Generating Paired Transliterated-cognates Using Multiple Pronunciation Characteristics from Web Corpora

Jin-Shea KUO [1, 2]
[1]Chung-Hwa Telecommunication Laboratories, 12, Lane 551, Sec. 5, Min-Tsu Rd., Yang-Mei, 326, Taoyuan, Taiwan
jskuo@cht.com.tw

Ying-Kuei YANG [2]
[2]Electrical Engineering Dept., National Taiwan University of Science and Technology, 43, Sec. 4, Keelung Rd., 106, Taipei, Taiwan
ykyang@mouse.ee.ntust.edu.tw

Abstract

A novel approach to automatically extracting paired transliterated-cognates from Web corpora is proposed in this paper. One of the most important issues addressed is that of taking multiple pronunciation characteristics into account. Terms from various languages may pronounce very differently. Incorporating the knowledge of word origin may improve the pronunciation accuracy of terms. The accuracy of generated phonetic information has an important impact on term transliteration and hence transliterated-term extraction. Transliterated-term extraction is a fundamental task in natural language processing to extract paired transliterated-terms in studying term transliteration. An experiment on transliterated-term extraction from two kinds of Web resources, Web pages and anchored texts, has been conducted and evaluated. The experimental results show that many transliterated-term pairs, which cannot be extracted using the approach only exploiting English pronunciation characteristics, have been successfully extracted using the proposed approach in this paper. By taking multiple language-specific pronunciation transformations into account may further improve the output of the transliterated-term extraction.

1. Introduction

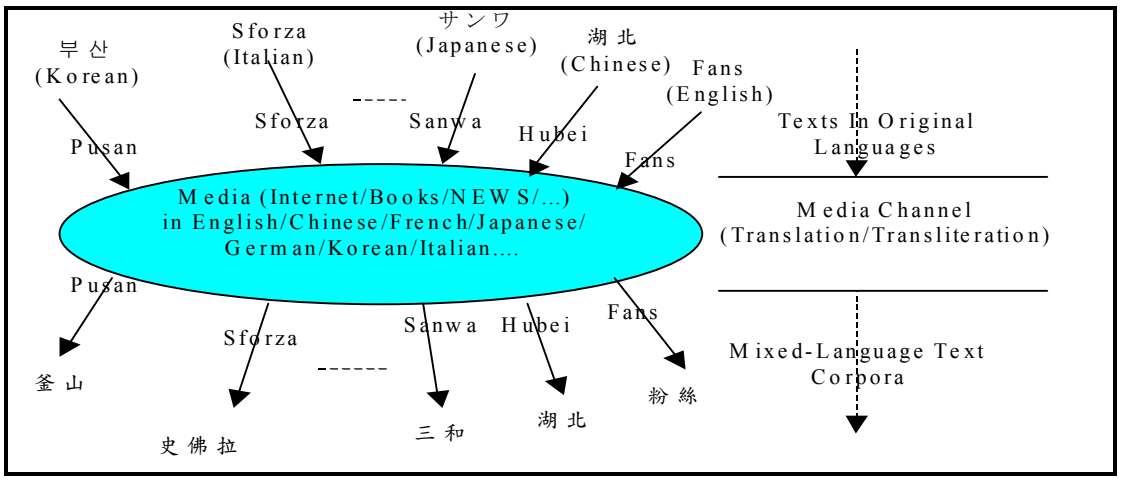

Figure 1. A conceptual flow for transliterating terms in multi-languages into Chinese.

Machine transliteration is one part of machine translation. Term transliteration in machine transliteration addresses the problem of converting terms in one language into their phonetic equivalents in the other language via spoken form. It is especially concerned with proper nouns, such as personal names, place names and organization names. Transliterated-term extraction, which is a fundamental task in studying term transliteration, has been focused on producing a large quantity of paired transliterated-cognates in order to observe various relations between cognate pairs. A transliteration lexicon, which is composed of many transliterated-term pairs, is important to the researches on term transliteration. Virga (2003) and Meng (2001) have explored relationships trained from a transliteration corpus collected manually to cross-language information retrieval and spoken document retrieval, respectively. However, it is time- and labor-consuming to prepare a transliteration

lexicon manually. It will be helpful if a transliteration lexicon can be compiled automatically.

