Proceedings of the ACL 2003 Workshop on Multilingual and
Mixed-language Named Entity Recognition, pp. 1-8.

Learning Formulation and Transformation Rules for Multilingual Named Entities

Hsin-Hsi Chen **Changhua Yang** **Ying Lin**
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, 106
{hh_chen, d91013, b88034}@csie.ntu.edu.tw

Abstract

This paper investigates three multilingual named entity corpora, including named people, named locations and named organizations. Frequency-based approaches with and without dictionary are proposed to extract formulation rules of named entities for individual languages, and transformation rules for mapping among languages. We consider the issues of abbreviation and compound keyword at a distance. Keywords specify not only the types of named entities, but also tell out which parts of a named entity should be meaning-translated and which part should be phoneme-transliterated. An application of the results on cross language information retrieval is also shown.

1 Introduction

Named entities are major components of a document. Capturing named entities is a fundamental task to understanding documents (MUC, 1998). Several approaches have been proposed to recognize these types of terms. For example, corpus-based methods are employed to extract Chinese personal names, and rule-based methods are used to extract Chinese date/time expressions and monetary and percentage expressions (Chen and Lee, 1996; Chen, Ding and Tsai, 1998). In the past, named entity extraction mainly focuses on general domains and is employed to various applications such as information retrieval (Chen, Ding and Tsai, 1998), question-answering (Lin, *et al*., 2001), and so on. Recently, several attempts have been extended to mine knowledge from biomedical documents (Hirschman, *et al*., 2002).

Most of the previous approaches dealt with monolingual named entity extraction. Chen *et al*. (1998) extended it to cross-language information retrieval. A grapheme-based model was proposed to compute the similarity between Chinese transliteration name and English name. Lin and Chen (2000) further classified the works into two directions – say, forward transliteration (Wan and Verspoor, 1998) and backward transliteration (Chen *et al*., 1998; Knight and Graehl, 1998), and proposed a phoneme-based model. Lin and Chen (2002) employed a machine learning approach to determine phonetic similarity scores for machine transliteration. AI-Onaizan and Knight (2002) investigated the translation of Arabic named entities to English using monolingual and bilingual resources.

The past works on multilingual named entities emphasizes on the transliteration issues. However, the transformation between named entities in different languages is not transliteration only. The mapping may be a combination of meaning translation and/or phoneme transliteration. The following five English-Chinese examples show this issue. The symbol $A \Leftrightarrow B$ denotes a foreign name A is translated and/or transliterated into a Chinese name B.

(s1) Victoria Fall
⇔ 維多利亞瀑布 (wei duo li ya pu bu)
(s2) Little Rocky Mountains
⇔ 小落磯山脈 (xiao luo ji shan mo)

(s3) Great Salt Lake ⇔ 大鹽湖 (da yan hu)
(s4) Kenmare ⇔ 康美爾 (kang mei er)
(s5) East Chicago ⇔ 東芝加哥 (dong zhi jia ge)

Example (s1) shows a name part (i.e., Victoria) and a keyword part (i.e., Fall) of a named location are transliterated and translated into "維多利亞" (wei duo li ya) and "瀑布" (pu bu), respectively. In Example (s2), the keyword part (i.e., Mountains) is still translated, i.e., "山脈" (shan mo), however, some part of name is translated (i.e., Little ⇔ "小" (xiao)) and another part is transliterated (i.e., Rocky ⇔ "落磯" (luo ji)). Example (s3) shows an extreme case. All the three words are translated (i.e., Great ⇔ "大" (da)), Salt ⇔ "鹽" (yan), Lake ⇔ "湖" (hu)). Examples (s4) and (s5) show two location names without keywords. The former is transliterated and the latter is a combination of transliteration and translation.

