Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 91-120 


# Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names

**Wen-Hsiang Lu*, Jiun-Hung Lin+, and Yao-Sheng Chang***

**Abstract**

Unknown term translation is important to CLIR and MT systems, but it is still an unsolved problem. Recently, a few researchers have proposed several effective search-result-based term translation extraction methods which explore search results to discover translations of frequent unknown terms from Web search results. However, many infrequent unknown terms, such as abbreviations and proper names (or named entities), and their translations are still difficult to be obtained using these methods. Therefore, in this paper we present a new search-result-based abbreviation translation method and a new two-stage hybrid translation extraction method to solve the problem of extracting translations of infrequent unknown abbreviations and proper names from Web search results. In addition, to efficiently apply name transliteration techniques to mitigate the problems of proper name translation, we propose a mixed-syllable-mapping transliteration model and a Web-based unsupervised learning algorithm for dealing with online English-Chinese name transliteration. Our experimental results show that our proposed new methods can make great improvements compared with the previous search-result-based term translation extraction methods.

**Keywords:** CLIR, Transliteration, Unknown Term Translation, Web Search Result, Machine Translation.

* Dept. of Computer Sci. and Eng., National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan (R.O.C.)
Tel.: +886-6-2757575 ext: 62545 Fax: +886-6-2747076
E-mail: whlu@mial.ncku.edu.tw, ys.chang1976@gmail.com
+ Penpower Technology Ltd. 7F, NO.47, Lane 2, Sec. 2, Guanfu Rd., Hsinchu City 300, Taiwan, R.O.C.
Tel: +886-3-5722691 Fax: +886-3-5716243
E-mail: hunter@penpower.com.tw

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

## 1. Introduction

Many existing cross-language information retrieval (CLIR) systems [Ballesteros and Croft 1997; Hull and Grefenstette 1996] encounter great difficulties in dealing with unknown term translation since these systems rely mostly on general-purpose bilingual dictionaries, which usually lack translations of abbreviations and proper names. Moreover, according to the report in a previous work [Cheng *et al.* 2004], even for frequent Web queries, about 64% of them are not covered in an English-Chinese lexicon with about 120K entries (provided by Linguistic Data Consortium). However, several automatic translation extraction methods based on parallel [Brown *et al*. 1993; Melamed 2000; Nie *et al*. 1999; Smadja *et al*. 1996] or comparable corpora [Rapp 1999; Fung and Yee 1998] eventually suffer from the problems of insufficient parallel texts and the shortage of translation accuracy of comparable corpora in various subject domains.

The Web has been expanded with an enormous amount of multilingual hypertext resources in diverse subjects. Recently, a number of studies in natural language processing (NLP) have concentrated on the use of Web resources to complement insufficient text corpora [Cao and Li 2002; Kilgarriff and Grefenstette 2003]. To automatically collect huge amounts of parallel corpora from the Web in various domains, some researchers have developed feasible techniques of utilizing similar file names, text length, and link structures to extract parallel text pages from bilingual Web sites [Nie *et al*. 1999; Resnik 1999; Yang and Li 2003]. On the other hand, Lu *et al.* [2002] made the first attempt of mining unknown term translations from Web anchor texts. Both Cheng *et al*. [2004] and Zhang and Vines [2004] have explored language-mixed search-result pages for extracting translations of frequent unknown queries. Although these approaches have successfully enhanced the performance of frequent unknown query translation, they still suffer from the problems of data sparseness and indirect association errors in finding translations of infrequent unknown query terms, particularly for abbreviations and proper names [Melamed 2000].

In this paper, we focus on dealing with two kinds of translation of unknown query terms, including proper names and abbreviations. According to the report in Davis and Ogden [1998], about 50% of unknown terms in queries are proper names. Most methods handling translations of proper names are based on name transliteration techniques [Knight and Graehl 1998; Lin and Chen 2002; Lin *et al*. 2003; Li *et al*. 2004]. One major drawback of these methods is that they do not consider semantic information. Lam *et al*. [2004] proposed a named entity matching model, which considers both semantic and phonetic information, and applied it in mining unknown named entity translations from online daily Web news. Huang *et al*. [2005] also presented a method to extract key phrase translations from the language-mixed search-result pages with phonetic, semantic and frequency-distance features. As for abbreviation translation, less attention has been put on this research topic in the past few years.

Different from the above works, our major goal is to solve the problems of query translation to help users access English/Chinese information in cross-lingual Web searches. In this paper, therefore, we concentrate our attention on the challenge of dealing with the translations of infrequent unknown abbreviations and transliterated names in Web search queries, *i.e.*, these unknown queries that appear infrequently in Web query logs. We present two new methods to effectively extract translations of these two kinds of infrequent unknown queries. First, we propose a search-result-based abbreviation translation method for handling bidirectional translation of abbreviations in Chinese/English. Second, a new two-stage hybrid translation extraction method, which combines Cheng *et al.*'s [2004] search-result-based term translation extraction method and a new Web-based transliteration method, is proposed to extract Chinese/English translations for infrequent unknown English/Chinese proper names. In addition, to train an effective transliteration model, we also present a Web-based unsupervised learning algorithm to automatically collect large amounts of diverse English-Chinese transliteration pairs from the Web. For application, we provide a real prototype website[1] for users to translate unknown terms in practice. Our experimental results show that the proposed new methods can make great improvements in extracting infrequent unknown term translation.

The remainder of this paper is organized as follows: Section 2 describes the problems of unknown term translation and our search-result-based term translation extraction approach. Section 3 evaluates the proposed approach. Section 4 provides a simple description and comparison with the related work. Section 5 gives our conclusions.

## 2. Search-Result-Based Unknown Term Translation

### 2.1 Problems

Cheng *et al.*'s search-result-based term translation extraction method (refer to Section 2.3) is effective in extracting translations for frequent unknown query terms. However, for a lot of infrequent abbreviations and proper names, their translations are still difficult to extract. For example, while submitting an English abbreviation "AMIA" to LiveTrans[2], an incorrect Chinese translation "系列" (series) is obtained. The reason might be that some abbreviations are semantically ambiguous and co-occur relatively infrequently with the correct Chinese translations of their full names (or original forms). However, we observe that for an English abbreviation, its full name may co-occur more frequently with its corresponding Chinese translation. Thus, to effectively extract correct translation for an infrequent abbreviation, our idea is to first identify its full name in search results, and then extract correct translation of its

[1] http://ws.csie.ncku.edu.tw/~jhlin/cgi-bin/index.htm

[2] http://livetrans.iis.sinica.edu.tw/: This website is developed based on the search-result-based term translation extraction method by Web Knowledge Discovery lab of Academia Sinica, Taiwan.

full name, using the search-result-based term translation extraction method mentioned above. Generally, it should be more feasible to extract the correct translation of an abbreviation via its full name. For example, if we can extract the full name of the abbreviation "AMIA", "American Medical Informatics Association", then we can get its correct Chinese translation "美國醫學資訊協會" via LiveTrans.

On the other hand, an English proper name might have multiple Chinese transliterated names which often vary with different translators due to phonetic variation and the lack of standard transliteration rules [Gao *et al*. 2004]. In other words, there may be several Chinese transliterated names corresponding to an English name. For example, the name "Disney" has various Chinese transliterated names, including "迪士尼", "迪斯尼", "迪斯奈", "狄斯奈", and "狄士尼"; the name "Hussein" also has several different Chinese transliterated names, including "海珊", "哈珊", and "侯塞因". Obviously, it will be helpful for query translation in cross-lingual Web search if we can collect all possible transliterated names from the Web for each unknown proper name. However, it is a real challenge to find all the various transliterated names. Thus, we consider integrating name transliteration techniques into the process of translation extraction for infrequent unknown proper names. Our idea is that we first extract high-frequency terms from the search-result pages as transliteration candidates, and then filter out impossible candidates by using a name transliteration model. In fact, it is still challenging to build an effective transliteration model while lacking sufficient transliteration pairs for training. Therefore, we propose a Web-based unsupervised learning algorithm to automatically collect large amounts of English-Chinese transliteration pairs from Web search results.