English is one of commonly used languages around the world. Many terms are translated or transliterated into English first and then disseminated to the rest of the world. People speaking non-English languages might frequently use borrowed terms that were transliterated from English, terms that actually originated in languages other than English. As international communications increase, many foreign words may be imported from other languages through the translation or transliteration processes. Figure 1 depicts this situation.

Different transliterated terms may appear in a language when foreign words were transliterated directly from the original languages or indirectly from cognates, which were transliterated into languages other than the original languages. People may not distinguish between these multiple cognates, which originated from the same term in the source language. For example, Firenze in Italian was transliterated into "斐冷翠" (fei-leng-cui in Han-yu pinyin) in Chinese; however, Firenze was also transliterated into "Florence" in English and then "Florence" was transliterated into "弗羅倫斯" (fu-luo-lun-si) in Chinese. People may not know "斐冷翠" and "弗羅倫斯" have been used to refer to the same place if they do not study these terms carefully.

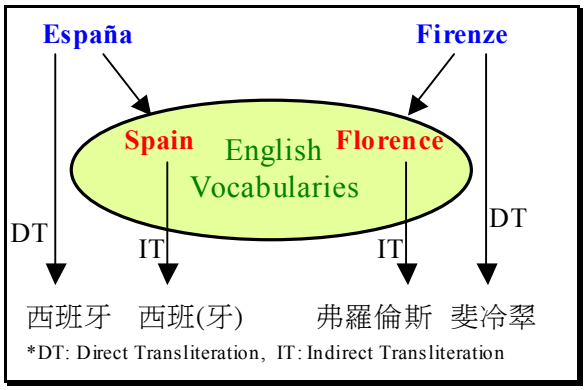

Figure 2. Chinese transliterated-terms transliterated directly from their original language terms or indirectly from their English cognates.

Another interesting example is the Chinese phonetic equivalents of Spain might be transliterated from Spanish or/and English. The problem is that "Spain" in English is syllabified into two syllables. These two syllables can be converted into "西班" (xi-ban) and cannot be mapped to "西班牙" (si-ban-ya), which has three syllables. The correct Chinese term is transliterated from Spanish, even though it is pronounced more like the English word. Figure 2 shows two examples of direct and indirect transliterations reflecting real cases. From these two examples, it implies that it will be helpful if source language terms originated from different languages should be pronounced using their native pronunciation system when extracting transliterated-term pairs.

Table 1. English terms transliterated from Mandarin-Chinese and its dialects.

Szechuan (四川)	Kung fu (功夫)	Guanxi (關係)	Feng shui (風水)	Tofu (豆腐)
Typhoon (颱風)	Qikong (氣功)	Taichi (太極)	Shaolin (少林)	Hong-Kong (香港)
Taiwan (台灣)	Whagwei (碗粿)	Gezaixi (歌仔戲)	Owanchian (蚵仔煎)	Jungtsu (粽子)

In addition to English, Chinese is also one of the commonly used languages around the world. Many English terms have been borrowed from Chinese and many Chinese terms have been imported

from English. Table 1[1] lists some English transliterated-terms and their Chinese counterparts. These terms have been prevalently used in English, especially when talking about Chinese issues. "Kung fu" and "Feng shui" are two typical examples.