Which part is translated and which part is transliterated depends on the type of named entities. For example, personal names tend to be transliterated. For a location name, name part and keyword part are usually transliterated and translated, respectively. The organization names are totally different. Most of constituents are translated. Besides the issue of the named entity types, different language pairs have different transformation rules. German named entity has decompounding problem when it is translated/transliterated, e.g., Bundesbahn ⇔ "聯邦鐵路局" (lian bang tie lu ju) and Bundesbank ⇔ "聯邦銀行" (lian bang yin hang).

This paper will study the issues of languages and named entity types on the choices of translation and transliteration. We focus on three more challenging named entities only, i.e., named people, named locations and named organizations. Three phrase-aligned corpora will be adopted – say, a multilingual personal name corpus and a multilingual organization name corpus compiled by Central News Agency (abbreviated CNA personal name and organization corpora hereafter), and a multilingual location name corpus compiled by National Institute for Compilation and Translation of Taiwan (abbreviated NICT location name corpus hereafter). We will extract transliteration/translation rules from these multilingual named corpora. This paper is organized as follows. Section 2 introduces the corpora used. Section 3 shows how to extract formulation rules and the transformation rules. Section 4 analyzes the results. Section 5 demonstrates the application of the extracted rules on cross language information retrieval. Section 6 concludes the remarks.

2 Multilingual Named Entity Corpora

NICT location name corpus which was developed by Ministry of Education of Taiwan in 1995 collected 19,385 foreign location names. Each entry consists of three parts, including foreign location name, Chinese transliteration/translation name, and country name, e.g., (Victoria Fall, "維多利亞瀑布" (wei duo li ya pu bu), South Africa), (Little Rocky Mountains, "小落磯山脈" (xiao luo ji shan mo), USA), *etc*. The foreign location names are in English alphabet. Some location names denoting the same city have more than one form like Firenze and Florence for a famous Italian city. The former is an Italian name and the latter is its English name. They correspond to two different transliterations in Chinese, respectively, i.e., "翡冷翠" (fei leng cui) and "佛羅倫斯" (fo luo lun si). The pronunciation of the foreign names in NICT corpus is based on Webster's New Geographic Dictionary. The foreign name itself may be a transliteration name. A Japanese city is transliterated in English alphabet, but its corresponding translation name is in Kanji (Hanzi in Japanese). It is hard to capture their relationships except dictionary lookup, so that Japanese location name is out of our discussion. We employ the country field to select the translation/transliteration pairs that we will deal with in this paper. Table 1 summarizes the statistics of NICT corpus based on country tags.

Table 1. Statistics of NICT Corpus

Country	Frequency	Percentage	Country	Frequency	Percentage
USA	3,012	15.5%	Korea	574	3.0%
UK	1,073	5.5%	Brazil	433	2.2%
Russia	961	5.0%	German	395	2.0%
Japan	796	4.1%	Italy	379	2.0%
Canada	692	3.6%	Spain	370	1.9%
France	679	3.5%	Mexico	324	1.7%
India	679	3.5%	Others	8,413	43.5%
Australia	603	3.1%	Total	19,385	100%

CNA personal name and organization corpora are used by news reporters to unify the name transliteration/translation in news stories. There are 50,586 pairs of foreign personal names and Chinese transliteration/translation in persona name corpus. Different from NICT corpus, there do not exist clear cues to identify the nationality of named people. Thus, we could not exclude the Japanese names like "Hayakawa" and the corresponding name "早川" (zao chuan) from our discussion automatically. There are 14,658 named organizations in CNA corpus. Some organization names are tagged with the country names to which they belong. For example, "Aachen Technical University" ⇔ 亞肯技術大學 (ya ken ji shu da xue) (Germany). But not all the organization names have such country tags. Comparatively, organization names are longer than the other two named entities. Table 2 shows the statistics of NICT organization name corpus. FL denotes the length of foreign names in words, CL denotes the length of Chinese names in characters, and Count denotes the number of foreign names of the specified length.

