

# A Phonetic Similarity Model for Automatic Extraction of Transliteration Pairs

JIN-SHEA KUO

National Taiwan University of Science and Technology

HAIZHOU LI

Institute for Infocomm Research

AND

YING-KUEI YANG

National Taiwan University of Science and Technology

---

This article proposes an approach for the automatic extraction of transliteration pairs from Chinese Web corpora. In this approach, we formulate the machine transliteration process using a syllable-based phonetic similarity model which consists of phonetic confusion matrices and a Chinese character  $n$ -gram language model. With the phonetic similarity model, the extraction of transliteration pairs becomes a two-step process of *recognition followed by validation*: First, in the *recognition* process, we identify the most probable transliteration in the  $k$ -neighborhood of a recognized English word. Then, in the *validation* process, we qualify the transliteration pair candidates with a hypothesis test. We carry out an analytical study on the statistics of several key factors in English-Chinese transliteration to help formulate phonetic similarity modeling. We then conduct both supervised and unsupervised learning of a phonetic similarity model on a development database. The experimental results validate the effectiveness of the phonetic similarity model by achieving an  $F$ -measure of 0.739 in supervised learning. The unsupervised learning approach works almost as well as the supervised one, thus allowing us to deploy automatic extraction of transliteration pairs in the Web space.

Categories and Subject Descriptors: H.2.4 [Database Management]: Systems – *Textual databases*; H.2.8 [Database Management]: Database Applications – *Data mining*; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Machine translation, machine transliteration, extraction of transliteration pairs, phonetic similarity modeling, phonetic confusion probability

## ACM File Format:

KUO, J.-S., LI, H., AND YANG, Y.-K. 2007. A phonetic similarity model for automatic extraction of transliteration pairs. *ACM Trans. Asian Language Inform. Process.*, 6, 2, Article 6 (September 2007), 24 pages. DOI = 110.1145/1282080.1282081 <http://doi.acm.org/10.1145/1282080.1282081>

---

## 1. INTRODUCTION

Machine transliteration plays an important role in the study of natural language processing on topics such as named entity recognition (NER), cross-language information retrieval (CLIR), question answering (QA) and machine translation (MT). It is a process

---

Authors' addresses: Jin-Shea Kuo, Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, 106, Taiwan, email: [d8807302@gmail.com](mailto:d8807302@gmail.com) ; Haizhou Li, Institute for Infocomm Research, Singapore 119613, email: [hli@i2r.a-star.edu.sg](mailto:hli@i2r.a-star.edu.sg) ; Ying-Kuei Yang, Department of Electrical Engineering, National Taiwan University of Science and Technology, Taiwan, email: [ykyang@mouse.ee.ntust.edu](mailto:ykyang@mouse.ee.ntust.edu)

Permission to make digital/hard copy of part of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date of appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Permission may be requested from the Publications Dept., ACM, Inc., 2 Penn Plaza, New York, NY 11201-0701, USA, fax: +1 (212) 869-0481, [permission@acm.org](mailto:permission@acm.org)  
© 2007 ACM 1530-0226/07/0600-ART 6 \$5.00 DOI 110.1145/1282080.1282081 <http://doi.acm.org/10.1145/1282080.1282081>

of translating a word in one language into another by preserving its pronunciation in the original language, otherwise known as *translation-by-sound*. In most cases, a precompiled bilingual lexicon serves as the training corpus for building a transliteration system, as shown in studies of NER [Chen and Lee 1996]; CLIR [Lee and Choi 1998; Qu et al. 2003; Virga and Khudanpur 2003]; machine transliteration [Knight and Graehl 1998; Wan and Verspoor 1998; Al-Onaizan and Knight 2002; Lin and Chen 2002]; and cross-language spoken document retrieval [Meng et al. 2001]. In the literature, most of the bilingual transliteration lexicons are small in scale and/or compiled manually. Considering the amount of potential transliteration pairs in the open domain, it is almost impossible to construct a comprehensive transliteration lexicon. To mitigate this problem, an automatic approach for extracting transliteration pairs from Web corpora could serve as a good solution.

In general, studies of transliteration fall into two categories: transliteration modeling (TM) and extraction of transliteration pairs (EX). There have been many reports on transliteration modeling for different language pairs, such as English-Chinese [Chen et al. 2003; Gao et al. 2004; Li et al. 2004]; English-Japanese [Knight and Graehl 1998; Qu et al. 2003]; English-Korean [Lee and Choi 1998; Jung et al. 2000; Kang and Choi 2000; Kang and Kim 2000; Oh and Choi 2002]; English-Arabic [Al-Onaizan and Knight 2002]; and French-Japanese [Tsuji et al. 2002]. In comparison, there are fewer works on the extraction of transliteration pairs from a corpus [Brill et al. 2001; Lee and Chang 2003; Kuo and Yang 2004a; 2004b; Lin et al. 2004; Kuo and Yang 2005]. The former approach models transliteration rules with a generative model that is trained from a large bilingual transliteration lexicon, with the objective of translating unknown words on-the-fly in the open domain. The latter approach uses a data-driven method to extract real-life transliteration pairs from a corpus, in an effort to construct a large, up-to-date transliteration lexicon from live text sources. In this article we approach the extraction problem through transliteration modeling.

Each language has its unique phonetic system, which consists of a phoneme inventory, phonic rules, and prosodic rules. Ambiguity arises when we attempt to map sounds across phonetic systems. In manual translation, the transliteration ambiguity can be alleviated if translators observe common rules which follow the *translation-by-sound* principle. For example, translation professionals in mainland China follow guidelines recommended by the *Xinhua News Agency* [1992]. Words transliterated by closely observing common guidelines are referred to as *regular* transliterations. However, in Web publishing, translators in different countries and regions may not observe the same guidelines. They sometimes skew the transliterations in different ways to create special flavors or to introduce semantic implications, also known as wordplay, resulting in *casual* transliterations. In other words, *casual* transliterations are those transliterations with multiple Chinese phonetic equivalents for an English word. For example, “Disney” and “Honeywell” are transliterated into Chinese in many ways as shown in Table I. In this article we use *Hanyu Pinyin*,<sup>1</sup> the official Chinese Romanization system of China, to denote Chinese words. The *Hanyu Pinyin* syllables are given in upper-case to differentiate from English words and syllables.

The rapidly growing Web is one of the largest distributed databases in the world. In this article, we propose an approach that extracts transliteration pairs from the Web. Instead of relying on a parallel corpus, we exploit techniques to acquire transliteration

---

<sup>1</sup> <http://www.romanization.com>

Table I. *Casual* Transliterations on the Web

Disney	迪士尼 /DI-SHI-NI/	狄士尼 /DI-SHI-NI/	迪斯耐 /DI-SI-NAI/	狄斯耐 /DI-SI-NAI/
Honeywell	漢尼威 /HAN-NI-WEI/	霍尼威 /HUO-NI-WEI/	霍尼偉 /HUO-NI-WEI/	霍尼韋爾 /HUO-NI-WEI-ER/

pairs from a nonparallel corpus, such as predominantly Chinese text mixed with English words or Chinese anchor text with English counterparts, which are obtained from a hyperlink analysis and have been used for mining multilingual translations [Lu et al. 2002]. In this way, we are able to tap the vast amount of Web data since a nonparallel corpus is generally more accessible than a parallel one.

The remainder of this article is organized as follows: In Section 2, we discuss the basics of English-Chinese transliteration. We refer to all Latin-script words as English words, although some of them originate from Romanization of Chinese. In Section 3, we formulate the extraction of transliteration pairs using the phonetic similarity model. In Section 4, we conduct several experiments and report the results. In Section 5, we discuss related and future work. We conclude with Section 6.

## 2. BASICS OF CHINESE TRANSLITERATION

The idea of transliteration is to preserve the sound of the original language, or to follow the *translation-by-sound* principle. Research on automatic transliteration has reported promising results for *regular* transliteration [Li et al. 2004], where transliterations observe rigid guidelines. However, in *casual* transliteration, the generative models fail in predicting the transliteration most of the time. In either case, the source English word serves as the basis for any transliterations that people can come up with under the *translation-by-sound* principle. In this section, we introduce the transliteration principle and provide an analytical study on a development database.

### 2.1 Transliteration Principle

To make good use of the English phonetic information, we have to establish a comparison between sounds in two languages. Note that English and Chinese have different syllable structures. Chinese is a syllabic language, with each Chinese character pronounced as a syllable in either consonant-vowel (CV) or consonant-vowel-nasal (CVN) structure. A Chinese word consists of a sequence of characters, or phonetically a sequence of syllables. As such, in the task of English-Chinese transliteration, it would be a natural choice to syllabify an English word by converting the phoneme sequence into a sequence of Chinese-like syllables so as to predict how its Chinese transliteration would be like.

