Improving Translation of Unknown Proper Names Using a Hybrid Web-based Translation Extraction Method

Min-Shiang Shia Jiun-Hung Lin Scott Yu Wen-Hsiang Lu

Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{foreverdream, jhlin, scottyu}@csie.ncku.edu.tw, whlu@mail.ncku.edu.tw

Abstract

Recently, we have proposed several effective Web-based term translation extraction methods exploring Web resources to deal with translation of Web query terms. However, many unknown proper names in Web queries are still difficult to be translated by using our previous Web-based term translation extraction methods. Therefore, in this paper we propose a new hybrid translation extraction method, which combines our pervious Web-based term translation extraction method and a new Web-based transliteration method in order to improve translation of unknown proper names. In addition, to efficiently construct a good quality transliteration model, we also present a mixed-syllable-mapping transliteration model and a Web-based semi-supervised learning algorithm to explore search-result pages further for collecting large amounts of English-Chinese transliteration pairs from the Web.

1 Introduction

In machine translation (MT) (Brown et al. 1993) or cross-language information retrieval (CLIR) (Jaleel and Larkey 2003; Pirkola et al. 2003), unknown term translation are still problematic and remain to be solved. Conventionally, most of the existing MT or CLIR systems rely mainly on general-purpose bilingual dictionaries, which usually lack translations of proper names or technical terms, and thus are unable to deal with such problems. We have proposed an effective Web-based approach to exploring abundant language-mixed texts on the Web like anchor texts and search-result pages for alleviating the difficulty of unknown query term translation (Lu et al. 2002, 2004; Cheng et al. 2004). However, the approach employing statistical techniques still suffers from the problem of data sparseness and indirect association errors in finding translations of low-frequency unknown terms (Melamed, 2000).

According to the report in previous research (Davis et al. 1998), around 50% of unknown terms are proper names. To improve translation of unknown proper names, in this paper, we propose a hybrid translation extraction method, which is composed of our pervious search-result-based term translation extraction method (Section 3.2) and a new Web-based transliteration method (Section 3.3). Transliteration is the process that converting a sequence of substrings or characters in the source language (e.g., English) into a pronunciation-approximate substring/character sequence in the target language (e.g., Chinese). Many researchers have proposed phoneme-based mapping techniques for proper name transliteration (Jung et al. 2000; Knight & Graehl 1998; Lin & Chen 2002; Meng et al. 2001; Virga & Khudanpur 2003), but converting an English word from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters may cause double errors. Taking this problem into consideration, we thus try to adopt direct orthographical mapping for proper name transliteration and propose a simple mixed-syllable-mapping transliteration model which can effectively increase the correct mapping between an English-Chinese transliteration pair with different number of transliteration unit (syllable), such as "Ericsson" (易利信) with four English transliteration units "e", "ri", "c", "sson" and three Chinese transliteration units "易", "利", "信" (Section 3.3). Additionally, to train a good quality transliteration model which is used to filter out impossible transliteration candidates in the process of extracting translation of unknown proper names, we also present a Web-based semi-supervised learning algorithm to collect large amounts of English-Chinese transliteration pairs from the Web. Experimental results show that our new approach can make improvements for translation of unknown proper names.

2 Related Work

2.1 Parallel-Corpus-based Term Translation Extraction

Term translation extraction is a significant research topic in the field of machine translation. A number of related researches (Gale and Church 1991; Kupiec 1993; Melamed 2000; Smadja et al. 1996) have used sentence-aligned parallel corpora to extract translations since the advent of statistical translation model (Brown et al. 1990, 1993). For example, Melamed (2000) proposed statistical translation models to improve the techniques of word alignment by taking advantage of pre-existing knowledge and overcome the problems of indirect association errors, i.e., erroneous translational correspondence arose from highly co-occurred relevant terms. Although high accuracy of translation extraction can be easily achieved by these techniques, sufficiently large parallel corpora for various subject domains and language pairs currently are not always available.

2.2 Comparable-Corpus-based Term Translation Extraction

However, less attention has been devoted to automatic extraction of term translations from comparable or even unrelated texts, since such methods encountered more difficulties due to lacking parallel correlation aligned between documents or sentence pairs. Rapp (1999) proposed an approach to utilizing non-parallel corpora based on the assumption that the contexts of a term should be similar to the contexts of its translation in any language pairs. Fung et al. (1998) also proposed a similar approach that uses vector-space model and takes a bilingual lexicon (called seed words) as feature set to estimate the similarity between a word and its translation candidates. These works are important for automatic extraction of new terminology and unknown proper names in diverse domains. It is a pity that comparable corpora are easier to obtain, however, how to achieve better performance for higher translation coverage is still a challenging task.