## 2.2 Overview of the Proposed Approach

Figure 1 demonstrates the process of our search-result-based query translation method. First, an unknown term is determined by a general-purpose dictionary. Then, an unknown term is recognized as an abbreviated term using our search-result-based abbreviation translation extraction methods. If the unknown term does not belong to an abbreviated term, we have to examine whether the unknown term is a transliteration based on our two-stage hybrid translation extraction method. To deal with unknown term translation, we employ the search-result-based term translation extraction method (described in Section 2.3) to handle translation of frequent (popular) unknown query terms, and propose two new infrequent unknown translation methods, namely the search-result-based abbreviation translation extraction method (Section 2.4) and two-stage hybrid translation extraction method (Section 2.5), to solve the problems of translation of abbreviated terms (*i.e*., abbreviations) and transliterated terms (*i.e*., proper names). To recognize the abbreviated terms in queries, we

collected an abbreviation list containing about 4K entries from the Wikipedia[3] website and then generated some pre-defined abbreviation patterns like those used in Park and Byrd (2001). Besides these, we used a Web-based transliteration model to recognize a transliterated term (Section 2.5).

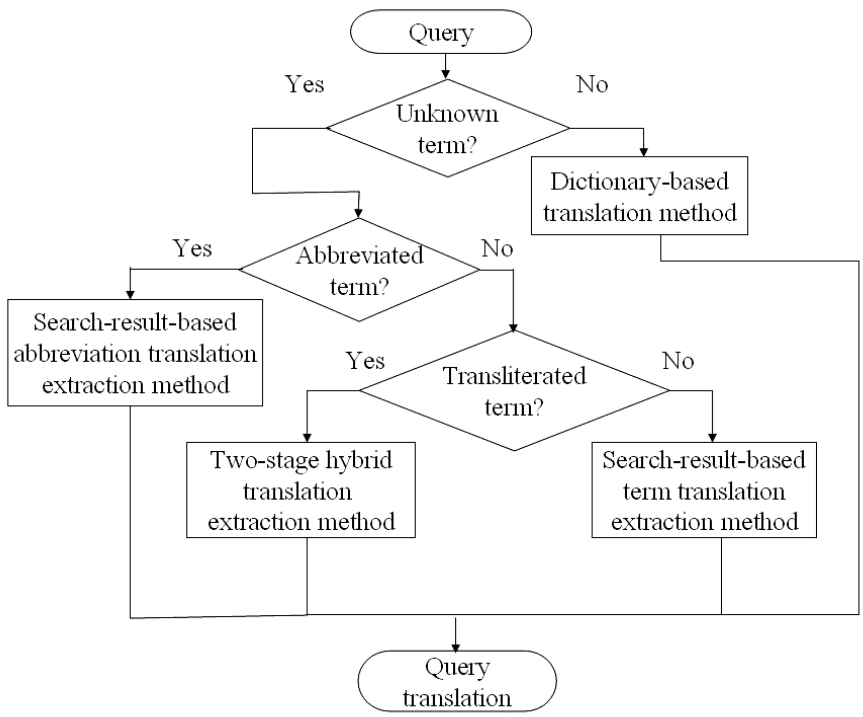


***Figure 1. The process of our search-result-based query translation method***

## 2.3 Search-Result-Based Term Translation Extraction Method

In this section, we will describe Cheng *et al.*'s [2004] search-result-based term translation extraction method, which explores search-result pages utilizing co-occurrence relation and contextual information for extraction of translations of unknown query terms.

**(1) Chi-square Test Method**

On the basis of co-occurrence analysis, chi-square test ($\chi 2$) is adopted to estimate semantic similarity between the source term *E* and the target translation candidate *C*. The similarity measure is defined as:

[3] http://en.wikipedia.org/wiki/List_of_acronyms_and_initialisms

$$S_{\chi^2}(E, C) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}, \quad (1)$$

where $a$, $b$, $c$ and $d$ are the numbers of pages retrieved from search engines by submitting Boolean queries: "*E* and *C*", "*E* and not *C*", "not *E* and *C*", and "not *E* and not *C*", respectively; $N$ is the total number of pages, i.e., $N = a + b + c + d$.

**(2) Context-Vector Analysis Method**

Due to the property of Chinese-English mixed texts often appearing in Chinese pages, the source term *E* and the target translation candidate *C* may share common contextual terms in the search-result pages. The similarity between *E* and *C* is computed based on their context feature vectors $E_{cv}$ and $C_{cv}$ in the vector-space model. The conventional tf-idf weighting scheme for each feature term $t_i$ in $E_{cv}$ and $C_{cv}$, $E_{cv} = \langle w_{e1}, w_{e2}, \ldots, w_{em} \rangle$, and $C_{cv} = \langle w_{c1}, w_{c2}, \ldots, w_{cm} \rangle$, is used and defined as:

$$w_{t_i} = \frac{f(t_i, p)}{\max_j f(t_j, p)} \times \log(\frac{N}{n}), \quad (2)$$

where $f(t_i, p)$ is the frequency of term $t_i$ in the search-result page $p$, $N$ is the total number of Web pages, and $n$ is the number of the pages containing $t_i$. Finally, we use the cosine measure to estimate the similarity between *E* and *C* as follows:

$$S_{CV}(E, C) = \frac{\sum_{i=1}^{m} w_{e_i} \times w_{c_i}}{\sqrt{\sum_{i=1}^{m} (w_{e_i})^2 \times \sum_{i=1}^{m} (w_{c_i})^2}}. \quad (3)$$

## 2.4 Search-Result-Based Abbreviation Translation Extraction Method

To effectively extract correct translations for infrequent abbreviated terms, we propose an integrated method in which an abbreviated term is transformed to its full name first, and then we extract the correct translation of the full name using the search-result-based term translation extraction method described above (Section 2.3). In the following, we describe two new proposed methods exploiting search results to extract full names for English and Chinese abbreviations, respectively.

### 2.4.1 Extracting Full Names for English Abbreviations

To deal with the full names for a given English abbreviation, we designed an efficient process of identifying full names, which consists of three major steps based on the hybrid text mining approach proposed by Park and Byrd [2001]. First, we use the contextual terms around an abbreviated term in the search results to extract possible full name candidates. Second, we use occurrence frequency and Part-of-Speech (POS) information of full name candidates to filter out some impossible candidates. Finally, we propose a simple adaptive co-occurrence model

which utilizes several different augmenting and decaying factors in selecting the best full name candidate. More details are described in the following.

**(1) Identifying Full Name Candidates**

To solve the problem of identifying full names without sufficient texts [Park and Byrd 2001], we take advantage of Web search results as a corpus. Our idea is to take the given abbreviated term as a search term to fetch the top 200 search result snippets from Google. To extract possible full name candidates by exploring the search result snippets, we utilize contextual information of the abbreviated term in the snippets. These full name candidates must appear in the same snippets with the abbreviated term, and should have a minimum word length between $|A|\times 2$ and $|A|+5$, where $|A|$ is the length of characters of the abbreviated term. In addition, to select more reliable full name candidates, we put a constraint on the identification process in which the first character of the first word of each full name candidate should match the first character of the abbreviated term.