There have been several effective algorithms for the syllabification of English words for transliteration. Typical syllabification algorithms first convert English graphemes to phonemes, referred to as G2P or letter-to-sound, then syllabify the phoneme sequence into a syllable sequence. Techniques for G2P are studied intensively, driven by the need in text-to-speech research. Successful techniques include heuristic rules [Wan and Verspoor 1998] and machine-learning methods [Pagel et al. 1998; Galescu and Allen 2001]. We adopt a syllabification approach similar to the one suggested in Jurafsky and Martin [2000].

Typically, an English syllable comprises of consonant-vowel, consonant-vowel-nasal, and consonant-vowel-consonant structures, denoted as CV, CVN, and CVC, whereas Chinese only has CV and CV-nasal (CVN) syllable structures. English phonetic rule

allows consonant clusters (CC) such as /-sk/ in “risk” and /-str-/ in “street”; Chinese, on the other hand, follows rigid CV or CVN rules where C is an optional single consonant. A Chinese syllable is also seen to have an initial-final structure where the initial is an optional single consonant while the final is either a vowel (V) or a vowel-nasal (VN) nucleus. When transliterating an English CV or CVN syllable into Chinese, we typically preserve the English syllable structure. To deal with consonant clusters and CVC structures in English, two processes are introduced in the syllabification. First, we decompose a consonant cluster into a sequence of individual consonants (ICs). Hereafter, we denote IC as a single consonant for clarity. Second, a CVC syllable in English is split into a CV and an IC.

To establish a syllable mapping between English and Chinese, we first syllabify the English words, and then apply syllable structure conversion rules. In this article we take the strategy of syllabification followed by a syllable conversion. For example, following the approach in Jurafsky and Martin [2000], the English word “Boulder” is converted into phonemes /b o l d ə/, which can be further syllabified into /bo-l-də/ or /Boul-der/, where syllables are separated by hyphens. The syllable conversion rules try to resolve the difference between English and Chinese phonetic systems by segmenting and finding the closest equivalents of syllable structures. They can be summarized as follows:

$$\begin{aligned} CVC (\text{English}) &\rightarrow CV \quad IC (\text{English}) \\ CC (\text{English}) &\rightarrow IC_1, IC_2, \dots, IC_I (\text{English}) \end{aligned} \quad (1)$$

An IC in English can either be augmented with a nucleus vowel to map into a Chinese CV syllable or be elided. The former is known as phonetic insertion, while the latter is known as phonetic elision. The elision can be formulated as a mapping to a null syllable  $\phi$ . They are given in rule (2). To summarize, we list the mapping rules in Table II.

$$\begin{aligned} IC (\text{English}) &\rightarrow C \quad \text{Nucleus (a Chinese CV syllable)} \\ IC (\text{English}) &\rightarrow \phi \quad (\text{elided in Chinese}) \end{aligned} \quad (2)$$

Let’s revisit the case of “Boulder,” the syllable sequence is now processed as /bo-l-də/  $\rightarrow$  /bo-l-də/ by rule (1) and /bo-l-də/  $\rightarrow$  /BO-ER-DE/ (波爾德) or /bo-l-də/  $\rightarrow$  /BO-DE/ (波德) by rule (2).

In this article, a syllable is considered a basic pronunciation unit in transliteration. The syllabification of an English word results in a sequence of Chinese-like syllables.

Table II. English Syllable Structures and Their Chinese Equivalents (the square brackets indicate multiple items)

English syllable	Segmented English syllable	Chinese-like equivalent	Chinese Initial-Final
CC	[IC]*	[CV or $\phi$ ]**	[C-V or $\phi$ ]
CVC	CV, IC*	CV, CV or $\phi$ **	C-V, C-V or $\phi$
CV	C-V	CV	C-V
CVN	C-VN	CVN	C-VN

\* Rule (1) is applicable. \*\* Rule (2) is applicable

The next question is how to compare English syllables with Chinese syllables; in other words, how to measure the similarity between syllables in two different phonetic systems. The formulation of PSM in this article will provide a solution.

Note that *casual* transliteration can be characterized in different categories, such as lexical variation, phonetic variation [Jurafsky and Martin, 2000], and wordplay, and so on. Chinese is an ideographical writing system that supports multiple dialectal/accented spoken languages. The dialectal/accented spoken Chinese and the choice of preferred characters may result in lexical variations for the same English word. For example, the English name “Bush” has Chinese transliterations of “布什/BU-SHI/” and “布希/BU-XI/” in China and Taiwan, respectively. On the other hand, transliteration is an artistic human endeavor. The creative process also leads to phonetic variations as a result of the combination of different syllabification, phonetic mapping, insertion and elision strategies in transliteration. For example, “Disney” has phonetic variants such as 迪士尼 /DI-SHI-NI/ and 狄斯耐 /DI-SI-NAI/ due to its English phonetic variations. As a result, the most prominent property of a *casual* transliteration is that there are multiple commonly used Chinese variants for an English word. In this article we propose a statistical framework for extracting transliterations, including both *regular* and *casual* cases.

## 2.2 Statistical Analysis

The Web is growing at a fast pace and providing a rich and live information source for researches. The proposed transliteration pair extraction method allows us to tap the Web for transliteration lexicons. To conduct our study, we first construct a development corpus from the Web. Initially, a Chinese predominant text corpus is collected using a Web spider. We start with submitting a list of 10 initial Web addresses (also called Universal Resource Locators, URLs). We then obtain new URLs from the returned Web pages by parsing the page contents. In this way, the new URLs are discovered and resubmitted iteratively. We aggregate a corpus of about 500MB Web pages in diverse topics, all normalized to plain text, referred to as SET1. We extract qualified sentences from SET1 for our transliteration study. A qualified sentence has at least one English word *EW*. With this criterion, 80,094 qualified sentences are extracted automatically.

Table III. Elision Rates for the Top Six Isolated Syllables in SET1

Individual consonant	Elision rate (%) in all cases	Elision rate (%) by position in all cases		Elision rate (%) by position in elided cases	
		Middle	End	Middle	End
/r/	61.1%	66.9%	33.7%	90.8%	9.2%
/l/	38.2%	47.6%	25.9%	72.3%	27.7%
/d/	29.2%	27.2%	31.9%	28.8%	71.2%
/t/	28.1%	43.5%	21.4%	49.0%	51.0%
/z/	9.8%	7.5%	10.2%	11.1%	88.8%
/s/	5.7%	5.7%	8.0%	45.6%	54.3%

Each qualified sentence is manually validated based on the following transliteration criteria: (i) if an *EW* is partly translated phonetically and partly translated semantically, only the phonetic transliteration constituent is extracted to form a transliteration pair; (ii) elision of English sound is accepted; (iii) multiple English-Chinese transliteration pairs can appear in one sentence; (iv) an *EW* can have multiple valid Chinese transliterations and vice versa. The validation process results in 8,898 qualified transliteration pairs, also referred to as distinct qualified transliteration pairs or DQTPs.

Before formulating the transliteration, it is beneficial to look into the statistics of three key factors that strongly impact the transliteration process, namely, phonetic confusion, elision, and insertion. Phonetic elision and insertion have to do with the individual consonants (ICs) resulting from English syllabification, while phonetic confusion arises due to mismatches between the two phonetic systems. As phonetic confusion will be studied in detail in the following sections, we only briefly study elision and insertion here.

Out of the 8,898 DQTPs in SET1, about 30% of them are found to have at least one elided IC. It is interesting to note that elisions take place much more often in the middle of a word than at the end of a word. We break down the elision rates of the top six ICs as shown in Table III in terms of total elision rate and elision rate by position. Each IC has its own elision pattern. For example, /r/ is elided much more often in the middle than at the end of a word, while /z/ behaves otherwise. We also observe the elision rate with respect to word length. The average numbers of syllables of the elided and non-elided English word are 3.8 and 2.7, respectively. It reveals that the longer the word is, the more likely that its ICs will be elided. Kuo and Yang [2005] have shown that the statistics of elision patterns are informative in helping to improve the extraction performance.

As in Table II, a Chinese syllable structure does not include an IC. Therefore, an IC in English could be either dropped by elision, blended with a neighboring consonant, or augmented with a nucleus vowel, known as insertion. In English, we can combine two or more ICs to form a *blend* such as /tr/, /dr/ and /pt/ which do not have Chinese equivalents. On the other hand, one can also insert a nucleus vowel into an IC to make it a Chinese-like syllable. For example, /k-/→/Ke-/ and /t-/→/Te-/ by inserting the nucleus vowel /e/. Although phonetic insertion adds a nucleus vowel to an IC, it does not increase the number of total Chinese-like syllables. Therefore, if we syllabify an *EW* into  $M$  syllables, then its Chinese transliteration would have  $N \leq M$  syllables. In the case of no elision, we have  $N = M$ .