2.3 Web-based Term Translation Extraction

The Web is becoming the largest data repository in the world, which consists of huge amounts of multilingual and wide-scoped hypertext resources. A number of studies have been concentrated in the use of the Web to complement insufficient corpora (Cao & Li 2002; Kilgarriff et al. 2003). How to utilize the Web resources to benefit translations of unknown terms is worthy to investigate.

As mentioned above, the conventional term translation methods suffer from the problems of the lack of large-size parallel corpora and the shortage of translation coverage of comparable corpora in medical domain. Thus, we have proposed several Web-based methods to effectively deal with translation of frequent Web query terms by exploring Web anchor text and search-result pages. Although the anchor-text-based approach has been proven effective in extracting multilingual translations (Lu et al. 2002, 2004), it requires crawling the Web to gather sufficient training data as well as more network bandwidth and storage. For the reason to reduce such costs, this paper only adopts the search-result-based approach to extract translation candidates for term translation (describes in Section 3.2). However, many proper names are still difficult to be translated correctly using the search-result-based approach. Therefore, in this paper we intend to further explore search results to collect English–Chinese transliteration pairs, and build a good quality transliteration model which can be used to filtered out impossible translation candidates to improve translation of unknown proper names.

2.4 Proper Name Transliteration

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 (Jung et al. 2000; Knight & Graehl 1998; Lin & Chen 2002; Meng et al. 2001; Virga & Khudanpur 2003). Knight and Graehl used an English-katakana dictionary, katakana-English phoneme mapping, and the CMU Speech Pronunciation Dictionary to deal with transliteration between English words and Katakana sequences. Lin et al. (2003) proposed a statistical transliteration model and apply the model to extract proper names and their transliterations in a parallel corpus with high average precision and recall rates. However, Li et al. (2004) have 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 (Pirkola et al. 2003), such as French and English. Second, converting English words to Chinese characters will need two steps: converting from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters. Two cascaded converting steps may cause double errors. Taking this problem into consideration, we try to adopt direct orthographical mapping for name transliteration (described in Section 3.3).

3 Extracting Translation of Unknown Proper Names

3.1 Problem and Challenge

Actually, search-result page is a good resource for extracting translation of frequent unknown query terms. However, a number of unknown proper names are still not extracted correctly due to the problems of data sparseness. Thus, our idea is to integrate name transliteration techniques into the process of extracting translation of proper names in order to filter impossible transliterated candidates for improving the performance of translation extraction. To deal with the problem, first we need to extract terms from the search-result pages as translation candidates, and then filter out impossible candidates based on the name transliteration model. In fact, it is challenging to build a good quality transliteration model while lacking sufficient transliteration pairs for training. We therefore propose a Web-based semi-supervised learning algorithm to collect large amounts of English-Chinese transliteration pairs from the Web (see Section 3.3).

3.2 Extracting Translation Candidates Using a Search-Result-based Translation Extraction Method

We have proposed an effective search-result-based method to explore language-mixed search-result pages and utilize co-occurrence relation and context information for extracting unknown query term translation. In this section, we will simply describe candidate selection methods using the search-result-based method. For more details, please refer to our previous work (Cheng et al. 2004).

(1) Chi-Square Test Method: On the basis of co-occurrence analysis, chi-square test (χ^2) is adopted to estimate semantic similarity between the source term *E* and the target candidate *C*. The similarity measure is defined as

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

where *a*, *b*, *c* and *d* are the numbers of pages retrieving 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 nature of Chinese-English mixed texts often appearing in Chinese pages, the source term *E* and the target candidate *C* may share common contextual terms in the search-result pages (Fung & Yee 1998; Rapp 1999). The similarity between *E* and *C* will be computed based on their context feature vectors in the vector-space model. The conventional *tf-idf* weighting scheme is used and defined as

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

where $f(t_i, p)$ is the frequency of term t_i in 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 as follows:

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

3.3 Filtering Translation Candidates Using a Web-Based Name Transliteration Method

(1) English Letter Substring 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 letter substrings. Each English substring is regarded as a transliteration unit (TU) in this paper and had at most one corresponding character of the Chinese transliterated name. Initially, we used only five rules 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 left vowel only if they are not followed by a vowel.
- A consonant and a following vowel are regarded as a TU.