**(2) Filtering Impossible Full Name Candidates**

To reduce computation time while extracting many full name candidates, we first select the top 20 frequent full name candidates and then filter out some impossible candidates whose first word or last word are prepositions, be-verbs, modal verbs, conjunctions, or pronouns [Park and Byrd 2001].

**(3) Selecting Best Full Name Candidate**

To select the best full name candidates, we propose an adaptive co-occurrence model by employing mutual information as well as four augmenting or decaying factors to compute the similarity between an abbreviated term $A$ and its full name candidates $F_C$.

**(A) Mutual Information:** In this step, mutual information is used to compute the similarity between an abbreviated term $A$ and its full name candidate $F_C$. Mutual information is defined as follows:

$$MI(A, F_C) = P(A, F_C) \times \log\left(\frac{P(A, F_C)}{P(A) \times P(F_C)}\right). \quad (4)$$

Here $P(A, F_C)$ is the probability of co-occurrence of $A$ and $F_C$. $P(A)$ and $P(F_C)$ are the probabilities of occurrence of $A$ and $F_C$ in the Web, respectively. We can get the occurrence frequencies from search engines by submitting queries: "$A$", "$F_C$", and "$A$ and $F_C$", respectively.

**(B) Syntactic Cues:** To augment the identification of full names, we utilize the information of orthographic and syntactic structure. $N_{SC}$ indicates the number of abbreviation-full-name pairs appearing in the same snippets. Several frequent patterns of abbreviation-full-name pair are used as the syntactic cues [Park and Byrd 2001], including:

- abbreviation (full name)
- full name (abbreviation)
- abbreviation, **or** full name
- full name, **or** abbreviation
- full name, abbreviation **for short**
- abbreviation … **stands/short/acronym** …full name

**(C) Similarity of Character:** To further determine correct full names, we add another augmenting factor to estimate the similarity between an abbreviated term and its full name candidates by adopting a fast and simple character matching method. We use two kinds of character matching: (1) first-letter matching is used to compute the total number $N_F$ of matching the first letter of each word in the full name candidate $F_C$ with each character in the abbreviated terms, and (2) non-first-letter matching is used to computer the total number $N_{NF}$ of matching the non-first letters of each word in the $F_C$ with each character in $A$. The score of character matching of $A$ and $F_C$ is defined as:

$$Overlap(A, F_C) = \alpha * N_F + (1-\alpha) * N_{NF} . \quad (5)$$

Here, the weighting parameter $\alpha$ is empirically set to 0.8. Basically, the first-letter matching should be reasonably assigned higher weight for each matching pair. The character similarity is defined as follows:

$$CharSim(A, F_C) = \frac{Overlap(A, F_C)}{|A|}, \quad (6)$$

where $|A|$ is the number of characters of the abbreviated term $A$.

**(D) Difference of Length:** The number $N_{LD}$ to represent the difference between character length $|A|$ of the abbreviated term $A$ and word length $|F_C|$ of the corresponding full name candidate $F_C$ as a decaying factor. $N_{LD}$ is defined as follows:

$$N_{LD} = \| A | - | F_C \| . \quad (7)$$

**(E) Number of Stop Words:** The number $N_{SW}$ of stop words in the full name candidate $F_C$ is also used as a decaying factor.

**(F) Adaptive Co-occurrence Model:** We adaptively integrate the above two augmenting and two decaying factors into the basic co-occurrence model to compute the similarity between $A$ and $F_C$. Our adaptive co-occurrence model is defined as follows:

$$S_{AC}(A, F_C) = \frac{MI(A, F_C) \times F_{Augument}}{F_{Decay}}, \quad (8)$$

where the augmenting factor $F_{Augument}$ is integrated as

$$F_{Augument} = CharSim(A, F_C) \times (\beta_1 + N_{SC}) ; \quad (9)$$

and the decaying factor $F_{Decay}$ is integrated as

$$F_{Decay} = (\beta_2 + N_{LD} + N_{SW}). \quad (10)$$

To avoid the product being zero, here, $\beta_1$ and $\beta_2$ are the adaptable parameters and set to 1 heuristically.

### 2.4.2 Extracting Full Names for Chinese Abbreviations

Due to language differences between Chinese and English, such as no space delimitation between Chinese words, it is more difficult to identify the full name for a given Chinese abbreviated term. Therefore, we designed a method slightly different from the method of extracting English full names described above. Our Chinese full name extraction method consists of three major steps. First, the possible full name candidates are extracted by using the PAT-tree-based keyword extraction method proposed by Chien [1997]. Second, we use the character similarity between an abbreviated term and its full name candidates to filter out some impossible candidates. Finally, to select the correct Chinese full name, we use the adaptive co-occurrence model (Equation (8)) but slightly modify the decaying factors. The following description will explain the different points in more details.

**(1) Identifying Full Name Candidates**

To identify the possible full name candidates for a given Chinese abbreviated term $A$, we adopt a PAT-tree-based keyword extraction method [Chien 1997] to extract Chinese phrases in the search results related to the abbreviated term $A$ as full name candidates. In addition, to select more reliable full name candidates, we put a length constraint on the candidates. These candidates should have more than ($|A|$ +2) characters, where $|A|$ is the number of characters of $A$.

**(2) Filtering Impossible Full Name Candidates**

According to our observations, the Chinese full name candidates extracted by the PAT-tree-based keyword extraction method generally have higher reliability. Thus, we just use Equations (5) and (6) with a threshold of character similarity to filter out some impossible candidates.

**(3) Selecting Best Full Name Candidate**

Like the above method of selecting the best English full name candidates, we still use the proposed adaptive co-occurrence model (Equation (8)) to select the best Chinese full name candidates. Please note, though, that the processing of augmenting/decaying factors is a little different. For example, we remove the decaying factor of stopword number since most stopwords seldom appear in Chinese full names. Some different points will be described below.

**(A) Syntactic Cues:** We also manually choose several syntactic patterns of Chinese abbreviation- full name pairs as the augmenting factor:

- abbreviation (full name)
- full name (abbreviation)
- abbreviation, 或 full name
- full name, 或 abbreviation
- abbreviation … 代表/簡稱/縮寫 …full name

Here the Chinese cues "或", "代表", "簡稱", "縮寫" correspond to the English words "or", "present", "short", and "acronym", respectively.

**(B) Similarity of Character:** First, we use the Chinese POS tagger to segment full name candidates. Then, we take character similarity (Equation (5) and (6)) as an augmenting factor.

**(C) Difference of Length:** Due to the fact that there is no space delimitation between Chinese words, we adopt a Chinese POS tagger[4] to do word segmentation for full name candidates. Then, we use the number $N_{LD}$ to represent the difference between character length $|A|$ of the abbreviated term $A$ and word length $|F_C|$ of the corresponding full name candidate $F_C$; this is considered a decaying factor (Equation (7)).

**(D) Adaptive Co-occurrence Model:** We adopt the same adaptive co-occurrence model (Equation (8)) with two augmenting factors and one decaying factor to compute the similarity between $A$ and $F_C$. The augmenting factors are the same as Equation (9), but the decaying factor in Equation (10) is modified adaptively by removing the stopword number as:

$$F_{Decay} = (\beta_3 + N_{LD}). \tag{11}$$

To avoid the product being zero, here $\beta_3$ is an adaptable parameter and set to 1, heuristically.