Chinese is an ideographic language. Each graphic character carries meanings. The choice of characters in a transliteration is decided from the cultural and aesthetic point of view, which implicitly follows the rule of avoiding offensive and irritating words and sounds. The evidence that demonstrates this rule is that, out of thousands of common Chinese characters, only 374 are used in *regular* transliterations [Xinhua 1992], and 1,210 of them are used for SET1, that include both *regular* and *casual* transliterations. The statistics of character usage in Chinese transliteration can help in the search for transliteration candidates. We propose using an  $n$ -gram language model to capture the statistics of character usage. A similar idea was reported earlier in Xiao et al. [2002]. To illustrate, we list the top ten unigrams and bigrams of characters in our development database in Table IV.

As the phonetic systems of Chinese and English are different, approximate matching takes place very often. This leads to multiple transliteration variants for an English word and vice versa. In Table V, we list four English words and their transliteration variants, with their counts from our development database SET1.

Table IV. The Top Ten Character Unigrams and Bigrams

Unigrams		Bi-grams	
斯 /SI/ (5.001%)	拉 /LA/ (1.937%)	斯特 /SI-TE/ (0.556%)	拉斯 /LA-SI/ (0.231%)
克 /KE/ (3.219%)	卡 /KA/ (1.378%)	克斯 /KE-SI/ (0.414%)	斯坦 /SI-TAN/ (0.209%)
爾 /ER/ (3.062%)	利 /LI/ (1.290%)	斯基 /SI-JI/ (0.356%)	拉克 /LA-KE/ (0.193%)
特 /TE/ (2.477%)	羅 /LUO/ (1.198%)	爾斯 /ER-SI/ (0.337%)	威爾 /WEI-ER/ (0.191%)
德 /DE/ (2.131%)	格 /GE/ (1.191%)	里斯 /LI-SI/ (0.251%)	羅斯 /LUO-SI/ (0.181%)

Table V. Transliteration Variants and Their Counts

English Words	Transliteration variants (count)			
Robert	羅伯 (31) /LUO-BO/	羅勃特 (11) /LUO-BO-TE/	羅伯特 (4) /LUO-BO-TE/	勞伯特 (1) /LAO-BO-TE/
Charles	查爾斯 (12) /CHA-ER-SI/	查理斯 (2) (CHA-LI-SI)	察爾斯 (2) /CHA-ER-SI/	查釐士 (1) /CHA-LI-SH/
Michael	麥可 (47) /MSI-KE/	麥克 (10) /MAI-KE/	邁可 (10) /MAI-KE/	邁克 (5) /MAI-KE/
Richard	理查 (34) /LI-CHA/	李察 (14) /LI-CHA/	睿哲 (3) /RUI-ZHE/	李查德 (1) /LI-CHA-DE/

In the 8,898 DQTPs from SET1, it is observed that 35.88% of them have multiple Chinese transliterations for an English word. That is, a large number of English words have been transliterated in different ways. The statistics confirm the fact that *casual* transliterations are not unusual in Web publishing. The observation also suggests that it is crucial to adequately model phonetic confusion to cover the widespread *casual* transliterations.

### 3. PHONETIC SIMILARITY MODELING

Assuming that Chinese transliterations always co-occur in proximity to their original English words, the proposed phonetic similarity modeling (PSM) approach aims to identify the transliteration pairs by measuring phonetic similarity between candidate

transliteration pairs. In a Chinese-predominant text, when an English word  $EW$  is spotted, the task is to search for a possible Chinese transliteration  $CW$  for the English word<sup>2</sup>  $EW$  in its *close context*. For the PSM approach to work, we need to address three issues: (1) the candidacy of transliterations; (2) the search strategy of candidates; and (3) the training of PSM.

### 3.1 Formulation of Chinese Transliteration

Let  $ES = \{es_1, \dots, es_m, \dots, es_M\}$  be a sequence of English syllables derived from  $EW$  and  $CS = \{cs_1, \dots, cs_n, \dots, cs_N\}$  be the sequence of Chinese syllables derived from  $CW$ , which is represented by a Chinese character string.

$$CW \rightarrow c_1, \dots, c_n, \dots, c_N. \quad (3)$$

The transliteration can be considered a generative process formulated by the noisy channel model [Brown et al. 1994], with  $EW$  as the input and  $CW$  as the output.  $P(EW | CW)$  is estimated to characterize the noisy channel, known as the transliteration probability.  $P(CW)$  is a language model to characterize the source language. Applying the Bayesian rule, we have

$$P(CW | EW) = \frac{P(EW | CW)P(CW)}{P(EW)}. \quad (4)$$

Following the *translation-by-sound* principle, the transliteration probability  $P(EW | CW)$  can be approximated by the phonetic confusion probability  $P(ES | CS)$ , which is given as

$$P(ES | CS) = \max_{\Delta \in \Lambda} P(ES, \Delta | CS), \quad (5)$$

where  $\Lambda$  is the set of all possible alignment paths between  $ES$  and  $CS$ . It is not trivial to find the best alignment path  $\Delta$ . We can resort to a dynamic programming algorithm, which is discussed in Section 3.3. Assuming conditional independence of syllables  $es_m$ ,

we have  $P(ES | CS) = \prod_{m=1}^M p(es_m | cs_m)$  in a special case where  $M = N$ . With PSM, Eq.(4)

can be rewritten as

$$P(CW | EW) \approx \frac{P(ES | CS)P(CW)}{P(EW)}. \quad (6)$$

The language model in Eq.(6) can be represented by Chinese character  $n$ -gram statistics, trained from a LDC corpus<sup>3</sup> representing the usage statistics of Chinese characters in transliterated names.

$$P(CW) = \prod_{n=1}^N p(c_n | c_1, \dots, c_{n-2}, c_{n-1}). \quad (7)$$

<sup>2</sup> In this article, an English word can be a single word or a phrase of multiple words.

<sup>3</sup> LDC (<http://www ldc.upenn.edu/>) Chinese-English Name Entity List (LDC2003E01 v1.beta)

In this article we adopt a bigram with unigram back-off, equation(7) can be rewritten as equation(8):

$$P(CW) \approx p(c_1) \prod_{n=2}^N p(c_n | c_{n-1}). \quad (8)$$

$P(CW | EW)$  in Eq.(6) can be used to rank a number of  $CW$  candidates. In this article the context of  $EW$  provides us with a set of competing Chinese transliteration candidates  $\Omega$ . We rank the candidates to find the most likely  $CW$  for a given  $EW$ . This is equivalent to the generative process that generates the most likely  $CW$  from  $EW$  in most of the machine transliteration literature. In the ranking process,  $P(EW)$  can be ignored because it is the same across all  $CW$  candidates. The  $CW$  candidate that gives the highest posterior probability is considered the most probable candidate  $CW'$ .

$$CW' = \arg \max_{CW \in \Omega} P(CW | EW) \approx \arg \max_{CW \in \Omega} P(ES | CS)P(CW). \quad (9)$$

Now that we have short-listed one candidate  $CW'$  once, the next step is to see if  $CW'$  indeed forms a genuine  $E-C$  pair with the English word  $EW$ . This can be achieved by a hypothesis test. We take the ratio between  $P(CW' | EW)$  and  $\sum_{\substack{CW \in \Omega \\ CW \neq CW'}} P(CW | EW)$ ,

as the confidence index. We have  $H_0$ , which hypothesizes  $CW'$  and  $EW$  forms an  $E-C$  pair, and  $H_1$ , which hypothesizes otherwise. The confidence index can be given as follows:

$$\sigma = \frac{P(CW' | EW)}{\sum_{\substack{CW \in \Omega \\ CW \neq CW'}} P(CW | EW)} = \frac{P(ES | CS')P(CW')}{\sum_{\substack{CW \in \Omega \\ CW \neq CW'}} P(ES | CS)P(CW)}, \quad (10)$$

where  $CS'$  is the syllable sequence of  $CW'$ . For each  $EW$ , we start with the ranking process to come up with the most probable  $CW'$ . We then qualify the  $E-C$  pair by examining the confidence index against a threshold  $\varepsilon$  in a hypothesis test. The confidence index shows how much the genuine  $E-C$  pair overtakes the rest. This is referred to as the *recognition followed by validation strategy*.

$$\begin{aligned} \text{if } \sigma > \varepsilon \text{ then } H_0 \text{ is true} \\ \text{else } \sigma \leq \varepsilon \text{ then } H_1 \text{ is true} \end{aligned} \quad (11)$$

The choice of threshold  $\varepsilon$  affects system performance in terms of precision, recall, and  $F$ -measure.

$$\begin{aligned} \text{precision} &= \frac{\# \text{extracted\_DQTPs}}{\# \text{extracted\_pairs}}, \text{recall} = \frac{\# \text{extracted\_DQTPs}}{\# \text{total\_DQTPs}}, \\ F\text{-measure} &= \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \end{aligned} \quad (12)$$

In this article we choose  $\varepsilon$  empirically, so that it maximizes the  $F$ -measure on SET1. The  $E$ - $C$  pair qualifying process can be summarized in a 4-step algorithm, as follows. (In the following sections we will discuss each of the steps in detail.)