For example, “amaya” (阿馬雅) is segmented into three substrings “a”, ”ma”, “ya”, and “mobley” (莫布里) is segmented into three substrings “mo”, “b”, “ley”. Currently, some English names may be segmented incorrectly, but it is easy to manually add new rules for improving English letter substring segmentation.

(2) Mixed-Syllable-Mapping Transliteration Model: To avoid double errors of converting English phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters, we thus adopted direct orthographical mapping to deal with the alignment between any English name, $E = e_1e_2 \ldots e_m$, and its Chinese transliterated name, $C = c_1c_2 \ldots c_n$. Each English TU e_i is mapped to a Chinese character c_i with the probability $P(c_i \mid e_i)$. Initially, to efficiently train a Web-based transliteration model based on the collected transliteration pairs from the Web for filtering out impossible transliteration candidates, we adopt a simple name transliteration model called **forward-syllable-mapping transliteration model**, which computes the forward syllable mapping score between E and C using the following formula:

$$S_{FSM}(E,C) = P(C \mid E) \approx \prod_{i=1}^{\min(m,n)} [(1-\alpha)P(c_i \mid e_i) + \alpha], \quad (4)$$

where α is the smoothing weight.

For an English-Chinese transliteration pair with different number of transliteration unit, such as “Rusedski” (魯塞斯基) with the five English segmented substrings “ru”, ”se”, “d”, “s”, “ki” and four Chinese characters “魯”, “塞”, “斯”, “基”, to increase the correct mapping between English TUs and Chinese characters, we propose an alternative transliteration mapping model called

reverse-syllable-mapping transliteration model, which is used to compute the reverse syllable mapping score as follows:

$$S_{RSM}(E,C) \approx \begin{cases} \prod_{i=m}^{m-n+1}[(1-\alpha)P(c_{i-(m-n)} \mid e_i)+\alpha], & m \geq n; \\ \prod_{i=n}^{n-m+1}[(1-\alpha)P(c_i \mid e_{i-(n-m)})+\alpha], & m < n. \end{cases} \qquad (5)$$

To cover all possibly correct mapping between English TUs and Chinese transliterated characters for the distinct types of English-Chinese transliteration pairs with the same or different transliteration units, we propose a simple **mixed-syllable-mapping transliteration model,** which combine the forward-syllable-mapping and reverse-syllable-mapping transliteration models, to estimate the mapping score as follows:

$$S_{MSM}(E,C) = \sqrt{S_{FSM}(E,C) \times S_{RSM}(E,C)}. \qquad (6)$$

(3) Web-based Semi-Supervised Learning Algorithm: We intend to take advantages of abundant language-mixed texts on the Web to collect English-Chinese transliteration pairs and then train a good quality transliteration model. Thus, we design a semi-supervised learning process of transliteration mapping. The process is composed of three main stages: extraction of Chinese transliterated names, extraction of English original names, and learning of transliteration mapping, and described below as well as the algorithm in Figure 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 of Chinese transliterated names by utilizing search-result pages. Inspired by Xiao et al., we design a different bootstrapping algorithm which uses the same seed character set to automatically find large amounts of Chinese transliterated names from search-result pages. Initially, we select two frequent Chinese transliterated characters from the seed character set, and then send them to search engines for getting search-results pages. To efficiently extract more Chinese transliterated names from the search-result pages, we use the CKIP tagger (Ma & Chen 2003), which is a representative Chinese POS tagger with the ability of segmenting Chinese texts into meaningful words and extracting unknown words.
- **Extraction of English Original Names:** We first use the search-result-based translation extraction method (Section 3.2) to find possible candidates of English original names, and then filter out the impossible candidates which are included in general-purpose bilingual dictionaries. Finally, to collect English-Chinese transliteration name pairs with high quality, we may need to take some manual efforts to examine the correct transliteration pairs.

```
Web-based Semi-Supervised Learning Algorithm for Collecting
English-Chinese Transliteration Pairs and Training a Transliteration Model

Input:    Chinese seed character set $C_s$ and a general-purpose bilingual dictionary $D$
Output:   English-Chinese transliteration pair set $V_{ec}$, and a transliteration model $T$

1.  Extraction of Chinese transliterated names:
    1.1.  Seed character selection: select two frequent characters from the Chinese
          seed character set $C_s$.
    1.2.  Search-result crawling: send the two selected characters to a search
          engine and get search-result pages.
    1.3.  Chinese transliterated name identification: use CKIP tagger to find
          unknown terms in the search-result pages, and then take the unknown
          terms containing the two Chinese seed characters as potential Chinese
          transliterated names and add them into $V_c$ .
    1.4.  Seed character set updating: update $C_s$ by adding the new characters
          from the new Chinese transliterated names.
    1.5.  Repeat step1 until the desired number of the Chinese transliterated name
          in the $V_c$ is reached.
2.  Extraction of English original names: for each potential Chinese transliterated
    name in $V_c$, perform the following sub-steps:
    2.1.  Potential English name extraction: use search-result-based translation
          extraction method (Section 3.2) to find potential candidates of English
          name.
    2.2.  Candidate filtering: filter out impossible English name candidates
          included in $D$.
    2.3.  English name identification: take some manual efforts to examine the
          correct original English names.
    2.4.  English-Chinese transliteration pair updating: update $V_{ec}$ by adding the
          new transliteration pair.
3.  Learning of English-Chinese transliteration mapping: use the proposed
    mixed-syllable-mapping transliteration model (equation (6)) to train a Web-based
    transliteration model $T$ based on the extracted English-Chinese transliteration
    pairs.
```