Table 2. Statistics of CNA Organization Corpus

FL	Count	CL	FL	Count	CL	FL	Count	CL
1	1,773	4.73	7	425	9.94	13	10	14.20
2	3,622	4.98	8	223	10.50	14	6	12.00
3	3,751	6.30	9	122	10.98	15	5	17.00
4	2,406	7.28	10	53	11.57	16	2	14.50
5	1,434	8.27	11	32	13.41	18	1	9.00
6	775	8.97	12	17	12.35	20	1	15.00

3 Rule Mining

3.1 Frequency-Based Approach with a Bilingual Dictionary

We postulate that a transliterated term is usually an unknown word, i.e., not listed in a lexicon and a translated term often appears in a lexicon. Under this postulation, a translated term occurs more often in a corpus, and comparatively, a transliterated term only appears very few.

A simple frequency-based method will compute the frequencies of terms and use them to tell out the transliteration and translation parts in a named entity. Because Chinese has segmentation problem, we start the frequency computation from the foreign name part in a multilingual named entity corpus. The method is sketched as follows.

(1) Compute the word frequencies of each word in the foreign name list.

(2) Keep those words that appear more than a threshold and appear in a common foreign dictionary (e.g., an English dictionary). These words form candidates of simple keywords.

(3) Examine the foreign word list again. Those word strings that are composed of simple keyword candidates are candidates of compound keywords. We find out the compound keyword set by using collocation metric by selecting the most frequently occurring compounds through the well-known elimination of prepositions.

(4) Because the experimental corpus is aligned, we can cluster the Chinese name list based on foreign keywords. For each Chinese name cluster, we try to identify the Chinese keyword sets. Here a bilingual dictionary may be consulted.

The above algorithm extracts foreign/Chinese keyword sets from a multilingual named entity corpus. In the meantime, formulation rules for foreign names and Chinese counterparts are mined. A complete foreign name and a complete Chinese name are mapped into name-keyword combination. By the way, which method, translation or transliteration, is used is also determined.

Take NICT location name corpus as an example. The terms of frequencies greater than 20 include River (河, he), Island (島, dao), Lake (湖, hu), Mountain (山, shan), Bay (灣, wan), Mountain (峰, feng), Peak (峰, feng), Islands (群島, qun dao), Mountains (山脈, shan mo), Cape (角, jiao), City (城, cheng), Range (嶺, ling), Peninsula (半島, ban dao), Point (角, jiao), Strait (海峽, hai xia), River (川, chuan), Gulf (灣, wan), Cape (岬, jia), Pass (山口, shan kou), Plateau (高原, gao yuan), Headland (岬, jia), Harbor (港, gang), Sea (海, hai), Promontory (岬, jia), and Hills (丘陵, qui ling). On the one hand, a foreign location keyword, e.g., "Mountain", may correspond to two Chinese location keywords, e.g., "山" (shan) and "峰" (feng). On the other hand, the same Chinese location keyword "峰" (feng) can be translated into two English location keywords "Mountain" and "Peak".

Similarly, suffix and prefix for organization names can be extracted from CNA organization

name corpus. Some high frequent keywords are shown as follows.

(1) Suffix

Party (黨, dang), Association (協會, xie hui), University (大學, da xue), Co. (公司, gong si), Committee (委員會, wei yuan hui), Company (公司, gong si), Bank (銀行, yia hang), *etc*.

(2) Prefix

International (國際, guo ji), World (世界, shi jie), American (美國, mei guo), National (全國, quan guo), Japan (日本, ri ben), National (國家, guo jia), Asian (亞洲, ya zhou), *etc*.

3.2 Keyword Extraction without a Bilingual Dictionary

At the step (4) of the algorithm in Section 3.1, a bilingual dictionary is required. Because abbreviation is common adopted in translation, dictionary-based approach is hard to capture this phenomenon. A named organization "World Taiwanese Association" which is translated into "世台會" (shi tai hui) is a typical example. The term "World" is translated into an abbreviated term "世" (shi) rather than a complete term "世界" (shi jie). Here another approach without dictionary is proposed. Suppose there are M pairs of (foreign name, Chinese name) in a multilingual named entity corpus. The j^{th} pair, $1 \le j \le M$, is denoted by $\{E_j, C_j\}$, where E_j is a foreign named entity, and C_j is a Chinese named entity. Then some Chinese segment $c \in C_j$ should be associated with some foreign segment e $\in E_j$. Consider the following examples.