Recognition followed by validation Algorithm

1. Spot an  $EW$  from a Chinese sentence;
2. Decide the candidacy of transliterations in the close context of  $EW$ ;
3. Convert  $E$ - $C$  candidates to syllables,  $ES$  and  $CS$  identify the most probable  $CS'$  that matches  $EW$  using Phonetic Similarity Model;
4. Qualify  $CS'$  through the hypothesis test and accept  $CW'$  as transliteration of  $EW$ .

### 3.2 Candidacy of Transliteration

We aim at extracting transliteration pairs in predominantly Chinese Web pages, where transliterated words are collocated closely with their original English words and the English words are often appositives of neighboring Chinese words in a *close context*. In similar research, it was found that Japanese translated terms are often accompanied by their original English words in parentheses [Nagata et al. 2001]. The cross-lingual apposition has been observed not only in Japanese but also in Korean and Chinese articles as well. The Web corpus of such appositions serves as an important source from which to extract transliteration pairs. We make the following assumptions:

- (1) The *close context* is within a sentence boundary delimited by punctuation such as full stop, question, and exclamation marks. A *close context* is a range of proximity where an English word and its Chinese transliteration collocate.
- (2) In a *close context* there could be word pairs of both semantic translation and phonetic transliteration; only transliteration pairs are extracted.

Let's look at an example:

『...經營 Kuro 庫洛 P2P 音樂交換軟體的飛行網，3 日發表 P2P 與版權爭議的解決方案—C2C (Content to Community)...』

In the example, C2C is not a transliteration of “Content to Community,” it is an acronym expansion. On the other hand, “庫洛 /KU-LUO/”, not in parentheses, presents a transliteration for “Kuro”. What is important is that the English words and their transliterations are always closely collocated. Inspired by this observation, we propose an algorithm that searches over a *close context* of the English word for its transliteration candidate.

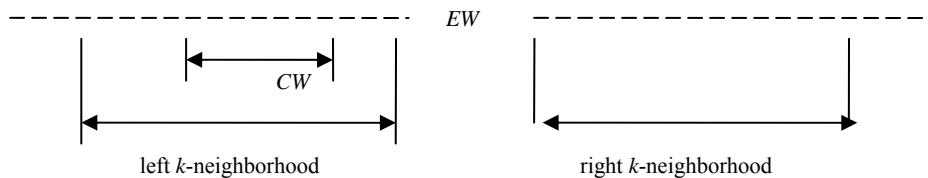


Fig. 1 Search space for transliteration of an anchor word  $EW$ .

Let's define several terms that we will use to facilitate the algorithm description. First, a predominantly Chinese Web page is segmented into sentences that are delimited by punctuation. Such a segmented sentence may or may not be a grammatical sentence. Second, we search for any English words  $EW$  in each sentence; there could be more than one English word in a sentence. Third, if an English word  $EW$  is recognized, then a  $k$ -neighborhood is defined as in Fig. 1, which serves as the *close context* of the recognized English word. Fourth, we define a Chinese transliteration candidate  $CW$  as a Chinese character string in the  $k$ -neighborhood. All the transliteration candidates in the  $k$ -neighborhood form a candidate set  $\Omega$  for  $CW$ . Suppose that we syllabify an English word  $EW$  into  $M$  syllables. We search in the candidate set  $\Omega$  over the  $k$ -neighborhood for a transliteration  $CW'$  of  $N$  Chinese syllables. Empirically,  $k$  can be set as five times the syllable number of  $EW$ . As discussed in Section 2.2, we typically have  $N \leq M$ ; in Section 3.3, we study how to account for phonetic elision in the phonetic similarity modeling.

In the example above, “Kuro” is recognized as an English word, “經營/JIN-YIN/” and “庫洛/KU-LUO/” are suggested in a *close context*, the left and right  $k$ -neighborhoods. Two candidate pairs, “Kuro-經營” and “Kuro-庫洛” will be examined further phonetically.

### 3.3 Search Strategy

We conducted a search to fulfill the *recognition* step in the *recognition followed by validation* strategy. First, given an English syllable sequence  $ES$  and its Chinese candidate  $CS$ , we find the best alignment path. Second, among the competing candidates, we identify the best  $CS'$  and therefore the best  $CW'$ .

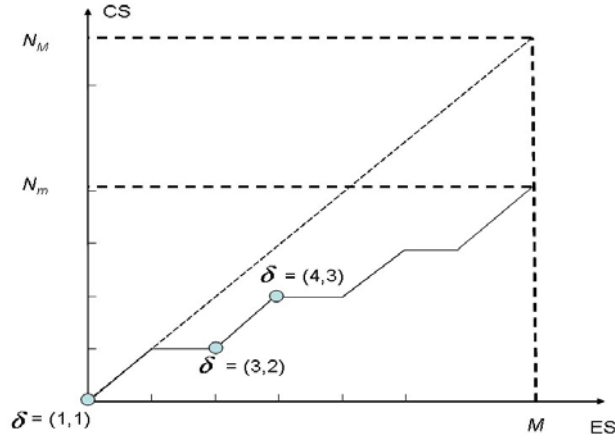
Token sequence alignment is a common issue in natural language processing [Brown et al. 1994]. Noting that  $M$  may or may not be equal to  $N$  due to syllable elision, we adopt a dynamic programming (DP) algorithm [Myers and Rabiner 1981] to eliminate the length differences between the two syllable sequences, which is also known as dynamic warping. To help understand the nature of dynamic warping, let us consider an  $M$ - $N$  plane, as in Fig. 2. The alignment between  $ES$  and  $CS$  can be depicted by a sequence of  $K$  points  $\delta = (m, n)$ :

$$\Delta = \delta_1, \dots, \delta_k, \dots, \delta_K, \quad (13)$$

where  $\delta_k = (m(k), n(k))$  denoting the  $k$ -th alignment point between  $es_m$  and  $cs_n$ , and  $\Delta \in \Lambda$ . The sequence can be considered to represent a function that approximately realizes a mapping from the time axis of  $ES$  to that of  $CS$ . Hereafter, it is called a warping function. Each warping function is the trace of a search path. All possible search paths form a search space  $\Lambda$ . In the case of no elision, the warping function coincides with the diagonal line  $n(k) = \frac{N}{M} \times m(k)$ . It deviates from the diagonal line as the elision happens more often, that is,  $N \leq M$ .

The probability of a mapping between  $es_m$  and  $cs_n$  is given as  $p(es_m | cs_n)$ . Assuming conditional independence of the syllables, the similarity between two syllable sequences given by a warping function  $\Delta$  is

$$P(ES | CS, \Delta) = \prod_{k=1}^K p(es_{m(k)} | cs_{n(k)}). \quad (14)$$

Fig. 2 Warping function between *ES* and *CS*.

The set of Chinese syllables  $\{cs_n\}$  includes the null symbol  $\phi$ . We have  $p(es_m | \phi)$  to model the elision  $es_m \rightarrow \phi$ , to address the zero fertility issue. Eq.(14) is also a measure for the goodness of the warping function  $\Delta$ , which attains its maximum value when warping function  $\Delta$  is determined to optimally align the transliteration pair to achieve the highest probability. Due to elision, there could be multiple *CSs* of different lengths in a *close context* for a given English word *ES* of  $M$  syllables. The length of *CS* ranges from  $N = 1$  of maximum elision to  $N = M$  of no elision. All the *CSs* in a *close context* of *ES* form a search space  $\Omega$ . (Please note that we have used  $\Omega$  to represent both the collection of *CWs* and their corresponding *CSs*.)

We identify the best *CS'* candidate by ranking the entire candidate set  $\Omega$  and taking all the possible warping functions into account. Equation(9) can be rewritten as equation (15), followed by equation(16). For each *CS* candidate, we first search for the best warping path  $\Delta_{CS}$

$$\Delta_{CS} = \arg \max_{\Delta \in \Lambda} P(ES | CS, \Delta), \quad (15)$$

then, we search for the best *CS'*, equivalently the best *CW'* transliteration, among all *CS* candidates,

$$CS' = \arg \max_{CS \in \Omega} P(ES | CS, \Delta)P(CW). \quad (16)$$

In the *validation* step, we qualify the resulting *CW'* with a hypothesis test that examines the posterior odds of the corresponding *E-C* pair as given in Eq.(11).