## 2.5 Search-Result-Based Transliteration Name Extraction Method

To improve the performance of unknown term translation extraction for infrequent proper names, we consider integrating name transliteration techniques into the process of translation extraction in order to filter out impossible transliterated name candidates. Our idea is to first extract terms from the search-result snippets as translation candidates (see Section 2.3), and then filter out impossible transliterated name candidates based on the name transliteration model (described in Section 2.5.2). Therefore, in this section we propose a two-stage hybrid translation extraction method, a Web-based transliteration model to deal with transliteration mapping between an English proper name and its corresponding Chinese, and a Web-based

[4] http://ckipsvr.iis.sinica.edu.tw/demo.htm, which is a Chinese POS tagger developed by Chinese Knowledge and Information Processing group of Academia Sinica.

unsupervised learning algorithm to automatically collect diverse English-Chinese transliteration name pairs from Web search results for transliteration model training (Section 2.5.3).

### 2.5.1 Two-Stage Hybrid Translation Extraction

Our proposed two-stage hybrid translation extraction method is composed of two major steps. First, we use the search-result-based translation extraction method (Section 2.3) to extract $k$ ($k$ = 20) terms with higher similarity scores as transliteration candidates. Second, some impossible candidates included in general-purpose bilingual dictionaries are filtered out, and then the rest of the candidates are ranked according to transliteration similarity with the source proper name, which is computed based on the proposed Web-based transliteration model below (Equation (15)).

### 2.5.2 Filtering Impossible Candidates Using Web-Based Transliteration Model

**(A) English Syllable Segmentation:** Wan and Verspoor [1998] have developed a fully rule-based algorithm to transliterate English proper names into Chinese names. We simplify their syllabification techniques to generate a few simple heuristic rules of segmenting an English name into a sequence of syllables. Each English syllable is regarded as an English transliteration unit (ETU) in this work and has at most one corresponding character of the Chinese transliterated name. Initially, we used only five rules for English syllable segmentation listed below:

- a, e, i, o, u are vowels, and y is also regarded as a vowel if it appears behind a consonant. All other letters are consonants.
- Separate two consecutive vowels except the following cases: ai, au, ee, ea, ie, oa, oo, ou, etc.
- Separate two consecutive consonants except the following cases: bh, ch, gh, ph, th, wh, ck, cz, zh, zk, ng, sc,ll, tt, etc.
- l, m, n, r are combined with the prior vowel only if they are not followed by a vowel.
- A consonant and a following vowel are regarded as an ETU.

For example, “Nokia” (諾基亞) is segmented into three ETUs “no”, “ki”, and “a”, and “Epson” (愛普生) is segmented into three ETUs “e”, “p”, and “son”. Currently, although some English names may be segmented incorrectly, it is easy to manually update new rules to improve English syllable segmentation.

**(B) Web-based Transliteration Model:** To avoid double errors of converting English phonetic representation to Chinese Pinyin and from Pinyin to Chinese characters, in this work, we adopted direct orthographic mapping for name transliteration. We use the probability $P(e_i, c_i)$ to estimate the possibility of the mapping between an ETU $e_i$ and a Chinese character $c_i$. Additionally, to build an efficient online name transliteration model, we propose a more simple transliteration

model. Our Web-based transliteration model is called forward-syllable-mapping transliteration model:

$$S_{FSM}(E,C)=\frac{P_{FSM}(E,C)}{D(E,C)}, \tag{12}$$

where $P_{FSM}(E, C)$ is the co-occurrence probability of $E$ and $C$ and defined as

$$P_{FSM}(E,C)\approx \prod_{i=1}^{\min(m,n)}[(1-\gamma_1)P(e_i,c_i)+\gamma_1], \tag{13}$$

and $\gamma_1$ is the smoothing weight. The decaying factor $D(E, C)$ indicates the number of syllable difference between an English name $E$ and a Chinese transliterated name $C$ and is defined as:

$$D(E,C)=\varepsilon+|m-n|. \tag{14}$$

Here $\varepsilon$ is a decaying parameter, $m$ is the total number of ETUs, and $n$ is the total number of Chinese characters.

To improve incorrect transliteration mapping between ETUs and Chinese characters while an English-Chinese transliterated name pair with different numbers of transliteration unit, we propose the reverse-syllable-mapping transliteration model to assist in learning more correct mapping, which is defined below:

$$S_{RSM}(E,C)=\frac{P_{RSM}(E,C)}{D(E,C)}, \tag{15}$$

where

$$P_{RSM}(E,C)\approx\begin{cases}\prod_{i=m-n+1}^{m}[(1-\gamma_2)P(c_{i-(m-n)},e_i)+\gamma_2], & m\geq n;\\ \prod_{i=n-m+1}^{n}[(1-\gamma_2)P(c_i,e_{i-(n-m)})+\gamma_2], & m<n.\end{cases} \tag{16}$$

Here $\gamma_2$ is the smoothing weight and $D(E, C)$ is the same as Equation (14).

Our alternative transliteration model will combine the forward-syllable-mapping and reverse-syllable-mapping transliteration model, which is called **mixed-syllable-mapping transliteration model**, and defined as:

$$S_{MSM}(E,C)=\sqrt{S_{FSM}(E,C)\times S_{RSM}(E,C)}. \tag{17}$$

### 2.5.3 Web-Based Unsupervised Learning Algorithm

To deal with the problems of the diversity of Chinese transliterated names to English proper names, we intend to take advantage of abundant language-mixed texts on the Web to collect various English-Chinese transliterated name pairs from the Web and build a an effective online transliteration model. Thus, we designed an unsupervised learning process for English-Chinese transliterated name mapping. The process is composed of three main stages:

extraction of Chinese transliterated names, extraction of English original names, and learning of transliterated name mapping. More details are described below and the unsupervised learning algorithm is illustrated as well in Figure 2.

**Web-based Unsupervised Learning Algorithm for Collecting English-Chinese Transliteration Pairs and Training a Transliteration Model**

Input: initial transliterated name pair set $V_{ec}$ and a general-purpose bilingual dictionary $D$.

Output: updating $V_{ec}$ and a transliteration model $T$.

1 Extraction of Chinese transliterated names: select a transliterated name pair ($E$, $C$) from $V_{ec}$, two characters from the Chinese name $C$ as seed characters $V_c$, and two corresponding English syllables from the English name $E$ as seed ETUs $V_e$.

1.1 Search-result crawling: send the two selected Chinese seed characters $V_c$ to a search engine and get search-result pages.

1.2 Chinese transliterated name identification: use a Chinese POS tagger to find unknown terms in the search-result pages, and then take the unknown terms containing the two seed characters $V_c$ as potential Chinese transliterated names $C_p$.

2 Extraction of English original names: for each potential Chinese transliterated name $C_p$ in $V_c$,, perform the following sub-steps:

2.1 Two-Stage hybrid translation extraction

2.1.1 English name candidate extraction: use search-result-based term translation extraction method to find English name candidates (see Section 2.3).

2.1.2 English name candidate filtering: first filter out impossible English name candidates included in $D$; second, compute transliteration mapping scores based on the English syllable segmentation rules and the name transliteration model $T$; third, choose the candidates with the highest scores as the possible English original names. Update $V_{ec}$ by adding the new transliterated name pairs extracted.

2.2 Learning of transliterated name mapping: update $T$ by computing the scores of transliterated name mapping of the new extracted transliterated name pairs (Equation (17)).