(s6) Aletschhorn **Mountain** ⇔ 阿利奇赫恩**山**
(s7) Catalan **Mountain** ⇔ 卡太蘭**山**
(s8) Cook **Strait** ⇔ 科克**海峽**
(s9) Dover, **Strait** of ⇔多佛**海峽**

We will align "山" (shan) and "海峽" (hai xia) to Mountain and Strait, respectively, from these examples.

We further decompose the named entities. If a named entity E_j comprises m words $\boldsymbol{w_1 w_2 \ldots w_m}$, then a candidate segment $e_{p,q}$ is defined as $\boldsymbol{w_p \ldots w_q}$, where $1 \le p \le q \le m$. If a Chinese named entity C_j has n syllables $\boldsymbol{s_1 s_2 \ldots s_n}$, then a candidate segment $c_{x,y}$ is defined as $\boldsymbol{s_x \ldots s_y}$, where $1 \le p \le q \le n$. Theoretically, we can get $\frac{m(m+1)}{2} \times \frac{n(n+1)}{2}$ pairs of $\{e_{p,q}, c_{x,y}\}$ from $\{E_j, C_j\}$. We then group the pairs collected from the multilingual named entity list and count the frequency for each occurrence. Those pairs with higher frequency denote significant segment pairs. In the above examples, both the two pairs {Mountain, "山" (shan)} and {Strait, "海峽" (hai xia)} appear twice, while the other pairs appear only once.

All the pairs $\{e, c\}$ whose frequency > 2 are kept. Two issues have to be addressed. The first is: redundancy which may exist in the pairs of segments should be eliminated carefully. If a pair $\{e, s_1\ s_2\ \ldots\ s_t\}$ occurs k times, then the frequency of $t \times (t+1)/2$ substrings ($1 \le u \le v \le t$) is at least k. The second is: e may be translated to more than one synonym, which has the same prefix, suffix, or infix. In examples (s10) and (s11), "Association" may be translated into "協會" (xie hui) and "聯誼會" (lian yi hui), where "會" (hui) is a common suffix of these two translation equivalents, so that its frequency is more than the translation equivalents.

(s10) World Trade Association ⇔ 世界貿易協會
(s11) North Europe Chinese Association ⇔ 北歐華人聯誼會

These two issues may be mixed together to make this problem more challengeable.

A metric to deal with the above issues is proposed. The concept is borrowed from *tf*×*idf* scheme in information retrieval to measure the alignment of each foreign segment and the possible Chinese translation segments. Assume there are N foreign segments. Term frequency (*tf*) of a Chinese translation segment c_i in e denotes the number of occurrences of c_i in e. Document frequency (*df*) of c_i is the number of foreign segments that c_i is translated to. We prefer to the Chinese translation segment that occur frequently in a specific foreign segment, but rarely in the remainder of foreign segments. Besides, we also prefer the longer Chinese segment, so that the length of a Chinese segment, i.e., $|c_i|$, is also considered.

$$score(\{e, c_i\}) = f(\{e, c_i\}) \times idf(c_i) \times \log_2(|c_i| + 1) \quad (1)$$

$$f(\{e, c_i\}) = \frac{tf(\{e, c_i\})}{\max_j (tf\{e, c_j\})} \quad (2)$$

$$idf(c_i) = \log_2 \left(\frac{N}{df(c_i)}\right), \quad (3)$$

For some e, the corresponding Chinese segment c is obtained by equation (4).

$$c = \arg\max_{c_i} score(\{e, c_i\}) \quad (4)$$

In this way, we can produce a ranking list of pairs of (foreign segment, Chinese segment), which form multilingual keyword pairs.