English terms, which originated from Chinese, have been used commonly in daily conversations. Some of these terms may pronounce very different from that of native English terms. If these terms input to an English letter-to-sound system, which are trained from a corpus composed of large English terms, a sequence of incorrect phonemes may be obtained. Attention should be paid to these terms when dealing with English-Chinese transliterated-term extraction. For example, "草屯" ("Cao-tun") may pronounce as "/cao-tun/" (in Hanyu). If the English counterpart of this term, "Cao-tun", is not recognized that it is represented in pinyin and it may pronounce erroneously as "/go-tʌn/" using an English letter-to-sound system. Incorporating the knowledge of word origin may improve the pronunciation accuracy of terms (Llitjos, 2001). The accuracy of generated phonetic information has an important impact on term transliteration and hence transliterated-term extraction. In order to improve the performance of the transliterated-term extraction, taking the knowledge of word origin into consideration when dealing with the extraction of English-Chinese paired transliterated-terms is the most important focus in this paper.

Transliterated-term extraction using parallel corpora has been conducted (Lee, 2003). Generally speaking, parallel corpora are smaller in scale and less versatile in coverage as compared to non-parallel corpora. Transliterated-term extraction using non-parallel corpora has also been conducted (Kuo, 2003). Kuo (2003) successfully extracted transliterated-term pairs from Web pages collected by a software spider with the aid of confusion matrices generated by a speech recognition system. Examining those cases, which failed to be extracted transliterated-term pairs, it showed that it is difficult to generate phonetic information from English terms, which were transliterated from Chinese and may not follow the western style pronunciation rules, correctly. "Yungan", which is such an English term, can be segmented into "Yun-gan" (雲岡) or "Yung-an" (永安).

In this paper, a novel approach, which uses multiple pronunciation transformations for terms originated from different languages, is proposed for transliterated-term extraction from Web corpora. Different pronunciation methods are used here to generate phonetic information when dealing with terms from various languages. If a term can be transliterated from Chinese into English, it may be represented in pinyin. However, various pinyin representations have been proposed and used to represent a Chinese term. A procedure is used to automatically detect which pinyin system is used to represent the term in order to generate correctly phonetic information in term extraction.

The remainder of the paper is organized as follows: Section 2 describes how English-Chinese transliterated term pairs can be extracted automatically using multiple pronunciation transformations. Experimental results obtained using Web corpora are presented in section 3. Section 4 provides an extensive discussion of transliterated-term extraction. Conclusions are drawn in section 5.

2. The Proposed Approach

An approach, which uses different pronunciation transformations for terms originated from different languages, is described in this section. English is the source language and Chinese is the target language referred to in this paper.

There are two pronunciation approaches used to process English terms, which may be generated natively or borrowed from other languages. MBRDICO (Pagel, 1998) is the English letter-to-sound system used to convert English strings into phonemes, and a Chinese pinyin detection algorithm, which is used first to segment this term into syllables using left-to-right longest matching algorithm and then to detect which pinyin scheme is used to represent a Chinese term in English. When a term is transformed into phonemes using an English letter-to-sound systems or a Chinese pinyin detection algorithm, then the degree of similarity between paired terms can be calculated.

[1] http://www.yellowbridge.com/language/chineseloan.html

Text-based syllabification algorithm (TSA) is used to handle Chinese terms encoded according to the pinyin rules and phoneme-based syllabification algorithm (PSA) is used to process English terms. TSA is a syllabification algorithm operated directly on text units represented in pinyin symbols in this paper. PSA is a syllabification algorithm operated on phonemes rendered from texts. Traditionally, an English syllable is composed of an initial consonant cluster followed by a vowel and with the option of a final consonant cluster. In order to convert English syllables to Chinese ones, the final consonant cluster is appended only when it is a nasal. The other consonants in the final consonant cluster are then segmented into isolated consonants. Such a syllable may be viewed as the basic pronunciation unit in transliterated-term extraction. By combing both TSA and PSA in this approach, we can obtain basic pronunciation units of terms, which are used to determine the degree of similarity between paired terms, in term extraction process.