Figure 1. Algorithm for collecting transliteration pairs and training a transliteration model.

- **Learning of Transliteration Mapping:** On the basis of the English letter substring segmentation rules and the proposed mixed-syllable-mapping transliteration model described above, we will train a Web-based transliteration model based on the collected transliteration pairs from the Web.

3.4 The Proposed Approach to Translation Extraction

Currently, for some unknown proper names, it is still difficult to effectively extract translation by using our previous search-result-based translation extraction method. Therefore, we try to combine a new Web-based transliteration method to enhance our previous search-result-based translation extraction method.

(1) Linear Combination Method: Intuitively, a simple method is to directly combine the above three different methods: the chi-square test method, the context-vector analysis method, and the Web-based transliteration method. Under consideration of the large difference of ranges of similarity values among the above methods, we would use a linear combination of inverse ranks to compute the similarity measure as follows:

$$S_{Combined}(E,C) = \sum_m \frac{\alpha_m}{R_m(E,C)}, \qquad (7)$$

where α_m is an assigned weight for each similarity measure S_m, and $R_m(E, C)$ represents the similarity rank of each target candidate C with respect to its source term E and is assigned to be from 1 to k (candidate number) according to similarity measure $S_m(E, C)$ in decreasing order.

Note that this liner combination method is only used as baseline in comparison with our proposed hybrid translation extraction method described below in the following experiments (Section 4.2).

(2) Hybrid Method: For some unknown proper names, the simple linear combination method might not make good improvements while these respective methods can't obtain high ranks for possibly correct transliteration candidates. Therefore, we propose a new hybrid translation extraction method in order to obtain better performance. First, we use the search-result-based translation extraction method described above to extract k (k = 20) terms with high similarity score as transliteration candidates. Second, some impossible candidates included in general-purpose bilingual dictionaries are filtered out, and then each of the rest transliterated candidates is ranked according to transliteration mapping score with the test proper name which is computed based on the Web-based transliteration model (Equation (4) and (6)).

4 Experimental Results and Analysis

We conducted the following experiments to examine the performance of the proposed hybrid translation extraction method and the comparison with the simple linear combination method. Particularly, the focus of the experiments is mainly emphasized on the effectiveness of translations of unknown proper names using the proposed mixed-syllable-mapping transliteration model and hybrid translation extraction method.

Collected data: Initially, our proposed Web-based semi-supervised learning algorithm is employed to efficiently collect about 11,000 English-Chinese transliteration pairs for training a transliteration model.

Test set: We constructed one test set of unknown English query terms, **NTCIR proper name set**, which contains 22 unknown transliteration names from a total of 100 NTCIR2 and NTCIR3 title queries that contain 175 and 183 unique query terms respectively (Chen & Chen 2001).

Evaluation Metric: The average top-*n* inclusion rate was adopted as a metric on the extraction of translation equivalents. For a set of terms to be translated, its top-*n* inclusion rate was defined as the percentage of the terms whose translations could be found in the first *n* extracted translations (Cheng et al. 2004).