### 3.4 Phonetic Similarity Model

In equation(5), we propose to examine the *E-C* pair's phonetic similarity by using the phonetic confusion probability matrix, or confusion matrix. A confusion matrix is a  $V_e \times V_c$  matrix with its element being the conditional probability  $p(es | cs)$ , and  $V_e$  and  $V_c$  being the syllable vocabulary size in the respective language. To establish the confusion matrix, we convert the words from their orthography to phonemes using G2P [Galescu and Allen 2001]. Then, we convert the phoneme sequence further into a

sequence of Chinese-like syllables [Jurafsky and Martin 2000]. As a Chinese character is rendered by a syllable, it is preferable to compare the PSM at the syllable level. There are several ways to estimate the syllable-based confusion matrix:

(1) *ASR-PSM*: We can estimate the syllable confusion probability  $p(es_m | cs_n)$  through an automatic speech recognition (ASR) system. We run a labeled English speech database through a Chinese ASR system. We syllabify the English transcripts and align them with the Chinese syllable recognition results. As such, the confusion probability  $p_{ASR}(es_m | cs_n)$  can be obtained to serve as an educated guess about how syllables across the two languages are confused. Suppose that we have  $\# \langle es_m, cs_n \rangle$  as the count of aligned pairs between  $\langle es_m \rangle$  and  $\langle cs_n \rangle$  and  $\#cs_n$  as that of all  $\langle cs_n \rangle$  tokens. We estimate  $p_{ASR}(es_m | cs_n)$  as follows:

$$p_{ASR}(es_m | cs_n) = \frac{\# \langle es_m, cs_n \rangle}{\# \langle cs_n \rangle}. \quad (17)$$

(2) *Syllable-PSM*: We can estimate the syllable confusion probability  $p(es_m | cs_n)$  through extracted transliteration pairs. Given a transliterated bilingual corpus, we first convert the bilingual entries into syllables and phonemes. We then align the syllables and obtain the counts  $\#cs_n$  for  $\langle cs_n \rangle$  and the pairing counts  $\# \langle es_m, cs_n \rangle$  between  $\langle es_m \rangle$  and  $\langle cs_n \rangle$ . We estimate  $p_{SYL}(es_m | cs_n)$  as follows:

$$p_{SYL}(es_m | cs_n) = \frac{\# \langle es_m, cs_n \rangle}{\# \langle cs_n \rangle}. \quad (18)$$

(3) *Subsyllable-PSM*: We can estimate the syllable confusion probability  $p(es_m | cs_n)$  by using subsyllable confusion probability. Representing the syllables in terms of initial-final,

$$\begin{aligned} es &\rightarrow ei + ef \\ cs &\rightarrow ci + cf. \end{aligned} \quad (19)$$

We have  $ei$  and  $ef$  as initial and final for an English syllable  $es$ , and similarly  $ci$  and  $cf$  for a Chinese syllable  $cs$ . Assuming conditional independence between initials and

finals, we have  $p(ei_m | ci_n) = \frac{\# \langle ei_m, ci_n \rangle}{\# \langle ci_n \rangle}$  and  $p(ef_m | cf_n) = \frac{\# \langle ef_m, cf_n \rangle}{\# \langle cf_n \rangle}$  and

$$p_{SS}(es_m | cs_n) = p(ei_m | ci_n)p(ef_m | cf_n). \quad (20)$$

The three phonetic confusion matrices that we discussed above,  $p_{ASR}(es_m | cs_n)$ ,  $p_{SYL}(es_m | cs_n)$  and  $p_{SS}(es_m | cs_n)$ , can be exploited in different stages. In the initial stage, only ASR-PSM is available. Therefore, we use ASR-PSM to bootstrap the extraction process. Then, the phonetic confusion matrices can be interpolated for parameter smoothing. The smoothed confusion matrix (CM) of PSM can be expressed by equation (21):

$$p_{CM}(es_m | cs_n) = \alpha p_{SS}(es_m | cs_n) + \beta p_{SYL}(es_m | cs_n), \quad \alpha + \beta = 1.0, \quad (21)$$

where  $\alpha > 0$  and  $\beta > 0$  are the weights of Syllable-PSM and Subsyllable-PSM, respectively, and can be obtained empirically through cross-validations. In the cross-validation, we split SET1 into 10 sets. We withhold 1 set and use the remaining 9 sets to train a confusion matrix. We then use the withheld set as the test data to obtain an  $F$ -measure. Such a process is called 10-fold cross-validation. We test the performance of

parameter settings for  $\alpha$  and  $\beta$ . After several 10-fold cross-validations, we empirically adopt the parameters that give the highest average  $F$ -measure.

### 3.5 Learning Strategy

We conduct learning to train the confusion matrix from statistics derived from a corpus. Given a transliteration pair, the alignment between cross-lingual syllables and phonemes can be established by the DP process, as discussed in Section 3.3. With the alignment statistics, PSM parameter  $p(es_m | cs_n)$  can be estimated by an *Expectation-Maximization* (EM) process [Dempster et al. 1977]. We first obtain an ASR-PSM through speech recognition. The ASR-PSM is used as the initialization of PSM parameters. In the *Expectation* step, we compute the counts of events such as  $\# \langle es_m, cs_n \rangle$ ,  $\# \langle ei_m, ci_n \rangle$ ,  $\# \langle ef_m, cf_n \rangle$  and  $\# \langle cs_n \rangle$ ,  $\# \langle ci_n \rangle$ ,  $\# \langle cf_n \rangle$  by force-aligning the *E-C* pairs in the training corpus  $\Psi$ . The force-alignment process is described by Eq.(15) in Section 3.3. To account for elision, the counts of elision events  $\# \langle es_m, \phi \rangle$ ,  $\# \langle ei_m, \phi \rangle$ , and  $\# \langle ef_m, \phi \rangle$  are also computed. In the *Maximization* step, we estimate the PSM parameters  $p(es_m | cs_n)$  by Eq.(18) or Eq.(20). As the EM process guarantees non-decreasing likelihood probability  $\prod_{\Psi} P(ES | CS)$ , we let the EM process iterate until  $\prod_{\Psi} P(ES | CS)$  converges. The EM process is summarized as follows:

*Start:* Bootstrap confusion matrix  $p(es_m | cs_n)$  by ASR-PSM as in equation(17);

*E-Step:* Force-align corpus  $\Psi$  using  $p(es_m | cs_n)$ , then estimate the counts of  $\# \langle es_m, cs_n \rangle$ ,  $\# \langle ei_m, ci_n \rangle$ ,  $\# \langle ef_m, cf_n \rangle$  and  $\# \langle cs_n \rangle$ ,  $\# \langle ci_n \rangle$ ,  $\# \langle cf_n \rangle$ ;

*M-Step:* Re-estimate  $p(es_m | cs_n)$  using the counts from E-Step;

*Iteration:* Repeat E-Step and M-Step until  $\prod_{\Psi} P(ES | CS)$  converges.

## 4. EXPERIMENTS

We implement the phonetic similarity model and conduct several experiments. We start with a set of handcrafted syllable-mapping rules for mapping between English and Chinese. The rules serve as a syllable confusion matrix for hard decision-making. The results are regarded as the baseline performance of the extraction. We then compare the three PSM training approaches, namely ASR-PSM, Syllable-PSM, and Subsyllable-PSM, following the formulation in Section 3.4. We also compare supervised and unsupervised learning strategies in the PSM training. The experiments are first carried out using the SET1 development database. We also carry out our experiments on an independent database SET2 from a different source. SET2 consists of 3GB of Web pages acquired in a similar way as SET1. As a result, 24,507 DQTPs are expected to obtain at the estimated precision of 82.6%. SET2 allows us to examine our proposed method on a corpus from different sources. The precision, recall, and  $F$ -measure, collectively referred to as extraction performance, of the *E-C* pair extraction are reported for the experiments.

In Section 4.1, we report an experiment on the Romanized Chinese names on SET1, in which we build a confusion matrix based on direct orthographical mapping. In Sections 4.2 and 4.3, we discuss two methods for using ASR-PSM to bootstrap a confusion matrix; In Section 4.4, we use the confusion matrix resulting from Section 4.3

as the initialization and re-estimate the PSM through an iterative *Expectation-Maximization* process; finally, we extend the search from local *close context* in predominantly Chinese Web pages to a hyperlinked English-Chinese bilingual Web.

#### 4.1 Orthographical Confusion Matrix for Romanized Chinese

Since we focus on Chinese-predominant text, it is common that many English words are in fact their Romanization Chinese equivalents, which follow certain Romanization rules; Table VI gives some examples. An effective way to link them up is through direct orthographical mapping [Li et al. 2004] between the *EW* and *CW*. It is noted that there are multiple common Chinese Romanization systems, such as *Pinyin*, *Tong-yong*, Wade-Giles, and so on. Romanization systems are invented to account for dialect and accent variations in Chinese-speaking regions. The general principle is to represent a Chinese character in the form of a syllable. For example, “中壩” was Romanized to “Chung-li”, “Jhong-li”, and “Zhong-li” using Wade-Giles, *Tong-yong*, and *Hanyu Pinyin*, respectively. Each Romanization system has a syllable vocabulary. As such, identifying the Romanization system is just as important as identifying the origin of words in general letter-to-sound applications [Litjens and Black 2001]. In doing so, two problems arise: One is that some syllable vocabularies are common across different Romanization systems. Hence we are unable to correctly identify a Romanization system based on only a few syllables. Another problem is that it is not uncommon to see a mixture of syllable vocabulary from different Romanization systems in a single transliterated word [Kuo and Yang 2004b].

