called lenient consonant-based skeleton strategy.

4. Experiment & Discussion

We have conducted two experiments to validate the lenient consonant-based skeleton method. The test data were collected from (1) English-Chinese Dictionary [6], and the revised version of *Names of the World's Peoples* [7]. About 2400 English personal names in the appendix of the former dictionary were extracted in conjunction with their phonetic transcriptions and the Chinese translations. While there is no original pronunciation information for a person name in the latter word list, CMUdict and other pronunciation dictionaries were construed for its real pronunciation.

We applied consonant-based skeleton algorithm for the automatic detection of same Chinese and English names. The result is shown in Table 1. The strict matching is operated on the exact string

matching, while lenient matching is based on the skeleton strategy empowered by correspondence between Chinese and English phonological constraints.

Table 1 Result of strict and lenient matching

Corpus	Total entries	Strict matching (positive)		Lenient matching(positive)	
		Match	Precision	Match	**Precision**
English-Chinese Dictionary	2388	719	31.37%	2112	**88.44%**
Names of the World's Peoples	**100**	**35**	**35.00%**	**95**	**95.00%**

Take a personal name pair of English name "Churchill" and Chinese name "QiuJier" as an example, "Churchill" is encoded as "ch@@ chi'l" in CMU dictionary. There is little similarity in the surface form between ARPAbet and Pinyin systems; therefore they are unlikely to be considered as identical with each other. While English consonant 'ch' resembles Chinese consonant 'q' and 'j', the English dark consonant "l" sounds like Chinese vowel 'er'. Based on the consonant-based skeleton strategy, English "Churchill" and Chinese "QiuJier" have consequently correspondent consonants 'ch' and 'q', 'ch' and 'j', and 'll' and 'r' literally, therefore, a fuzzy regular expression of two consonant clusters /ch.*ch.*l/ and /q.*j.*r/ could be implemented to evaluate the similarity between two strings. Some previous studies reported about 90% recognition of transcription pairs with much computing work [5]. Our experiments showed precision was 88.44% and 95.04% in two different test beds respectively, achieving the state-of-the-art results.

5. Conclusion

Our linguistically-inspired method makes use of phonetic mapping rules across different languages. Knowledge base pertaining to sound parallelism between two languages, and the phonological constraints insider one particular language, is built offline in advance. As in online matching process, sound similarity is computed based on consonant-cluster skeletons of the two strings. A simplistic fuzzy regular expression matching technique works well in this task.

Due to relatively small size of the knowledge base built in this pivot study, and the coarse-grained hand-crafted correspondence rules, there still is room for much improvement. In the future, more data will be collected for the further solid construction of a sound knowledge base. Some other machine learning algorithms can be adopted for accuracy and efficiency.

Acknowledgments

This work was supported by National Social Science Fund (11BYY051) and Beijing Social Science Fund (16YYB018).

References

[1] H. Li, A. Kumaran, V. Pervouchine, et al. Report of NEWS 2009 machine transliteration shared task. *The ACL/IJCNLP-2009 Named Entities Workshop*. Singapore, August 7, 2009, p.1-18.

[2] P. Simon, C. Huang and S. Hsieh. Transliterated named entity recognition based on Chinese Word sketch. *International Journal of Computer Processing of Languages*, Vol.1 (2008), p.230-236.

[3] J. Oh, K. Choi and H. Isahara. A comparison of different machine transliteration models. *Journal of Artificial Intelligence Research.* Vol. 27 (2006) 119-151.

[4] K. Knight and J. Graehl. Machine transliteration. *Comput. Linguist.* Vol. 4(1998), p. 599-612.

[5] J. Halpern. Some linguistic issues in the machine transliteration of Chinese, Japanese, and Arabic names. *The ACL-2016 Named Entity Workshop*. Berlin, August 12, 2016, p.47-48.

[6] G. Lu. *The English-Chinese Dictionary*. (Shanghai Translation Publishing House, China 2006).

[7] Xinhua News Agency. *Names of the World's Peoples*: A Comprehensive Dictionary of Names in Roman-Chinese, 2nd edition (China Translation & Publishing Corporation, China 2006)