3 Repeat from step1 until the desired number of transliteration pairs is reached.

***Figure 2. Web-based unsupervised learning algorithm for collecting English-Chinese transliterated name pairs and building a transliteration model***

**(1) Extraction of Chinese Transliterated Names:** Xiao *et al*. [2002] have proposed a bootstrapping algorithm that uses only five frequent Chinese transliterated characters as initial seed character set: {阿, 爾, 巴, 斯, 基} to automatically collect over 100,000 Chinese transliterated names by utilizing search-result pages. Inspired by this work, we further propose a bootstrapping algorithm to automatically find English-Chinese transliterated name pairs from search-result pages. Initially, we need at least one English-Chinese transliterated name pair containing two frequent Chinese transliterated characters as seed transliteration pair set $V_{ec}$, *e.g*., $V_{ec}$ = {(Bush, 布希)}. We select two Chinese characters from the Chinese name of the seed pair, and then send them to search engines for getting search-results pages. To efficiently extract more Chinese transliterated names from search-result pages, we use the CKIP tagger (Section 2.4.2), which is a representative Chinese POS tagger and performs well in segmenting Chinese texts into meaningful words and extracting unknown words.

**(2) Extraction of English Original Names:** We use the proposed two-stage hybrid translation extraction method described above (Section 2.5.1) to find possible English original names.

**(3) Learning of Transliterated Name Mapping:** On the basis of the rules of English syllable segmentation, we will gradually train an English-Chinese name transliteration model by computing the scores of the transliterated name mapping of the new extracted transliterated name pairs (Equation (17)).

## 3. Experimental Results

We conducted the following experiments to evaluate the performance of our proposed search-result-based abbreviation translation extraction method and two-stage hybrid translation extraction method.

**Evaluation Metric:** For the following experiments on full name identification of abbreviations and translation of abbreviations, the average top-*n* inclusion rate is adopted as a metric. For a set of abbreviated terms to be expanded/translated, its top-*n* inclusion rate was defined as the percentage of the abbreviated terms whose correct full names/translations could be found in the first *n* extracted full name candidates/translation candidates [Cheng *et al*. 2004].

**Correct Translation / Transliteration:** The correct translation / transliteration or correct definition is judged by us according to more popular sense in general cases.

### 3.1 Evaluation for the Search-Result-Based Abbreviation Translation Extraction Method

In this experiment, we intend to compare the performance of our proposed search-result-based abbreviation translation method with that of the search-result-based term translation extraction method proposed by Cheng *et al*. [2004].

### 3.1.1 Translation Extraction Results for English Abbreviations

**Test data:** Four test sets of English abbreviated terms are prepared in the following.

- FA-Dreamer-E: 28 frequent English abbreviated terms which have correct Chinese translations were manually selected from about 20K frequent queries with occurrence frequency over 10 in the Dreamer query log[5] which contains 228,566 unique queries. (The partial test data is listed in Appendix).
- IA-Dreamer-E: 27 infrequent English abbreviated terms (frequency < 3 in Dreamer query log) which have correct Chinese translations were manually selected from infrequent English queries in the Dreamer query log (about 40K entries). (The partial test data is listed in Appendix).
- FA-Wiki-E: 62 popular English abbreviated terms which have correct Chinese translations were manually selected from Wikipedia abbreviation list containing about 4k entries (Section 2.2). (The partial test data is listed in Appendix).
- RA-Wiki-E: 25 English abbreviated terms which have correct Chinese translations were randomly selected from Wikipedia abbreviation list due to the list without frequency information. (The partial test data is listed in Appendix).

**(1) Results for English Full Name Extraction**

Table 1 shows that our full name extraction method is effective for the test abbreviated terms with various subjects. Our method can achieve the top-1 inclusion rate of over 85% and the top-5 inclusion rate of over 92% for all test sets. Different from existing methods, our full name extraction method is very promising even for infrequent abbreviated terms by utilizing search results from Web search engines. However, some errors still result from the problem of data sparseness. For example, given the abbreviated term "MPEG", its correct full name "Motion Picture Experts Group" might appear quite rarely in the top 200 search results snippets. Therefore, the correct full name is filtered out by the filtering step and this causes trouble in extracting incorrect full names.

***Table 1. Inclusion rates on full name extraction for different test sets of English abbreviated queries***

| Test Set | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| FA-Dreamer-E | 93% | 96% | 96% |
| IA-Dreamer-E | 85% | 96% | 96% |
| FA-Wiki-E | 90% | 94% | 94% |
| RA-Wiki-E | 88% | 88% | 92% |

[5] http://www.dreamer.com.tw, which was a popular Chinese search engine and is closed now.

**(2) Search-Result-based Abbreviation Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method**

Tables 2 to 5 show that the proposed search-result-based abbreviation translation extraction method actually performs better than the previous search-result-based translation extraction method proposed by Cheng *et al*. For example, for the infrequent English abbreviated queries from the Dreamer query log, the search-result-based abbreviation translation extraction method achieve the top-1 inclusion rate of 48% (see Table 3) but the search-result-based translation extraction method achieve the top-1 inclusion rate of 0%. Given the example query "ISS", the search-result-based term translation extraction method cannot obtain the correct Chinese translation "國際太空站" among the top five extracted candidates. However, our search-result-based abbreviation translation extraction method can extract the correct full name "International Space Station", and then extract correct Chinese translation "國際太空站" via the full name "International Space Station". As mentioned in Section 2.1, the reason might be that the abbreviated terms are semantically more ambiguous and co-occur relatively infrequently with the correct translations of their full names.

**(3) Linear Combination Results**

To further improve the performance of our search-result-based abbreviation translation extraction method, we intuitively intend to combine our method and Cheng *et al*.'s method. We expect that such a combination would make both methods mutually complementary by extracting translations from abbreviations and their full names simultaneously. Tables 2 to 5 show that the linear combination method is effective in improving the top-5 inclusion rate. For example, for the abbreviated query "AOL", its correct full name "America Online" is correctly extracted via our abbreviation expansion method. It fails to find the correct translation among the top five extracted candidates using our search-result-based abbreviation translation method, but the correct translation "美國線上" can be ranked at third place using the linear combination method.

***Table 2. Inclusion rates on translation of frequent English abbreviations from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 43% | 54% | 57% |
| Search-result-based Abbreviation Translation Extraction Method | 75% | 82% | 86% |
| Linear Combination | 71% | 82% | 93% |

***Table 3. Inclusion rates on translation of infrequent English abbreviations from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 0% | 19% | 19% |
| Search-result-based Abbreviation Translation Extraction Method | 48% | 59% | 63% |
| Linear Combination | 44% | 63% | 67% |

***Table 4. Inclusion rates on translation of frequent English abbreviations from Wikipedia abbreviation list***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 24% | 40% | 44% |
| Search-result-based Abbreviation Translation Extraction Method | 65% | 79% | 79% |
| Linear Combination | 65% | 77% | 81% |

***Table 5. Inclusion rates on translation of randomly selected English abbreviations from Wikipedia abbreviation list***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 24% | 36% | 36% |
| Search-result-based Abbreviation Translation Extraction Method | 64% | 76% | 76% |
| Linear Combination | 64% | 72% | 80% |

### 3.1.2 Translation Extraction Results for Chinese Abbreviations

**Test data:** Two test sets of Chinese abbreviated terms are prepared in the following.

- FA-Dreamer-C: 35 frequent Chinese abbreviated terms with correct English translations were manually selected from about 20K frequent queries with occurrence frequency over 10 in the Dreamer query log. (The partial test data is listed in Appendix).
- IA-Dreamer-C: 28 infrequent Chinese abbreviated terms (frequency < 3 in Dreamer query log) with correct English translations were manually selected from infrequent Chinese queries in the Dreamer query log (about 115K entries). (The partial test data is listed in Appendix).