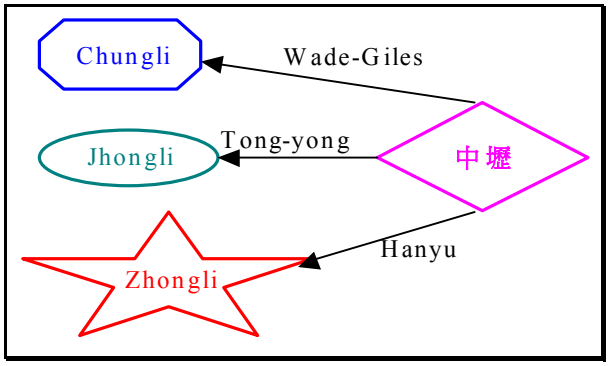

Figure 3. Multiple English terms have been transliterated from the same Chinese term using different pinyin systems.

Multiple pinyin systems have been used to romanize Mandarin-Chinese terms. For example, Hanyu pinyin has been used uniformly in mainland China; on the other hand, Wade-Giles, Tong-yong and Hanyu have been used in Taiwan in different situations. For example, "中壢" has been romanized into "Chungli", "Jhongli", and "Zhongli" using Wade-Giles, Tong-yong, and Hanyu, respectively. People may think that these three terms refer to three different places; actually, they all refer to the same town in northern Taiwan. Figure 3 depicts this phenomenon that multiple English terms have been transliterated from the same Chinese term using different pinyin systems. Actually, more pinyin systems in addition to the romanization systems mentioned above have been used[2]. In this paper, we only focus on how to disambiguate the results produced by these three romanization approaches and to determine the possible pinyin system. This disambiguation procedure is also applicable to the cases of taking other pinyin systems into consideration.

The proposed pinyin detection algorithm is described as follows:

1. Segmenting the input term into pinyin tokens by using left-to-right longest matching algorithm.
2. Tagging each pinyin token with one or more tags. Each tag represents a pinyin scheme and is assigned a score in order to determine the possible pinyin scheme. Each pinyin scheme has the same score.
3. Selecting the pinyin scheme with the highest score by accumulating and sorting the scores assigned to pinyin tokens in descending order.

A segmentation procedure used in the pinyin detection algorithm, which exploits left-to-right longest matching algorithm, is to decompose English terms into segments. By trying to match the longest string from left to right using the basic syllables specified in pinyin systems, tokens used by different pinyin systems might be matched and found. When a term composed of one or more such tokens, it is possible that this English term was transliterated from Chinese. For example, "Tienanmen" can be

[2] http://pinyin.info

segmented into "Tie-nan-men" (鐵南門) or "Tien-an-men" (天安門). Only the first term can be obtained according to the western pronunciation rules (Jurafsky, 2000), which combines a leading consonant cluster, a vowel and a following consonant cluster together, using an English letter-to-sound system and syllabifying English syllables. Therefore, in the case of "Tienanmen", "Tien-an-men" is selected to generate phonetic information in extracting terms transliterated from Chinese using pinyin detection algorithm.

After finishing segmentation, a disambiguation procedure is used to determine the possible pinyin system used to represent the term. An English term, which was transliterated from Chinese, is possibly composed of syllables in different pinyin systems. For example, "Chang-hwa" of the Bank of Chung-hwa (彰化銀行) can be decomposed into two syllables, namely, "Chang" and "Hwa". "Chang" may be encoded in Wade-Giles, Yale, Tong-yong and Hanyu pinyin systems. If "Chang" is encoded in Wade-Giles, it is pronounced as /zhang/ (in Hanyu pinyin; or "彰"); however, if "Chang" is encoded in Yale, Tong-yong and Hanyu, then it is sounded as /chang/ (in Hanyu pinyin; or "昌"). The problem is "Hwa" is detected only in Yale, Therefore "Chang-hwa" is supposed to be encoded in Yale. However, according to the Chinese pronunciation rules, "彰" should be read as /zhang/. It means that "Chang-hwa" might be encoded using different pinyin systems. This kind of romanization problems will worsen the determination of the possible pinyin system. In order to take these cases into consideration, combinations of tokens in different pinyins are also taking into account in transliterated-term extraction. The confusion matrices generated or trained can be used to alleviate this problem when dealing with the extraction of paired cognates (Kuo, 2003).