3.3 Extraction of Transformation Rules

We apply the keyword pairs extracted in the last section to the original named entity list. In (s6)-(s9), (mountain, 山 (shan)) and (strait, 海峽 (hai xia)) are significant keyword pairs. We replace the non-keywords of E_j and C_j with patterns γ and δ, respectively, get the following rules.

(s6') γ mountain ⇔ δ 山
(s7') γ mountain ⇔ δ 山
(s8') γ Strait ⇔ δ 海峽
(s9') γ, Strait of ⇔ δ 海峽

(s6') and (s7') can be grouped into a rule. As a result, a set of transformation rules can be formulated. From these examples, Chinese location name keyword tends to be located in the rightmost and the remaining part is a transliterated name. On the counterpart, foreign location name keyword tends to be either located in the rightmost, or permuted by some prepositions, comma, and the transliterating part.

3.4 Extraction of Keywords at a Distance

The algorithm proposed in Section 3.2 can deal with single keywords and connected compound keywords. Now we will extend it to keywords at a distance. Consider examples (s12)-(s15) at first.

(s12) ***American*** Podiatric medical ***Association***
⇔ **美國**足病醫療**學會**
(s13) ***American*** Public Health ***Association***
⇔ **美國**公共衛生**學會**
(s14) ***American Society*** for Industrial Security
⇔ **美國**工業安全**協會**
(s15) ***American Society*** of Newspaper Editors
⇔ **美國**報紙編輯人**協會**

(s12) and (s13) show that an English compound keyword is separated and so is its corresponding Chinese counterpart. In contrast, the English compound keyword is connected in (s14) and (s15), but the corresponding Chinese translation is separated. The phenomenon appears quite often in the translation of organization names.

We introduce a symbol Δ to cope with the distance issue. The original algorithm is modified as follows. A candidate segment $c_{p,q}$ is defined as a string that begins with $\boldsymbol{s_p}$ and ends with $\boldsymbol{s_q}$. Each syllable from $\boldsymbol{s_{p-1}}$ to $\boldsymbol{s_{q-1}}$ can be replaced by Δ. Therefore, both $e_{p,q}$ and $c_{x,y}$ are extended to $2^{(p-q-1)}$, and $2^{(x-y-1)}$ instances, respectively. For example, the following shows some additional instances for "American Civil Liberties Union".

"American Δ Liberties Union"
"American Civil Δ Union"
"American Δ Union"

The scoring method, i.e., formulas (1)-(4), is still applicable for the new algorithm. Nevertheless, the complexity is different. The complexity of the original algorithm is $O(m^2n^2)$, but the complexity of the algorithm here is $O(2^m2^n)$, where m is the word count for a foreign named entity and n is the character count for a Chinese named entity.

The mining procedure is performed only once, and the mined rules are employed in an application without being recomputed. Thus, the running time is not the major concern of this paper. Besides, the N is bounded in a reasonable small number because the length of a named entity is always rather shorter than that of a sentence. Table 2 shows that 93.88% of foreign names in CNA organization name corpus consist of less than 7 words.

4 Experimental Results

The algorithm in Section 3.2 was performed on NICT location name corpus, and CNA personal name and organization corpora. With this algorithm, we can produce a ranking list of pairs of (foreign segment, Chinese segment), which form multilingual keyword pairs. Individual foreign segments and Chinese segments are regarded as formulation rules for foreign languages and Chinese, respectively. When both the two

Table 3. Learning Statistics

	NICT LOC	CNA ORG	CNA PER
# of records in corpus	18,922	14,658	50,586
# of records for learning	5,714	12,885	100
Vocabulary size	18,220	11,542	50,315
# of keyword pairs	122	5,229	12
# of transformation rules	230		
# of successful records	4,262		

segments are considered together, they form a transformation rule. Table 3 summarizes the results using the frequency-based approach without dictionary. For named locations, there are 18,922 records, of which, only 5714 records consist of more than one foreign word. In other words, 13,208 named locations are single words, and they are unique, so that we cannot extract keywords from these words. Total 122 keyword pairs are identified. We classify these keyword pairs into the following types:

(1) Meaning translation
Total 69 keywords belong to this type. It occupies 56.56%. They are further classified into three subtypes.
   (a) common location keywords
   Besides the English location keywords mentioned in Section 3.1, some location keywords in other languages are also captured, including Bir ⇔ 井 (jing), Ain ⇔ 泉 (quan), Bahr ⇔ 河 (he), Cerro ⇔ 山 (shan), *etc*.
   (b) direction (e.g., Low ⇔ 下 (xia), Central ⇔ 中 (zhong), East ⇔ 東 (dong), *etc*.), size (e.g., Big ⇔ 大 (da)), length (e.g, Long ⇔ 長 (zhang)), color (e.g., Black ⇔ 黑 (hei), Blue ⇔ 藍 (lan), *etc*.)
   (c) the specificity of place or area such as Crystal ⇔ 結晶 (jie jing), Diamond ⇔ 鑽石 (zuan shi), *etc*.

(2) Phoneme transliteration keywords
Some morphemes are transliterated such as el ⇔ 拉 (la), Dera ⇔ 德拉 (de la), Monte ⇔ 蒙特 (meng te), Los ⇔ 洛斯 (luo si), Le ⇔ 勒 (le), and so on. Besides, some common transliteration names are also regarded as keywords, e.g., Elizabeth ⇔ 伊利莎白 (yi li sha bai), Edward ⇔ 愛德華 (ai de hua), *etc*. Total 39 terms belong to this type. It occupies 31.97%.

(3) Some keywords in type (1) are transliterated. For example, Bay ⇔ 貝 (Bay), Beach ⇔ 比奇 (bi qi), mountain ⇔ 蒙坦 (meng tan), Little ⇔ 利特 (li te), *etc*. Total 14 keywords (11.48%) are extracted.

Total 230 transformation rules are mined from the NICT location corpus. On the average, a keyword pair corresponds to 1.89 transformation rules. Consider a keyword pair *mountain* ⇔ 山 (shan) as an example. Four transformation rules shown as follows are learned, where α and β denote keywords for foreign language and Chinese, respectively; δ is a Chinese transliteration of a foreign fragment γ; the number enclosed in parentheses denotes frequency the rule is applied.

(1) γα ⇔ δβ (234)
(2) γ, α ⇔ δβ (45)
(3) γ, αγ ⇔ δβ (1)
(4) γαγ ⇔ δβ (1)

When we apply the 230 transformation rules back to the 5,714 named locations, we can tell out which part is transliterated and which part is translated from 4,262 named locations. It confirms our postulation that a named location is composed of two parts, i.e., one is translated and the other one is transliterated.

Comparatively, there are 50,586 personal names in CNA personal names, but only 100 named people are composed of more than one word. The number of keywords extracted is only a few. They are listed below.

De ⇔ 戴 (dai), La ⇔ 拉 (la), De La ⇔ 戴拉 (dai la), Van Der ⇔ 范德 (fan de), Du ⇔ 杜 (du), David ⇔ 大衛 (da wei), Khan ⇔ 汗 (han), Del ⇔ 戴 (dai), Le ⇔ 勒 (le), Van Den ⇔ 范登 (fan deng), Di ⇔ 迪 (di)

It shows that personal names tend to be transliterated and the CNA personal name corpus is suitable for training the similarity scores among phonetic characters (Lin and Chen, 2002).

Finally, we consider the named organizations. There are 14,658 records in CNA organization corpus. Total 12,885 organization names are composed of more than one word. The percentage, 87.90%, is the highest among these three corpora. Besides that, 5,229 keyword pairs are extracted.