Table 1. Comparison of translation results between the forward-syllable-mapping model and the mixed-syllable-mapping model.

Translation Method	Forward-Syllable-Mapping Transliteration Model			Mixed-Syllable-Mapping Transliteration Model		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Name Transliteration	14%	27%	27%	27%	32%	32%
Hybrid	36%	41%	45%	45%	50%	55%

Table 2. Comparison of translation results between the different translation methods.

Translation Method	Mixed-Syllable-Mapping Transliteration Model		
	Top-1	Top-3	Top-5
Search-Result-based	36%	36%	45%
Name Transliteration	27%	32%	32%
Linear Combination	32%	41%	55%
Hybrid	45%	50%	55%

Table 3. Effective results of translation extraction using the hybrid translation extraction method (underlined terms indicate correct translation).

Test Query	Translation Method	Top 5 Translation Candidates	
		Forward-Syllable-Mapping Transliteration Model	Mixed-Syllable-Mapping Transliteration Model
Michael	Search-Result-based	麥可布雷,麥克傑克森,施文彬,華納,個人	麥可布雷,麥克傑克森,施文彬,華納,個人
	Name Transliteration	蜜可艾爾,麥可艾爾,蜜可艾德,蜜可埃爾,蜜高艾爾	蜜可艾爾,密可艾爾,麥可艾爾,蜜可埃爾,蜜麥艾爾
	Linear Combination	麥可布雷,麥克傑克森,蜜可艾爾,施文彬,華納	麥可布雷,麥克傑克森,蜜可艾爾,施文彬,華納
	Hybrid	麥可傑克森,麥可布雷,麥克傑克森,麥克,喬丹	麥可布雷,麥可傑克森,麥克,麥克傑克森,舒馬克
Kosovar	Search-Result-based	譯音無限次，發行公司，散財，科索沃，譯音無限	譯音無限次，發行公司,散財,科索沃，譯音無限
	Name Transliteration	可索瓦,克索瓦,茉索瓦,可蘇瓦,可喬瓦	可索雷,克索雷,可索瓦,克索瓦,科索雷
	Linear Combination	譯音無限次，發行公司，可索瓦，散財，科索沃	譯音無限次，發行公司,可索雷，散財，科索沃
	Hybrid	科索沃，譯音無限次，發行公司，散財，譯音無限	科索沃,譯音無限次，發行公司，散財，譯音無限

4.1 Mixed-Syllable-Mapping Transliteration Model vs. Forward-Syllable-Mapping Transliteration Model

To test the effectiveness of the mixed-syllable-mapping transliteration model, we carried out a comparative experiment with different ranking. The results are shown in Table 1. Actually, the mixed-syllable-mapping transliteration model is effective to improve the top-*n* inclusion rate. For translation extraction of the NTCIR proper names, the mixed-syllable-mapping transliteration model can achieve 27% and 45% top-1 inclusion rates for the name transliteration method and the hybrid translation method, respectively. Obviously, the reason is that for many English-Chinese transliteration pairs with different number of TU, reverse-syllable-mapping transliteration model can aid in learning correct mapping between English substrings and Chinese characters. Additionally, the model has the same assist effect to many partially matching transliteration pairs collected by using our proposed Web-based transliteration method. For the given proper name “Michael” (麥克) shown in Table 3, the better rank of its correct translation can be obtained by using the mixed-syllable-mapping transliteration model.

4.2 Hybrid Translation Extraction Method vs. Linear Combination Method

To determine the effectiveness of the proposed hybrid translation extraction method compared with other methods, we also did several comparative experiments with different ranking. The results are also shown in Table 2. For the NTCIR test set, surprisingly, the hybrid translation extraction method made a great improvement compared with the search-result-based translation extraction method, name transliteration method, or linear combination method. The hybrid translation extraction method with mixed-syllable-mapping transliteration model can achieve 45% top-1 inclusion rate. The main reason is that most of the incorrect translation candidates extracted by using the search-result-based translation extraction method can be filtered out by using the Web-based transliteration method. For example, given the proper name “Kosovar” (see Table 3), the correct Chinese transliterated name “科索沃” can be ranked to the top one from the fourth rank using only the search-result-based translation extraction method. However, the simple linear combination method seems not effective to improve translation performance since the name transliteration method is still limited in generating correct transliterated candidates even though it can generate many pronunciation-proximate candidates.