The steps, which used to extract paired transliterated cognates, can be referred to (Kuo, 2003) and are described briefly here. Generally, a sentence separated by a pronunciation mark is selected from the training text corpora. Then the candidates of the cognates in the target language are obtained from the contexts of the located source-language string in the selected sentence. Both strings in the source language and target language are converted into phonemes of the same representation in order to calculate the degree of similarity between these two terms. English phonemes are then syllabified into consonant-vowel pairs. The converted English syllables are transformed into Chinese syllables by using basic English-to-Chinese phoneme conversion with hand-coded rules initially when ASR (Automated Speech Recognition)-generated confusion matrices are used or the phoneme conversion obtained from the extracted paired transliterated cognates. Then the similarity degree between syllables is calculated, and a pair of transliterated terms can be extracted, depending on whether the similarity degree is larger than a predefined threshold or not.

When the above algorithm, which uses an English letter-to-sound transformation first, fails to extract paired transliterated terms. A procedure is used to try to exploit Chinese pronunciation characteristics to extract paired transliterated terms. A left-to-right longest matching algorithm is used to syllabify an English term. If all the syllables of the term can be found using multiple pinyin systems, then this term may be borrowed from Chinese. This term is then segmented into phonemes. The syllables generated from pinyin segmentation are used to calculate the degree of similarity between two terms. The overall performance achieved by using the proposed approach will be better than that achieved by only using an English letter-to-sound system. This is because the proposed approach in this paper used only when the approach using an English letter-to-sound system failed to extract paired terms.

3. Experimental Results

Two kinds of Web resources are used in the experiment. An English-Chinese text corpus of 500MB in 15,822,984 pages, which was collected from the Internet using a web spider and was converted to plain text, was used as a testing set. This corpus is called SET1. From SET1, 80,094 qualifying sentences that occupied 5MB were extracted. A qualifying sentence was a sentence composed of at least one English string. This corpus was used in Kuo (2003), which successfully used ASR (Automated speech recognition)-generated confusion matrices to extract transliterated-term pairs. The other Web resource, anchored texts, has been successfully used to extract multilingual translation

terms (Lu, 2002) and has not been applied to transliterated-term extraction so far.

3.1 Transliterated-term Extraction from Web Pages

The results obtained by applying CASCM (Chinese ASR-generated syllable-based confusion matrix), CAPCM (Chinese ASR-generated phoneme-based confusion matrix) and both CASCM and CAPCM (CACM for abbreviation) are depicted in Table 2. Generally speaking, phonemes have finer-grained controls on pronunciation than that by syllables do initially, and hence, the CAPCM extracted more pairs than that produced by CASCM. Those pairs produced by using CACM got the best results in generating term pairs. All the distinct qualifying term pairs (DQTP) reported in this paper were verified manually.

Table 2. The results produced by applying ASR-generated confusion matrices.

	CASCM	CAPCM	CACM
DQTP	1,971	3,353	3,831

The collection of extracted term pairs produced by using ASR-generated confusion matrices was a "parallel" corpus and reflected the real cases of term transliteration. The approaches using the ASR-generated confusion matrices and taking multiple pronunciation characteristics into account were called CASCM-MP and CAPCM-MP, respectively. If both syllable and phoneme confusion matrices are used, the approach is called CACM-MP. The results obtained by CASCM-MP, CAPCM-MP and CACM-MP are displayed in Table 3.

Table 3. The results produced by applying cross-linguistic syllable-phoneme conversion and taking multiple pronunciation characteristics into consideration.