**(1) Results for Chinese Full Name Extraction**

Table 6 shows that our Chinese full name extraction method is effective and can achieve top-1 inclusion rate of over 86% for the two test sets. We observed that some errors resulted from

incorrect matching between the abbreviated query terms and their highly related full name candidates in the search results. For example, given the abbreviated term "中影" (Central Motion Picture Corporation), our method extracted the incorrect full name "中國電影" (Chinese Movie) at first place. Since the correct full name "中央電影公司" co-occurs infrequently with the abbreviated query term "中影" in the search results, it can't be extracted by the PAT-tree-based keyword extraction method. As a result, our method extracted the incorrect full name "中國電影" because the abbreviated term "中影" and the incorrect full name candidate "中國電影" have stronger correlation in the search results and higher character similarity.

***Table 6. Inclusion rates on full name extraction for two test sets of Chinese abbreviated queries***

| Test Set | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| FA-Dreamer-C | 94% | 100% | 100% |
| IA-Dreamer-C | 86% | 89% | 89% |

**(2) Performance Comparison between Search-Result-based Abbreviation Translation Extraction Method and Term Translation Extraction Method**

Tables 7 and 8 show that, for the extraction of Chinese abbreviation translation, the proposed search-result-based abbreviation translation extraction method still performs better than the previous search-result-based translation extraction method proposed by Cheng *et al*. For example, for the infrequent Chinese abbreviated queries from the Dreamer query log, Cheng *et al*.'s method performs very poorly with a top-5 inclusion rate of 4%, but our method achieves great improvement with the top-5 inclusion rate of 29%. For example, given the Chinese abbreviated query "國安局", Cheng *et al*.'s method cannot obtain the correct English translation "National Security Bureau" among the top five extracted candidates. However, our method can extract the correct Chinese full name "國家安全局", and then extract the correct English translation "National Security Bureau", which is ranked at second place.

In addition, Table 8 shows that the linear combination method just achieves the same performance as our method, and is unable to further improve the top-*n* inclusion rates. In fact, we need larger amounts of test data to determine the effectiveness using the linear combination method in the future.

***Table 7. Inclusion rates on translation of frequent Chinese abbreviations from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 17% | 46% | 54% |
| Search-result-based Abbreviation Translation Extraction Method | 40% | 66% | 71% |
| Linear Combination | 49% | 63% | 71% |

***Table 8. Inclusion rates on translation of infrequent Chinese abbreviations from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 4% | 4% | 4% |
| Search-result-based Abbreviation Translation Extraction Method | 11% | 21% | 29% |
| Linear Combination | 11% | 21% | 29% |

## 3.2 Evaluation for the Two-Stage Hybrid Translation Extraction Method

The following two experiments are focused on the evaluation of the performance of extracting translations for infrequent unknown English and Chinese proper names, respectively, using the proposed mixed-syllable-mapping transliteration model and the two-stage hybrid translation extraction method.

### 3.2.1 Translation Extraction Results for English Proper Names

**Test data:** Two test sets of unknown English proper names are prepared, including:

- FP-Dreamer-E: 28 frequent unknown English proper names are manually selected from the 169 unknown terms out of the 430 frequent English queries in the Dreamer query log. (The partial test data is listed in Appendix).
- IP-Dreamer-E: 41 infrequent unknown English proper names (frequency < 3 in the query log) are manually selected from the Dreamer query log. (The partial test data is listed in Appendix).

**(1) Two-Stage Hybrid Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method**

According to the results shown in Tables 9 and 10, we can obtain the following findings. For the two test sets, the proposed two-stage hybrid translation extraction method made great improvements compared with the search-result-based translation extraction method and the general name transliteration method [Wan and Verspoor 1998; Knight and Graehl 1998; Lin

and Chen 2002; Virga and Khudanpur 2003; Gao *et al*. 2004; Li *et al*. 2004]. In this work, we just use our proposed transliteration model as a "Name Transliteration" method for performance comparison. For example, the two-stage hybrid translation extraction method can achieve the top-1 inclusion rate of 41% (Table 10) for infrequent unknown English proper names, but the search-result-based translation extraction method only achieved 17%. The main reason is that most of the incorrect translation candidates extracted via the search-result-based translation extraction method can be filtered out based on our mixed-syllable-mapping transliteration model. For example, given the English proper name "Pamela", the correct Chinese transliterated name "潘蜜拉" can be extracted and ranked at second place (see Table 11).

**(2) Linear Combination Results**

Tables 9 and 10 also demonstrate that the simple linear combination method obtained slight improvement on transliterated name performance since the general name transliteration method is still limited in generating correct transliteration candidates. However, note that for many English-Chinese transliteration pairs with different numbers of transliteration units, the mixed-syllable-mapping transliteration model is still effective to learn correct transliteration mapping between English syllables and Chinese characters.

***Table 9. Inclusion rates on translation of frequent unknown English proper names from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 32% | 71% | 82% |
| Name Transliteration | 11% | 18% | 21% |
| Linear Combination | 32% | 50% | 86% |
| Two-Stage Hybrid Translation Extraction Method | 61% | 64% | 68% |

***Table 10. Inclusion rates on translation of infrequent unknown English proper names from Dreamer query log***

| Translation Extraction Method | Inclusion Rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 17% | 32% | 37% |
| Name Transliteration | 15% | 15% | 17% |
| Linear Combination | 17% | 37% | 44% |
| Two-Stage Hybrid Translation Extraction Method | 41% | 46% | 46% |

***Table 11. Effective results of translation extraction using the two-stage hybrid translation extraction method (underlined terms indicate correct translation)***

| Test Query | Translation Extraction Method | Top 5 Translation Candidates |
|---|---|---|
| Pamela | Search-result-based Translation Extraction Method | 最後發表，發表文章，派米拉路，發表，討論區 |
| | Name Transliteration | 帕麥拉，帕亞拉，帕雲拉，帕麥斯，柏麥拉 |
| | Linear Combination | 最後發表，發表文章，帕麥拉，派米拉路，發表 |
| | Two-Stage Hybrid Translation Extraction Method | 彭美拉，潘蜜拉，派米拉路，安德森，尤德夫人 |

### 3.2.2 Translation Extraction Results for Chinese Proper Names

**Test data:** Two test sets of unknown Chinese proper names are prepared, including:

- FP-Dreamer-C: 28 frequent unknown Chinese proper names are obtained from the transliterated terms of the frequent unknown English proper name set FP-Dreamer-E (described in Section 3.2.1). (The partial test data is listed in the Appendix).
- IP-Dreamer-C: 41 infrequent unknown Chinese proper names are obtained from the transliterated terms of the infrequent unknown English proper name set IP-Dreamer-E (described in Section 3.2.1). (The partial test data is listed in the Appendix).

**(1) Two-Stage Hybrid Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method**

Table 12 shows that our two-stage hybrid translation extraction method obtains the top-1 inclusion rate of 64%. Surprisingly, it performs worse than the search-result-based translation extraction method at 70%. This means that our candidate filtering method based on our trained Web-based transliteration model is unable to improve the performance of extracting translations for frequent unknown Chinese proper names in Web queries. We will investigate the possible reasons in the following discussion. However, for the test set of infrequent unknown Chinese proper names, the two-stage hybrid translation extraction method made effective improvements compared with the search-result-based translation extraction method (Table 13). For example, the two-stage hybrid translation extraction method can achieve the top-1 inclusion rate of 46% for infrequent unknown Chinese proper names, whereas the search-result-based translation extraction method only achieved 27%. It shows that most of the incorrect translation candidates extracted via the search-result-based translation extraction method can be filtered out using our mixed-syllable-mapping transliteration model. For

example, given the Chinese transliterated name “艾立克”, its correct English original name “Eric” can be extracted and ranked at first place (Table 14).