Most of the keyword pairs are meaning translated. This set is also the largest among the three corpora. Thus, the keyword pairs are too small and too large to find suitable transformation rules for personal names and organization names, respectively.

Although the original idea of our algorithm is universal for languages, it should be modified slightly for some specific languages. The following takes German as examples. German words have cases and genders. Most of German words are compound. Consider examples (s16)-(s19).

(s16) **Neu***e* Osnabruecker ⇔ **新**奧斯納布律報
(s17) **Neu***es* Deutschland ⇔ **新**德國
(s18) **Bund**esbahn ⇔ **聯邦**鐵路局
(s19) **Bund**esbank ⇔ **聯邦**銀行

The first two examples show the German adjective **Neu** (New) has different suffixes such as "*-e*" and "*-es*" according to the case and gender of the noun. The last two examples suggest that morphological analysis for decompounding the words into meaningful segments is necessary before our algorithm.

5 Application on CLIR

Cross language information retrieval (CLIR) facilitates using queries in one language to access documents in another. Because named entities are key components of a document, they are usually targets that users are interested in. Figure 1 shows an application of the extracted formulation rules and transformation rules on Chinese-Foreign CLIR. For each document in the Foreign collection, named entities are recognized and classified by using formulation rules. They form important indices for the related documents. When a Chinese query is issued, the system extracts the possible Chinese named entities according to Chinese formulation rules. If keywords are specified in a query, we know the structure and the type of the named entity. The lexical structure tells us which part is translated and which part is transliterated.

The backward transliteration method proposed by Lin and Chen (2000, 2002) was followed to select the most similar English named entity and the related documents at the same time. In Lin and Chen's approach, both Chinese name and English candidates will be transformed into a canonical form in terms of International Phonetic Alphabets. Similarity computation among Chinese query term and English candidates are done on phoneme level.