We adopt an approach to establish the PSM in the format of an orthographical confusion matrix. The confusion matrix establishes a correspondence between Romanized syllable codes and their Chinese characters for the above-mentioned three Romanization systems. As a result, we build a confusion matrix of 1,200 syllables with 13,000 characters. The 1,200 syllables encompass all the syllables in the three Romanization systems. A *CW* is parsed from left to right using a substring longest-match approach to arrive at a single segmentation of syllables. If multiple segmentations are possible, the most common one is retained according to unigram counts of syllables. Once the syllable segmentation is reached, the match between the *EW* and the Chinese word *CW* can be easily validated. In SET1, we have 135 Romanized Chinese names which constitute 1.5% of the total 8,898 DQTPs. Using the orthographical confusion matrix, we correctly extract 134 Romanized Chinese DQTPs in SET1. In Table VII we report the performance on the Romanized Chinese subset of SET1. From the result, we can see that reasonably good precision and recall can be obtained by extracting 159 candidates, of which 134 are correct DQTPs. This translates to 84.2% (134/159) preci-

Table VI. Romanization as Transliteration

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 (臺灣)	Whagwei (碗粿)	Gezaixi (歌仔戲)	Owanchian (蚵仔煎)	Jungtsu (粽子)

Table VII. Qualifying *E-C* Pairs by the Orthographical Confusion Matrix for Romanized Chinese

#DQTPs in SET1	#Extracted DQTPs	Precision	Recall
135	134	84.2%	99.3%

Table VIII. Results Using the Confusion Matrix with Hard Decision

	Rule	Rule with Elision
#DQTPs	996	1,165
Precision	76.1%	76.5%
Recall	11.2%	13.1%
<i>F</i> -measure	0.20	0.22

on and 99.3% (134/135) recall for Romanized Chinese. If the syllable segmentation is unsuccessful, we follow the *E-C* pair qualifying process, as in Section 3.1.

#### 4.2 Phonetic Confusion Matrix with Hard Decision

Human translators rely on phonetic rules that convert English phonemes into their Chinese equivalents [Wan and Verspoor 1998]. To establish the baseline performance, we generate a set of handcrafted syllable-mapping rules; a similar approach was reported by Tsuji [2002]. The rules can be translated into a confusion matrix in the format of equation(18), but with *hard decisions*. That is,  $p(es_m | cs_n)$  is set to 1 if there exists a mapping  $es_m \rightarrow cs_n$  and  $p(es_m | cs_n)$  is set to 0 otherwise. Assuming no-elision, we call this the *rule-based approach*. To account for the elision of English sounds, we conduct a search, as formulated in equations (15) and (16), with the *rule with elision approach*. The results on SET1 are reported in Table VIII. We observe that the *rule with elision approach* slightly improves the *F*-measure. However, either approach gives too low a recall rate to be practical for any real applications. In this experiment,  $\epsilon$  is empirically set to 0.3.

#### 4.3 Phonetic Confusion Matrices with Soft Decision

The handcrafted rules are good for most *regular* transliterations [Xinhua 1992]. In today's press, most of the transliterations are considered *casual*, and hence fail to generalize to account for variations. So a learning mechanism to automatically derive the soft decision from a corpus is desirable.

In automatic speech recognition (ASR), we report the confusion matrix across sound units for error analysis. The confusion matrix provides an invaluable source for bootstrapping cross-lingual syllable-based PSM. A confusion matrix reports the confusion statistics of syllables between two languages. The diagonal line in the confusion matrix reflects the desired matching results. An ASR system typically provides a confusion matrix with strong correlation values along the diagonal and weak correlation off the diagonal. When English utterances pass through a Chinese ASR system, we obtain a confusion matrix between English phonemes and Chinese phonemes, and similarly between English syllables and Chinese syllables. It is expected that similar sounds across

Table IX. Results Using Phonetic Confusion Matrices with *Soft Decision*

	SCM	PCM	CM
#DQTP	1,802	3,050	3,608
Precision	75.7%	80.3%	76.0%
Recall	20.3%	34.3%	40.5%
<i>F</i> -measure	0.32	0.48	0.53

languages will be characterized by strong correlation values in the cross-lingual confusion matrix. This process helps establish the phonetic relevance across languages. We conduct three experiments here. First, we generate a cross-lingual syllable-based PSM in the format of a confusion matrix with *soft decision* using the ASR-PSM approach, referred to as the SCM approach. Second, we generate a similar phoneme-based PSM using the ASR-PSM approach, referred to as PCM. Third, we use linear interpolation to combine PCM and SCM, as in Eq. (21), for model smoothing, referred to as CM. We adopt the cross-validation approach as discussed in Section 3.4 to estimate linear interpolation parameters. A set of parameters  $\alpha = 0.3$  and  $\beta = 0.7$  are chosen as a result of the 10-fold cross-validation on SET1. The hypothesis test threshold  $\epsilon$  is empirically set to 0.3. The test results on SET1 are shown in Table IX. Compared to the results in Table IIX, we see that the *soft decision* confusion matrix significantly improves extraction performance over the *hard decision* strategy.

#### 4.4 Learning Confusion Matrices through EM Process

We have discussed how to establish confusion matrices from linguistic knowledge such as an orthographical confusion matrix and from phonetic knowledge such as ASR results. The collection of extracted transliteration pairs by the soft decision approach above serves as an informative parallel corpus for model learning as well. In Section 3.5, we discussed a model training method by means of an EM process. Now we would like to see how such a confusion matrix learning process can help improve performance.

In the first experiment, we use the 8,898-DQTP parallel corpus as the gold standard. We first split the 8,898 DQTPs in SET1 into 10 sets and then conduct 10 cross-validations. For each cross-validation, we withhold 1 set of data for testing and use the remaining 9 sets for the training of syllable and phoneme confusion matrices, as formulated in Syllable-PSM and Subsyllable-PSM in Section 3.4. We consider this experiment a *supervised* training process, referred to as a *supervised learning*.

In the second experiment, we carry out *unsupervised* learning on the same development corpus, SET1. In this experiment, we do not rely on the gold standard as the parallel corpus. We start with the CM resulting from ASR-PSM to extract *E-C* pairs. Instead of using the gold standard, we use the newly extracted *E-C* pairs to re-estimate the confusion matrix. Note that there are many falsely detected *E-C* pairs on the list. In each iteration, the surviving *E-C* pairs after screening are used to align and re-estimate the PCM, SCM, and CM. Multiple iterations are carried out until a convergence criterion is met. We can see that performance improves over iterations, as shown in Fig. 3. It is noteworthy that PCM outperforms SCM at the beginning, whereas SCM overtakes PCM

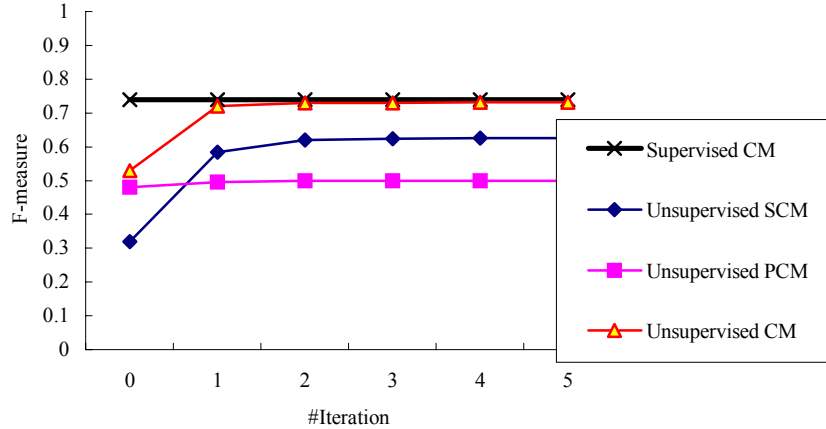


Fig. 3. Learning phonetic confusion matrices through an EM process.

after a few iterations. This is probably due to the fact that there are fewer parameters in the PCM than in the SCM, which allows PCM to converge quicker. As expected, SCM eventually captures the syllabic characteristics of Chinese transliteration well.

The results of *supervised* and *unsupervised* learning are reported in Table X. It is interesting to find that *supervised* and *unsupervised* learning lead to close extraction performance, with *supervised* learning giving slightly better results. The *unsupervised* learning process allows us to get rid of the tedious manual validation of DQTPs at the cost of a slight degradation in performance. It learns PSM as it discovers new transliteration pairs.