***Table 12. Inclusion rates on translation of frequent unknown Chinese proper names from Dreamer query log***

| Translation Extraction Method | Inclusion rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 71% | 89% | 93% |
| Name Transliteration | 14% | 21% | 25% |
| Linear Combination | 71% | 82% | 86% |
| Two-Stage Hybrid Translation Extraction Method | 64% | 71% | 75% |

***Table 13. Inclusion rates on translation of infrequent unknown Chinese proper names from Dreamer query log***

| Translation Extraction Method | Inclusion rates | | |
|---|---|---|---|
| | Top-1 | Top-3 | Top-5 |
| Search-result-based Translation Extraction Method | 27% | 44% | 51% |
| Name Transliteration | 12% | 22% | 22% |
| Linear Combination | 27% | 47% | 57% |
| Two-Stage Hybrid Translation Extraction Method | 46% | 51% | 51% |

***Table 14. Effective results of translation extraction using the two-stage hybrid translation extraction method (underlined terms indicate correct translation)***

| Test Query | Translation Extraction Method | Top 5 Translation Candidates |
|---|---|---|
| 艾立克 | Search-result-based Translation Extraction Method | Blog, Doll Edward, card, ebay, Eric Benet |
| | Name Transliteration | Elic, Eddoc, Alic, Addoc, Eric |
| | Linear Combination | Blog, Doll Edward, Elic, card, ebay |
| | Two-Stage Hybrid Translation Extraction Method | Eric, Alex, Eric idle, Clapton Eric, Eric Clapton Tears, KKBox Eric |

**(2) Discussion**

According to our further analyses of the results shown in Tables 12 and 13, we obtain the following interesting findings.

- Our test set FP-Dreamer-C (frequent unknown Chinese transliterated terms) contains a number of company names, *e.g.*, "銳跑" (Reebok) and "新浪" (Sina). In fact, these Chinese characters like "銳", "跑", and "浪" are rarely used as transliterated characters in general cases. Thus, these characters are certainly difficult to be matched with those possibly correct ETUs since they have never appeared in the training data of our collected English-Chinese transliterated name pairs from search-result pages.
- The probabilities of some correct transliteration mapping between Chinese characters and English ETUs are lower than those of incorrect transliteration mapping trained from incorrect or partial matching transliteration pairs. However, our training data of about 10k potential transliterated name pairs extracted via our Web-based unsupervised learning algorithm should contain a number of incorrect transliteration mapping pairs and still be insufficient to build a good-quality transliteration model.
- The search-result-based term translation extraction method perform well for the test set of frequent unknown Chinese proper names while our two-stage hybrid translation extraction method is effective in improving the translation performance for infrequent unknown Chinese proper names. Therefore, we consider adding the information of term occurrence frequency in the query log into the process of unknown term translation. For a query with frequent Chinese proper names in the query log, we can use the previous search-result-based term translation extraction method to translate it. On the other hand, for queries with infrequent Chinese transliterated terms, we can use the proposed two-stage hybrid translation extraction method to translate them.

However, utilizing Web search results to translate unknown terms would lead to only partial representative candidates, which are the most popular ones. Therefore, we should continuously collect much more English-Chinese transliterated name pairs for training a better transliteration model in the future, and at the same time improve the techniques of extracting and filtering English name candidates to further collect larger amounts of correct transliterated name pairs for building a high quality transliteration model. In addition, there are still a number of cases which are difficult to be dealt with by using the simple mixed-syllable-mapping transliteration model and need to be further improved in the future.

## 4. Related Work

In previous works on identifying full names of abbreviations, AFP (Acronym Finding Program) [Taghva and Gilbreth 1995] used free texts to find English abbreviations and their full names. Park and Byrd [2001] used contextual information around abbreviations to extract potential full name candidates based on their pre-defined rules. However, these methods might suffer from the problem of insufficient texts. Our proposed method exploiting search results can extract English full names for abbreviations in various domains, and then effectively

extract correct Chinese translations via their full names.

Also, Leah *et al*. [2000] tried to find full name candidates from a small number of Web pages, and they used lots of syntax rules to select full name candidates of English acronyms. Instead of using many syntax rules, we propose an adaptive co-occurrence model to select the best full name candidates based on the co-occurrence relation and the integration of several augmenting and decaying factors.

For name transliteration between Latin-alphabet languages and some Asian languages with different writing forms, such as English and Chinese, researchers have proposed phoneme-based mapping techniques [Knight and Graehl 1998; Lin and Chen 2002; Meng *et al*. 2001]. Lin *et al*. [2003] proposed a statistical transliteration model and apply the model to extract English proper names and their Chinese transliterated names in a parallel corpus with high average precision and recall rates. However, Li *et al*. [2004] pointed out that the transliteration precision of the phoneme-based approaches could be limited by two main constraints. First, Latin-alphabet foreign names from different origins have different phonic rules, such as French and English. Second, transforming English words to Chinese characters will need two steps: transforming from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters. Two cascaded transforming steps may cause double errors. To avoid this problem, we propose a Web-based mixed-syllable-mapping transliteration model for dealing with online English-Chinese name transliteration based on the concept of direct orthographic mapping.

Both Cheng *et al*. [2004] and Zhang and Vines [2004] have exploited language-mixed search-result pages for extracting translations of frequent unknown queries. Moreover, Huang *et al.* [2005] takes advantage of cross-language query expansion to retrieve more relevant search-result pages and then extract translations by combining with phonetic, semantic and frequency-distance features. However, these methods haven't solved the problems of translation extraction for infrequent unknown abbreviations and proper names. Currently, our search-result-based methods presented in this paper can effectively mitigate such kinds of translation problems.

## 5. Conclusions

In this paper we presented two new search-result-based methods to extract unknown term translation based on the previous method proposed by Cheng *et al*., including the search-result-based abbreviation translation extraction method and the two-stage hybrid translation extraction method. Our experimental results demonstrate the effectiveness of improving translation extraction for infrequent unknown abbreviations and proper names. Additionally, our proposed adaptive co-occurrence model is effective in aiding the process of selecting the correct full name candidates for the best abbreviated terms. However, currently,

the search-result-based abbreviation translation extraction method can perform well in the first stage of extracting the full names of those test abbreviated terms but can hardly extract correct translations via the extracted full names in the second stage. In the future, we are investigating to integrate the cross-language query expansion techniques proposed by Huang *et al*. into our search-result-based abbreviation translation extraction method.

As for the two-stage hybrid translation extraction method, we will continuously collect larger amounts of English-Chinese transliterated name pairs via our proposed Web-based unsupervised learning algorithm to build a more reliable transliteration model. In the future, referring to the methods proposed by both Lam *et al*. [2004] and Huang *et al*. [2005], we will extend our method by involving both semantic and phonetic information and expect that it can be more robust in extracting translations of unknown proper names.

## References

Ballesteros, L. and W. B. Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," In *Proc. of 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, Philadelphia, USA, pp. 84-91.

Brown, P. F., S. A. D. Pietra, V. D. J. Pietra and R. L. Mercer, "The Mathematics of Machine Translation," *Computational Linguistics*, 19(2), 1993, pp. 263-312.

Cao, Y.-B. and H. Li, "Base noun phrase translation using Web data and the EM algorithm," In *Proc. of COLING*, 2002, Taipei, Taiwan, pp. 127-133.

Cheng, P.-J., J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, L.-F. Chien, „Translating unknown queries with web corpora for cross-language information retrieval," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 146-153.

Davis, M. W. and W. C. Ogden, "Free Resources and Advanced Alignment for Cross-Language Text Retrieval," In *Proc. of the Sixth Text Retrieval Conference* (*TREC 6*), 1998, Gaithersburg, Maryland, pp. 385-394.