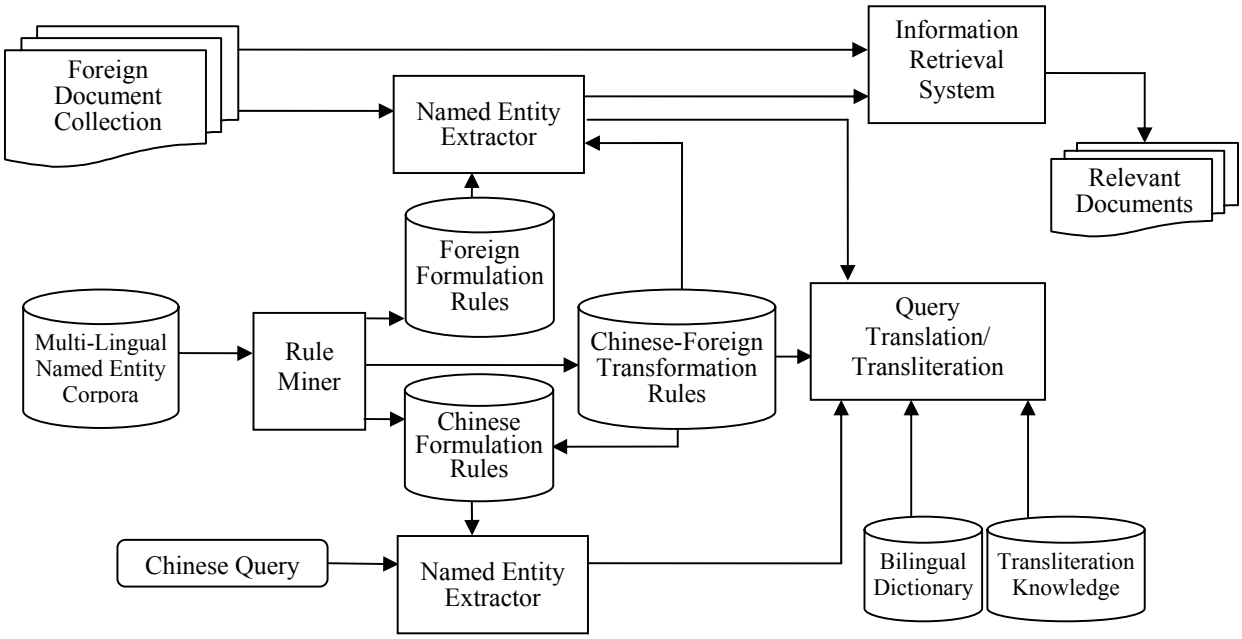

Figure 1. A Chinese-Foreign CLIR System

That is an expensive operation. Hopefully, the type of Chinese named entity will help to narrow down the number of candidate.

6 Conclusion and Remarks

This paper proposes corpus-based approaches to extract the formulation rules and the translation/ transliteration rules among multilingual named entities. Simple frequency-based method identifies keywords of named entities for individual languages and their correspondence. The modified *tf×idf* scheme deals with the issues of abbreviation and compound keyword at a distance.

Since the corpora are already phrase-aligned, the mined rules cover at least a significant number of instances. That is, they seem to be significant, but further evaluation is needed. Two types of evaluation are being conducted, i.e., direct and indirect approaches. In the former, we will partition the corpora into two parts, one for training and the other one for testing. In the latter, we are integrating our method in a cross language information retrieval system. Given a query consisting of Chinese named entity, the Chinese formulation rules will tell us its type and lexical structures. The transformation rules show which parts should be translated and transliterated. Our previous works on phoneme transliteration is integrated. The transformation result may be submitted to an information retrieval system to access documents in another language. In the ongoing evaluation, the test bed is supported by CLEF (2003). The result will be reported in CLEF2003 after evaluation by CLEF organizer. Further applications will be explored in the future and the methodology will be extended to other types of named entities.

References

Al-Onaizan, Yaser and Knight, Kevin (2002) "Translating Named Entities Using Monolingual and Bilingual Resources," *Proceedings of 41st Annual Meeting of Association for Computational Linguistics*, 2002, pp. 400-408.

Chen, Hsin-Hsi and Lee, Jen-Chang (1996) "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.

Chen, Hsin-Hsi; Ding, Yung-Wei and Tsai, Shih-Chung (1998) "Named Entity Extraction for Information Retrieval," *Computer Processing of Oriental Languages*, Special Issue on Information Retrieval on Oriental Languages, **12**(1), 1998, pp. 75-85.

Chen, Hsin-Hsi *et al*. (1998) "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of 17th COLING and 36th ACL*, pp. 232-236.

CLEF (2003) *Cross-Language Retrieval in Image Collections*, Pilot Experiments, 2003.

Hirschman, L.; Park, J.C.; Tsujii, J.; Wong, L. and Wu, C.H. (2002) "Accomplishments and Challenges in Literature Data mining for Biology," *Bioinformatics*, **18**(12), pp. 1553-1561.

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

Lin, Chuan-Jie; Chen, Hsin-Hsi; *et al*. (2001) "Open Domain Question Answering on Heterogeneous Data," *Proceedings of ACL Workshop on Human Language Technology and Knowledge Management*, 2001, pp. 79-85.

Lin, Wei-Hao and Chen, Hsin-Hsi (2000) "Similarity Measure in Backward Transliteration between Different Character Sets and Its Application to CLIR," *Proceedings of Research on Computational Linguistics Conference XIII*, pp. 79-113.

Lin, Wei-Hao and Chen, Hsin-Hsi (2002) "Backward Machine Transliteration by Learning Phonetic Similarity," *Proceedings of 6th Conference on Natural Language Learning*, 2002.

MUC (1998) *Proceedings of 7th Message Understanding Conference*, 1998, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Wan, Stephen and Verspoor, Cornelia Maria (1998) "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources," *Proceedings of 17th COLING and 36th ACL*, pp. 1352-1356.