We further apply the *unsupervised* learning technique to a test corpus SET2 that we acquired from an independent source. By examining 400 of the extracted *E-C* pairs, we report a precision rate of 82.6%, which is close to the results in CM, as shown in Table X for SET1; some of the extracted *E-C* pairs are shown in Table XI. In general, words of Japanese and Korean origin are translated to Chinese through direct orthographic map-

Table X. Supervised vs. Unsupervised Learning of Confusion Matrices

	Supervised/ Unsupervised SCM	Supervised/ Unsupervised PCM	Supervised/ Unsupervised CM
#DQTPs	5,401/4,379	3,481/3,485	5,899/5,742
Precision	80.2%/85.7%	69.4%/68.6%	83.4%/84.3%
Recall	60.7%/49.2%	39.1%/39.2%	66.3%/64.5%
<i>F</i> -measure	0.691/0.625	0.500/0.499	0.739/0.731

Table XI. Transliteration Pairs Extracted Using an Unsupervised Approach from SET2

Word Origin/Type	Extracted <i>E-C</i> Pairs				
Chinese (Mandarin)	Hubei (湖北)	Sohoo (搜弧)	Ilan (宜蘭)	Harbin (哈爾濱)	Kunming (昆明)
Japanese	Seibu (西武)	Iga (伊賀)	Kanji (漢字)	Sanyo (三洋)	Nikkan (日刊)
Korean	Ssangyong (雙龍)	Kyonggi (京畿)	Hyundai (現代)	Hynix (海力士)	Pusan (釜山)
Taiwanese	Kavannan (蛤仔難)	Tanangan (打那岸)	Auvunken (翁文卿)	Huwei (滬尾)	Qauqaut (猴猴)
Western Languages	Vanderbilt (范德比)	Wucherer (伍賀瑞)	Sforza (史佛拉)	Sphinx (斯陰克斯)	Sydney (悉尼)
Western Languages	Sgismund (司基世蒙德)	Paradiz (派拉吉斯)	Capek (恰彼克)	Weinstein (溫斯坦)	Limousin (利穆贊)
Newly Transliterated	Logo (漏狗)	Homework (洪沃客)	Style (史黛爾)	Fans (粉絲)	Lightning (雷霆)
Drug	Lamivudine (拉美芙錠)	Xenica (讓你酷)	Elisa (一粒沙)	Ribavirin (雷巴威林)	Ritalin (利他能)

ping. It is interesting to note that some of them are correctly extracted by our search strategy due to similarities in phonetic systems among the Chinese, Japanese, and Korean languages.

#### 4.5 Learning Confusion Matrices from a Bilingual Web

Web pages are connected through hyperlinks and are woven into a vast network. An iterative method was proposed to identify hubs and authorities in this hyperlinked environment and to refine search topics by using information on hub pages and authoritative pages [Kleinberg 1998]. Hyperlink analysis has been widely used in information retrieval research [Brin and Page 1998], and has achieved promising results in statistical term translation [Lu et al. 2002]. As hyperlinks establish the association of *E-C* pairs across different languages, we extend our *close context* proximity from sentences to hyperlinked cross-lingual Web pages. We further collect 1.98 million Web pages using a Web spider. Among these pages, 109,416 linked text pairs (anchored texts [Lu et al. 2002]) were extracted. This corpus is known as SET3, which is obtained in a similar way to SET1. We use confusion matrices from Section 4.3 for bootstrapping, and apply the *unsupervised* learning technique to extract *E-C* pairs from SET3. To understand the extraction performance, we randomly select 400 extracted *E-C* pairs for manual validation. Results are reported in Table XII, where we see that precision is comparable to rates for SET1 and SET2. The results further validate that the *unsupervised* learning technique can be reliably extended to transliteration extraction from a bilingual Web.

Table XII. Learning about Confusion Matrices from SET3

	Unsupervised CM
#Extracted DQTPs	314
Precision	78.5%

## 5. RELATED WORK

Let us revisit the two categories of machine transliteration research that we discussed in Section 1, the transliteration modeling (TM) and the extraction of transliteration pairs (EX). The proposed PSM benefits from the both studies: it adopts the TM approach to the EX problem.

The phonetic similarity model (PSM) is formulated by a generative model under the noisy channel modeling framework [Brown et al. 1994]. It learns from multiple levels of knowledge sources, such as orthography in Section 4.1, and syllables and phonemes in Section 4.2 and 4.3. This is inspired by the finding that different levels of knowledge sources help in the transliteration process. For example, Kuo and Yang [2004a] and Lam et al. [2004] used phonetic similarity in matching name entity translations. Meng et al. [2001] proposed a phoneme-based name entity transliteration. Li et al. [2004] proposed a direct orthographical mapping framework. The PSM framework effectively fuses the three levels of knowledge sources in the decision making process.

The effort in EX research is motivated by the idea that transliteration pairs exist in bilingual corpora such as bilingual text, interlinked bilingual Web pages, etc. Al-Onaizan and Knight [2002] and Qu et al. [2003] reported using Web pages as a live thesaurus to validate generated translations and transliterations. The EX approach allows us to extract real life transliteration pairs. The problem in EX research centers on how to establish the transliteration correspondence between words.

Prior work in EX research falls into two categories. One work category extracts the context of a bilingual word pair, with the assumption that if the contexts of two words in their respective languages are similar in a comparable/parallel bilingual corpus, then the word pair is considered to be a translation pair. The same approach, known as the context-based approach, is valid for transliteration of new words as well. This approach does not make use of the phonetic link between words. Lu et al.'s approach [2002] falls into this category. The success of such techniques relies on the availability of a comparable/parallel corpus.

Another category attempts to explore the phonetic similarity between a transliteration pair. In this way, the generative model, as in a TM study, can be used to establish the phonetic correspondence between bilingual words. Many efforts have been reported along this line: Brill et al. [2001] proposed extracting transliteration pairs from query logs; Lee and Chang [2003] and Lam et al. [2004] attempted to acquire transliteration pairs from parallel corpora and comparable corpora, respectively. In this article, for the first time, we enlarge the scope of corpus sources for EX by looking into monolingual texts, which are more accessible than either query logs or comparable corpora. We start with the targeted monolingual texts where Chinese transliterations are collocated closely with their appositives in English. We use the collocation proximity to constrain the search space and propose a two-stage extraction process, known as *recognition followed by validation*.

## 6. DISCUSSION

To understand the SET1 task further, we study the SET1 statistics and its DQTPs. As described in Section 3.1, the Chinese words are first converted to syllables for phonetic similarity comparison (it is interesting to know how complex the syllable-mapping task will be). In Table XIII, we show the word and syllable statistics for the entire SET1 and its DQTPs. It is not surprising that Chinese DQTPs use a smaller subset of Chinese characters (1,210) for transliteration than the entire SET1 does (3,595), since the Chinese vocabulary is a few times larger than English because SET1 is a Chinese-predominant corpus, while the mapping between *E-C* pairs is roughly one-to-one. The ratio between Chinese and English syllables reflects the complexity of the PSM. We find that there are 394 and 1,012 distinct Chinese and English syllables in the entire SET1, and 333 and 824 distinct Chinese and English syllables in the DQTPs, respectively. In other words, each Chinese syllable corresponds to about 2.5 English syllables on average, so it is a great challenge to manually code the syllable-mapping rules. Our experiments show that the SCM and PCM models allow us to learn the rules effectively from the data.

We have extracted a large quantity of transliteration pairs from SET2 and SET3 in our experiments. The extracted transliteration pairs form a bilingual lexicon, and it is desirable to look into the statistics of the extracted lexicon as well.

Comparing the English words in the extracted bilingual lexicon to the *CMU Pronunciation Dictionary*<sup>4</sup> and the *Shorter Oxford English Dictionary*, we discovered that 31.1% and 47.8% of the English words were not found in the respective dictionaries. Not only did the PSM framework extract new, real-life transliteration pairs, but it also acquired a significant amount of new English vocabulary from the Web. The *unsupervised* learning mechanism also allows us to acquire new transliteration pairs in a more scalable and cost-effective manner than the manual one.<sup>5</sup> Furthermore, the PSM adopts a probabilistic framework that offers multiple transliteration variants for each English word.

We have seen that the proposed PSM model effectively handles *casual* transliteration. Despite the fact that statistics show that 68.48% of *casual* transliterations in SET1 were successful, extracting *casual* transliteration remains a challenging task that is affected by wordplay, regional dialect, and the gender of English personal names. Note that sound inventories and phonic rules differ from language to language. Words in the Latin alphabet in English may not entirely be of English origin. As pointed out by Li et al. [2004], a single G2P system does not work for English words of different origins. In this article, we show that prior knowledge about the Romanization system of a Chinese name helps direct orthographical mapping. Similarly, if the information about word origins is available, it will improve the G2P system and the PSM framework in turn, as a whole. Error analysis on the extracted lexicon shows that the recall rate is low for foreign words

Table XIII. Number of Distinct Entries in SET1 and in its DQTPs

	Chinese Words	Chinese Syllables (Characters)	English Words	English Syllables
Whole SET1	80,501	394 (3,595)	21,353	1,012
DQTPs	7,902	333 (1,210)	7,005	824

<sup>4</sup> <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

<sup>5</sup> <http://client.cna.com.tw/name/>

of Korean and Japanese origins, which is probably due to the fact that Romanization rules of Korean and Japanese names were not adequately captured in a PSM model trained mainly on English-Chinese pairs.