4.3 Discussions

Our proposed mixed-syllable-mapping model and hybrid translation extraction method is effective to improve performance in extracting translation of unknown proper names. However, the hybrid translation extraction method sometimes performs not good as linear combination method. An example such as “Viagra” (威而剛) is shown in Table 4. Currently, our Web-based semi-supervised learning algorithm is limited by insufficient transliteration training from our collected transliteration pairs which are still in the need of examining by large amounts of manual labor. In the future, we will develop an unsupervised learning algorithm to automatically collect much more amounts of English-Chinese

Table 4. Ineffective results of translation extraction using the hybrid translation extraction method (underlined terms indicate correct translation).

Test Query	Translation Method	Top 5 Translation Candidates
		Mixed-Syllable-Mapping Transliteration Model
Viagra	Search-Result-based	偉哥,食品藥物,威而剛,藥物管理局,藥物
	Name Transliteration	薇阿格拉,薇亞格拉,薇艾格拉,薇阿葛拉,薇亞葛拉
	Linear Combination	偉哥,食品藥物,薇阿格拉,威而剛,藥物管理局
	Hybrid	萬艾可,藥物管理,食品管理,輝瑞,威而剛

transliteration pairs from the Web for training good quality transliteration model. Besides them, there are still a number of cases that are still difficult to deal with by using the simple mixed-syllable-mapping transliteration model and need to be further investigated in the future.

5 Conclusions

We have presented a new hybrid translation extraction method that works well for improving extraction of translation of known proper names by effectively combining a previous search-result-based translation extraction method and our proposed Web-based name transliteration method. Additionally, our proposed simple mixed-syllable-mapping transliteration model and Web-based semi-supervised learning algorithm are also effective to collect English-Chinese transliteration pairs and then train a transliteration model for filtering out incorrect transliteration candidates in the process of extracting proper name translation.

References

N. A. Jaleel and L. S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. *CIKM* 2003: 139-146.

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

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

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

K.-H. Chen and H.-H. Chen. 2001. The Chinese Text Retrieval Tasks of NTCIR Workshop 2. In *Proc. of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

P.-J. Cheng, Y.-C. Pan, W.-H. Lu, L.-F. Chien. 2004. Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora. In *Proc. of ACL* 2004: 535-542.

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

P. Fung and L.-Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of ACL* 1998: 414-420.

W. A. Gale and K. W. Church. 1991. Identifying Word Correspondances in Parallel Texts, In Proc. of DARPA Speech and Natural Language Workshop.

W. Gao, K.-F. Wong and W. Lam. 2004. Phoneme-based Transliteration of Foreign Names for OOV Problem. *In Proc. of IJCNLP* 2004: 274-381.

J. Halpern. 2000. Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In *Proc. of Workshop on Asian Language Resources and International Standardization*.

S. Y. Jung, S. L. Hong and E. Paek. 2000. An English to Korean Transliteration Model of Extended Markov Window. In *Proc. of COLING* 2000.

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

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

J. M. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of ACL* 1993: 17-22.

H. Li, M. Zhang and J. Su. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proc. of ACL* 2004: 160-167.

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

W.-H. Lin and H.-H. Chen. 2002. Backward machine transliteration by learning phonetic similarity. In *Proc. of CONLL* 2002: 139-145.

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

W.-H. Lu., L.-F. Chien and H.-J. Lee. 2004. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems* 22(2): 242-269.

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

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

H. Meng, W.-K. Lo, B. Chen and K. Tang. 2001. *Generate Phonetic Cognates to Handle Name Entities in English-Chinese Cross-Language Spoken Document Retrieval*, ASRU 2001.

A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala and K. Jarvelin. 2003. Fuzzy Translation of Cross-Lingual Spelling Variants, In *Proc. of SIGIR* 2003: 345-352.

Y. Qu and G. Grefenstette. 2004. Finding Ideographic Representations of Japanese Names Written in Latin Script via Language Identification and Corpus Validation In *Proc. of ACL* 2004: 184-191.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora, In *Proc. of ACL* 1999: 519-526.

R. Schwartz and Y.-L. Chow. 1990. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypothesis. In *Proc. of ICCASP* 1990: 81-84.

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

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

S. Wan and C. M. Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proc. of ACL* 1998: 1352-1357.

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