Fung, P. and L.-Y. Yee, "An IR approach for translating new words from nonparallel, comparable texts," In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics*, 1998, Montreal, Quebec, Canada, pp. 414-420.

Hull, D. A. and G. Grefenstette, "Querying across Languages: A Dictionary-based Approach to Multilingual Information Retrieval," In *Proc. of 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996, Zurich, Switzerland, pp. 49-57.

Gao, W., K.-F. Wong and W. Lam, "Phoneme-based Transliteration of Foreign Name for OOV Problem" In *Proc. of the first International Joint Conference on Natural Language Processing* (*IJCNLP*), 2004, Hainan Island, China, pp. 274-381.

Huang, F., Y. Zhang and S. Vogel, "Mining Key Phrase Translations from Web Corpora," In *Proc. of HLT-EMNLP*, 2005, Vancouver, B.C., Canada, pp. 483-490.

Kilgarriff, A. and G. Grefenstette, "Introduction to the special issue on the web as corpus," *Computational Linguistics*, 29(3), 2003, pp. 333-348.

Knight, K. and J. Graehl, "Machine Transliteration," *Computational Linguistics* 24(4), 1998, pp. 599-612.

Lam, W., R. Huang, P.-S. Cheung, "Learning phonetic similarity for matching named entity translations and mining new translations," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 281-288.

Leah, L., P. Ogilvie, A. Price, and B. Tamilio, "Acrophile: An Automated Acronym Extractor and Server," In *Proc. of the 5th ACM Digital Libraries Conference*, 2000, San Antonio, TX, pp. 205-214.

Li, H., M. Zhang and J. Su, "A Joint Source-Channel Model for Machine Transliteration," In *Proc. of 42th Annual Meeting of the Association for Computational Linguistics*, 2004, Forum Convention Centre, Barcelona, pp. 160-167.

Lin, T., C.-C. Wu, J.-S. Chang, "Word-Transliteration Alignment," In *Proc. of ROCLING XV*, 2003, Hsinchu, Taiwan, pp. 1-16.

Lin, W.-H. and H.-H. Chen, "Backward machine transliteration by learning phonetic similarity," In *Proc. of CONLL*, 2002, Taipei, Taiwan, pp. 139-145.

Lu, W.-H., L.-F., Chien, H.-J. Lee, "Translation of Web Queries using Anchor Text Mining," *ACM Transactions on Asian Language Information Processing*, 1(2), 2002, pp. 159-172.

Ma, W.-Y. and K.-J. Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," In *Proc. of ACL workshop on Chinese Language Processing*, 2003, pp. 31-38.

Melamed, I. D., "Models of translational equivalence among words," *Computational Linguistics*, 26(2), 2000, pp. 221-249.

Meng, H., W.-K. Lo, B. Chen and K. Tang, "Generate Phonetic Cognates to Handle Name Entities in English-Chinese Cross-Language Spoken Document Retrieval," In *Proc. of Automatic Speech Recognition and Understanding Workshop* (ASRU), 2001, Italy, pp. 311-314.

Nie, J.-Y., P. Isabelle, M. Simard, and R. Durand, "Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web," In *Proc. of 22th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, University of California, Berkeley, pp. 74-81..

Park, Y. and R. J. Byrd, "Hybrid text mining for finding abbreviations and their definitions," In *Proc. of Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 2001, pp. 126-133.

Rapp, R., "Automatic identification of word translations from unrelated English and German corpora," In *Proc. of 37th Annual Meeting of the Association for Computational Linguistics*, 1999, College Park, Maryland, USA, pp. 519-526.

Resnik, P., "Mining the Web for Bilingual Text," In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*,.1999, College Park, Maryland, USA, pp. 527-534.

Smadja, F., K. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: a statistical approach," *Computational Linguistics*, 22(1), 1996, pp. 1-38.

Taghva, K. and J. Gilbreth, "Recognizing Acronyms and their Definitions. Technical Report 95-03," *ISRI* (*Information Science Research Institute*), *UNLV*, June, 1995.

Virga, P. and S. Khudanpur, "Transliteration of Proper Names in Cross-Lingual Information Retrieval," *ACL 2003 workshop MLNER*.

Wan, S. and C. M. Verspoor, "Automatic English-Chinese name transliteration for development of multilingual resources," In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics*, 1998, Montreal, Quebec, Canada, pp. 1352-1357.

Xiao, J., J. Liu and T.-S. Chua, " Extracting pronunciation-translated names from Chinese texts using bootstrapping approach," In Proc. of *the 1st SIGHAN workshop on Chinese Language Processing*, 2002, Taipei, Taiwan, pp. 1-6.

Yang, C. C. and K. W. Li, "Automatic Construction of English/Chinese Parallel Corpora, " *Journal of the American society for Information Science and Technology*, 54(8), 2003, pp. 730-742.

Zhang, Y. and P. Vines, "Using the web for automated translation extraction in cross-language information retrieval," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 162-169.

## Appendix

**Partial English abbreviation test data**

| FA-Dreamer-E | IA-Dreamer-E | FA-Wiki-E | RA-Wiki-E |
|---|---|---|---|
| EDI | ADSM | ACM | NFL |
| ERP | AMIA | AMD | ABS |
| FMEA | ALSA | AOL | ACARS |
| TSMC | ATN | BBS | AGP |
| VLSI | BFI | CAD | ALTE |
| OTC | BGP | CDMA | BBS |
| VSAT | CGS | CEO | CICS |
| AIT | BSI | CMMI | DOM |
| CPR | CGMH | CS | DSP |

**Partial Chinese abbreviation test data**

| FA-Dreamer-C | IA-Dreamer-C |
|---|---|
| 台銀<br>(Bank of Taiwan) | 中影<br>( Central Motion Pictures Company) |
| 日亞航<br>(Japan Asia Airways) | 中選會<br>( Central Election Commission) |
| 中信銀<br>(Chinatrust Commercial Bank) | 智財權<br>( Intellectual Property Right) |
| 勞保<br>(Labor Insurance) | 台啤<br>( Taiwan Beer) |
| 證交稅<br>( Securities Exchange Transaction Tax) | 國衛院<br>( National Health Research Institutes) |
| 竹科<br>(Hsinchu Science Park) | 央銀<br>(Central Bank) |
| 華航<br>(China Airlines) | 兒福<br>(Child Welfare) |
| 中研院<br>(Academia Sinica) | 國台辦<br>( Taiwan Affairs Office of the State Council) |
| 台大<br>(National Taiwan University) | 客服<br>(Customer Service) |

**Partial English and Chinese transliteration test data**

| FP-Dreamer-E | IP-Dreamer-E | FP-Dreamer-C | IP-Dreamer-C |
|---|---|---|---|
| Alex | Athena | 法拉利<br>(Ferrari) | 雅典娜<br>(Athena) |
| Benz | Austin | 古奇<br>(Gucci) | 奧斯汀<br>(Austen) |
| Betty | Kournikova | 辛吉斯<br>(Hingis) | 庫妮可娃<br>(Kournikova) |
| Bosch | Bond | 義大利<br>(Italy) | 龐德<br>(Bond) |
| Calvin Klein | Brandy | 肯尼<br>(Kenny) | 布蘭蒂<br>(Brandy) |
| Ferrari | Charles | 托福<br>(Tofel) | 查爾斯<br>(Charles) |
| Gucci | David Robinson | 泰迪<br>(Teddy) | 大衛羅賓森<br>(David Robinson) |
| Hingis | Damon | 茱蒂<br>(Judy) | 達蒙<br>(Damon) |
| Italy | Duncan | 迪士尼<br>(Disney) | 鄧肯<br>(Duncan) |