Translation pairs such as “Korea-韓國/HAN-GUO/” and “America-美國/MEI-GUO/” made up of both semantic translation and phonetic transliteration, constitute a large percentage of translated name entities; the PSM framework in this article does not handle such cases. There are some results reported by Al-Onaizan and Knight [2002]; Chen et al. [2003]; and Huang et al. [2004] on mixed translation and transliteration modeling. We believe that the combination of context-based and PSM approaches will provide a good solution to the problem of mixed translation and transliteration pair extraction.

## 7. CONCLUSIONS

We have proposed a novel PSM model for extracting transliteration pairs; our contribution can be summarized as follows. (i) we propose a PSM model that accounts for both *regular* and *casual* transliterations by exploiting orthographic, syllabic, and phonetic information; (ii) we propose both *supervised* and *unsupervised* training algorithms for PSM modeling in an EM framework. The *unsupervised* training algorithm performs closely to the *supervised* one, and provides a low-cost alternative to tap the live and dynamic Web for new transliteration pairs; (iii) we propose an effective EX process, *recognition followed by validation*. Recognition is done via a dynamic programming search strategy; *validation* is achieved through a hypothesis test; and (iv) we propose an EX strategy that exploits the lexical collocation information between Chinese transliterations and English appositives in a monolingual corpus. The new EX strategy expands the scope of corpus sources, as monolingual corpora are much more prevalent than parallel corpora.

The results shown in this article confirm the effectiveness of the PSM framework in *E-C* transliteration pair extraction. Without loss of generality, the same framework is applicable to other language pairs such as English-Japanese and English-Korean. Although we start with experiments on monolingual Chinese text, we also explore extending the framework to EX in anchored texts.

## ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable comments and suggestions. We also thank Yu Chen at the Institute for Infocomm Research, Singapore, for her efforts in improving the manuscript; Wen-Hsiang Lu at the National Cheng-Kung University for providing hyperlink and Web page corpora; and Wern-Jun Wang at Chung-Hwa Telecommunication Laboratories for providing speech data.

## REFERENCES

- AL-ONAIZAN, Y. AND KNIGHT, K. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 400-408.
- BRILL, E., KACMARCIK, G., AND BROCKETT, C. 2001. Automatically harvesting Katakana-English term pairs from search engine query logs. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, 393-399.
- BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7<sup>th</sup> International World Wide Web Conference*, 107-117.
- BROWN, P. F., DELLA PIETRA, S. A., DELLA PIETRA, V. J., AND MERCER, R. L. 1994. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19, 2, 263-311.
- CHEN, H. H. AND LEE, J. C. 1996. Identification and classification of proper nouns in Chinese texts. In *Proceedings of the 16<sup>th</sup> International Conference on Computational Linguistics*, 222-229.
- CHEN, H. H., YANG, C. H., AND LIN, Y. 2003. Learning formulation and transformation rules for multilingual entities. In *Proceedings of 41<sup>st</sup> ACL Workshop on Multilingual and Mixed-language Named Entity Recognition*, 1-8.

- DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Ser. B. Vol. 39*, 1-38.
- GALESCU, L. AND ALLEN, J. 2001. Bi-directional conversion between graphemes and phonemes using a joint N-gram model. In *Proceedings of the International Speech Communication Association Tutorial and Research Workshop of Speech Synthesis*, 103-108.
- GAO, W., WONG, K. F., AND LAM, W. 2004. Phoneme-based transliteration of foreign names for OOV problem. In *Proceedings of the 1<sup>st</sup> International Joint Conference on Natural Language Processing*, 374-381.
- HUANG, F., VOGEL, S., AND WAIBEL, A. 2004. Improving name entity translation combining phonetic and semantic similarities. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting*, 281-288.
- JUNG, S. Y., HONG, S. L., AND PAEK, E. 2000. An English to Korean transliteration model of extended Markov window. In *Proceedings of the 18<sup>th</sup> International Conference on Computational Linguistics*, 383-389.
- JURAFSKY, D. AND MARTIN, J. H. 2000. *Speech and Language Processing*. Prentice-Hall, Englewood Cliffs, NJ, 91-188.
- KANG, B. J. AND CHOI, K. S. 2000. Automatic transliteration and back-transliteration by decision tree learning. In *Proceedings of the 2<sup>nd</sup> International Conference on Language Resource and Evaluation*, 1135-1411.
- KANG, I.H. AND KIM, G. C. 2000. English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks. In *Proceedings of the 18<sup>th</sup> International Conference on Computational Linguistics*, pp. 418-424.
- KLEINBERG, J. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9<sup>th</sup> ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, 14-20.
- KNIGHT, K. AND GRAEHL, J. 1998. Machine transliteration. *Computational Linguistics* 24, 4, 599-612.
- KUO, J.S. AND YANG, Y.K. 2004a. Constructing transliterations lexicons from Web corpora. In *The Companion Volume to Proceedings of the 42<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics*, 102-105.
- KUO, J. S. AND YANG, Y. K. 2004b. Generating paired transliterated-cognates using multiple pronunciation characteristics from Web corpora. In *Proceedings of the 18<sup>th</sup> Pacific Asia Conference on Language, Information and Computation*, 275-282.
- KUO, J. S. AND YANG, Y. K. 2005. Incorporating pronunciation variation into extraction of transliterated-term pairs from Web corpora. In *Proceedings of the International Conference on Chinese Computing*, 131-138.
- LAM, W., HUANG, R. Z., AND CHEUNG, P. S. 2004. Learning phonetic similarity for matching named entity translations and mining new translations. In *Proceedings of the 27<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 289-296.
- LEE, C. J. AND CHANG, J. S. 2003. Acquisition of English-Chinese transliterated word pairs from parallel-aligned texts using a statistical machine transliteration model. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting Workshop on Building and Using Parallel Texts Data-Driven Machine Translation and Beyond*, 96-103.
- LEE, J. S. AND CHOI, K. S. 1998. English to Korea statistical transliteration for information retrieval. *Computer Processing of Oriental Languages* 12, 1, 17-37.
- LI, H., ZHANG, M., AND SU, J. 2004. A joint source channel model for machine transliteration. In *Proceedings of the 42<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics*, 159-166.
- LIN, W. H. AND CHEN, H. H. 2002. Backward machine transliteration by learning phonetic similarity. In *Proceedings of the Sixth Conference on Natural Language Learning*, 139-145.
- LIN, T., WU, J. C., AND CHANG, J. S. 2004. Extraction of name and transliteration in monolingual and parallel corpora. In *Proceedings of the 6<sup>th</sup> Conference of the Association for Machine Translation in the Americas*, 177-186.
- LLITJOS, A. F. AND BLACK, A. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proceedings of Eurospeech'2001, Vol. 3*, 1919-1922.
- LU, W. H., CHIEN, L. F., AND LEE, H. J. 2002. Translation of Web queries using anchor text mining. *ACM Trans. on Asian Language Information Processing* 1, 2, 159-172.
- MENG, H., LO, W. K., CHEN, B., AND TANG, K. 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*, 311-314.
- MYERS, C.S. AND RABINER, L.R. 1981. A comparative study of several dynamic time-warping algorithms for connected word recognition. *Bell System Technical J.* 60, 7, 1389-1409.
- NAGATA, M., SAITO, T., AND SUZUKI, K. 2001. Using the Web as a bilingual dictionary. In *Proceedings of the 39<sup>th</sup> ACL Workshop on Data-Driven Methods in Machine Translation*, 95-102.
- OH, J. H. AND CHOI, K. S. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *Proceedings of the 19<sup>th</sup> International Conference on Computational Linguistic*, 758-764.

- PAGEL, V., LENZO, K., AND BLACK, A. 1998. Letter to sound rules for accented lexicon compression. In *Proceedings of the International Conference on Spoken Language Processing*, 2015-2020.
- QU, Y., GREFFENSTETTE, G., AND EVANS, D. 2003. Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of the 26<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 353-360.
- TSUJI, K., DAILLEY, B., AND KAGEURA, K. 2002. Extracting French-Japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the 3<sup>rd</sup> International Conference on Language Resources and Evaluation*, 499-502.
- VIRGA, P. AND KHUDANPUR, S. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the 41<sup>st</sup> ACL Workshop on Multilingual and Mixed Language Named Entity Recognition*, 57-64.
- WAN, S. AND VERSPOOR, C.M. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proceedings of the 17<sup>th</sup> International Conference on Computational Linguistics and the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 1352-1356.
- XIAO, J., LIU, J., AND CHUA, T.S. 2002. Extracting pronunciation-translated names from Chinese texts using a bootstrapping approach. In *Proceedings of the 1<sup>st</sup> SIGHAN Workshop on Chinese Language Processing*, 1-6.
- XINHUA NEWS AGENCY. 1992. *Chinese Transliteration of Foreign Personal Names*. The Commercial Press.

Received May 2005; revised September 2006; revised December 2006; accepted March 2007