	CASCM-MP	CAPCM-MP	CACM-MP
DQTP	2,374	3,753	4,373

The performances achieved by CASCM-MP, CAPCM-MP and CACM-MP boost about 20.45%, 11.93% and 14.15% with respect to CASCM, CAPCM and CACM, respectively. The overall performance of each approach is shown graphically in Figure 4. From Figure 4, it is found that the performance achieved by using ASR-generated confusion matrices (AGCM) and taking multiple pronunciations into account was better than that produced by only using AGCM.

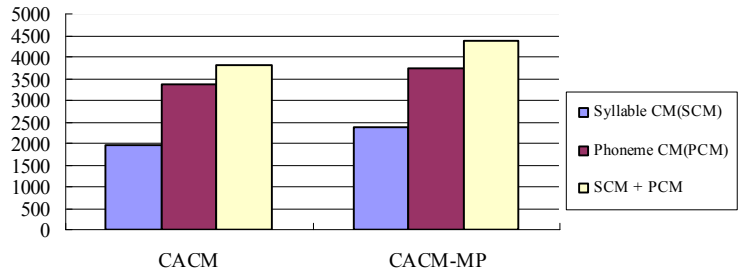

Figure 4. The overall performances achieved by using two different methods.

3.2 Transliterated-term Extraction from Anchored Texts

The Web is growing at a fast pace. It is a rich information source for researches. Many data of

different types, such as texts, pictures, animations etc. are distributed through the Internet. The Web pages are connected through hyper-links and are weaved into a vast network. An iterative method was proposed to identify hubs and authorities in this hyper-linked environment and to then refine search topics by using the information of hub pages and authoritative pages (Kleinberg, 1998). Web pages have a mechanism similar to reader citations in the form of links that provide very useful information in different areas. Hyperlink analysis has been widely used in information retrieval research and commercial systems (Brin, 1998). It has also applied to statistical term translation (Lu, 2002) and achieves good performances. However, information obtained through hyper-link analysis has not been exploited in transliterated-term extraction so far. Hyperlinks provides a mechanism to associate multiple cognates together. Through link analysis and transliterated-term extraction, many possible transliteration term pairs can be obtained.

1,980,816 web pages were collected using a web spider. Among these pages, 109,416 anchored text pairs were extracted from the collected corpus, which was used as the testing corpus. This corpus was called SET2. Terms were extracted from these web pages, the results generated by using the original AGCM and the proposed approach in this paper are depicted in Table 4.

Table 4. Results produced by applying CACM and CACM-MP to anchored texts, respectively.

	CACM	CACM-MP
DQTP	1,515	1,816

The performance achieved by taking multiple pronunciations into account is outperformed the algorithm without taking Chinese pronunciation transformations into consideration by 19.87%.

4. Discussion

A large quantity of transliterated-term pairs were extracted successfully using the proposed approach in this experiment. Using ASR-generated confusion matrices as a basis initially provided the initial phoneme conversion for terms in two completely different languages. By combining the ASR-generated confusion matrices and multiple pronunciations together, more paired transliterated-cognates from Web corpora was extracted.

Some examples of qualifying transliterated-term pairs, which were extracted using the proposed approach, are shown in Table 5. One important point worth of noting is that some transliterated-term pairs, which were not found using the method in Kuo (2003), were extracted using the approach proposed in this paper. Many terms are out-of-vocabulary in existing dictionaries.

Table 5. Newly extracted transliterated-term pairs, most of which were transliterated from Chinese, using the proposed approach

Tzuchi (慈濟)	Lanyu (蘭嶼)	Xiangqi (象棋)	Luantan (亂彈)	Yushan (玉山)	Hwalian (花蓮)
Chienkuo (建國)	Zhaoming (昭明)	Ouyang (歐陽)	Pingju (平劇)	Hsitou (溪頭)	Hualien (花蓮)
Siqing (思清)	Lianhe (聯合)	Xiaoniu (小牛)	Waishuangxi (外雙溪)	Xinjiang (新疆)	Kwanghwa (光華)
Kunqu (崑曲)	Chitou (溪頭)	Kwaninn (觀音)	Xiaoao (笑傲)	Jinyong (金庸)	Guanghua (光華)

Another worthy point is that two paired transliterated-cognates, "Hwalian" and "Hualien", are shown in Table 5. These terms, which are in different pinyin systems and have been used on Web pages, were extracted by the proposed approach. From the experimental results, it also revealed that several pinyin systems, such as Hanyu, Tong-yong and Wade-Giles, have been used in Taiwan.

Collecting the terms, which were generated by incorporating the knowledge of Chinese word origin into consideration, may help to translate articles written in English but talking about Chinese issues into Chinese. Kwok (2003) proposed a system to back-transliterate place names. One of the important steps in back-transliteration is to look-up a bilingual place name list. The automatically extracted transliterated-terms in this paper can be used to mitigate the problem of collecting bilingual proper noun lists in these applications. It will be also helpful if we can exploit the pronunciation characteristics of multiple languages such as Mandarin-Chinese, Taiwanese, Cantonese and Japanese when dealing with transliterated-term extraction in the future.

5. Conclusions

In this paper, a novel approach, which incorporates different pronunciation characteristics into transliterated-term extraction when dealing with English-Chinese transliterated-term extraction, has been proposed. Many transliterated-term pairs, which are composed of English terms imported from Chinese, were successfully extracted from Web pages. These terms cannot be extracted in the approach only using an English letter-to-sound system. The overall performance achieved by exploiting the English and Chinese pronunciation characteristics are better than that achieved only using English pronunciations. Because English vocabularies actually consists of many terms imported from other languages, incorporating the pronunciation characteristics of Japanese, Korean and Cantonese in addition to the Chinese pinyin system into term extraction can improve the output of transliterated-term extraction further.

6. References

Brin S. and L. Page 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine, In *Proceedings of 7th International World Wide Web Conference,* pp. 107-117.

Jurafsky D. and J. H. Martin 2000. Speech and Language Processing, pp. 102-120, Prentice-Hall, New Jersey.

Kleinberg, 1998, Authoritative Sources in a Hyperlinked Environment, In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*, pp. 14-20.

Knight K. and J. Graehl 1998. Machine Transliteration, Computational Linguistics, Vol. 24, No. 4, pp.599-612.

Kuo J. S. and Y. K. Yang 2003. Automatic Transliterated-term Extraction Using Confusion Matrices from Non-parallel Corpora, In *Proceedings of ROCLING XV Computational Linguistics Conference*, pp.17-32.

Kwok K. L. and Q. Deng 2003. GeoName: a system for back-transliterating pinyin place names, In *Proceedings of Analysis Geographic References Workshop of HLT-NAACL*, pp. 26-30.

Lee C. J. and J. S. Chang, Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model, In *Proceedings of HLT-NAACL*, Edmonton, Canada, 2003, 96-103.

Llitjos A. F and A. Black 2001 Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names, In *Eurospeech'2001*, Vol. 3, pp. 1919-1922.

Lu W. H., L. F. Chien, and H. J Lee. 2002, Translation of Web Queries Using Anchor Text Mining, ACM transaction on Asia Language and Information Processing, Volume 1, Issue 2, pp. 159- 172.

Meng H., W. K. Lo, B. Chen and K. Tang 2001. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval, in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Trento, Italy.

Pagel V., K. Lenzo, and A. Black 1998. Letter to Sound Rules for Accented Lexicon Compression, In *Proceedings of ICSLP*, pp. 2015-2020.

Virga P. and S. Khudanpur 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval, In *Proceedings of ACL Workshop on Multilingual and Mixed Language Named Entity Recognition*, pp. 57-64.