# Research on the Extraction of Wikipedia-Based Chinese-Khmer Named Entity Equivalents

Qing Xia[1,2], Xin Yan[1,2(✉)], Zhengtao Yu[1,2], and Shengxiang Gao[1,2]

[1] School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650051, China
kg_yanxin@sina.com
[2] The Intelligent Information Processing Key Laboratory,
Kunming University of Science and Technology, Kunming 650051, China

**Abstract.** Named entity equivalent has been playing a significant role in the processing of cross-language information. However limited by the corpora resource, few in-depth studies have been made on the extraction of the bilingual Chinese-Khmer named entity equivalents. On account of this, this paper proposes a Wikipedia-based approach, utilizes the internal web links in Wikipedia and computes the feature similarity to extract the bilingual Chinese-Khmer named entity equivalents. The experimental result shows that good effect has been achieved when the entity equivalents are acquired through the internal web links in Wikipedia with F value up to 90.67%. Also it shows that the result is quite favorable when the bilingual Chinese-Khmer named entity equivalents are acquired through the computation of feature similarity, turning out that the method proposed in this paper is able to give better effect.

**Keywords:** Named entity equivalents · Chinese-Khmer bilingual · Wikipedia · Transliteration model · Translation model

## 1 Introduction

Name entity equivalents have been widely applied in the processing of natural language. In a MTS, the targeted processing of named entities has played an important role in the improvement of the overall translation quality, therefore it's of great application value. Currently most of the researches on the acquisition of named entity equivalents are focused on the adoption of the parallel corpora to mine the named entity equivalents with high accuracy rate having been achieved[1]. In Literature[2], it has proposed such a method to extract the relevant named entity equivalents from the source language corpus. Meng[3] has extracted the named entity equivalents through transliteration, while Huang[4] proposes a multi-feature-based minimum cost approach to extract automatically the named entity equivalents. However since the resource of parallel corpora is extremely limited, it will cost a lot. Compared with the parallel corpora, the comparable corpus can be obtained more conveniently with more abundant contents. Cao[5] proposes an approach to mine the Chinese-English translation equivalence from Chinese web pages. However the effect is not so favorable through such an approach

© Springer International Publishing Switzerland 2015
J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 372–379, 2015.
DOI: 10.1007/978-3-319-25207-0_32

with few entity equivalents having been acquired. Although this method has brought some positive effect when comparable corpus is utilized to acquire the bilingual named entity equivalents[6], it never takes the characteristics of named entity itself into account. However since the Chinese- Khmer bilingual corpus resources are extremely limited on internet, currently no researches have ever been made on the extraction of Chinese- Khmer named entity equivalences.

On Wikipedia, all of the entries are provided with links to the other languages, as the correlation between different languages and the same entity is implied in the organizational structure[7]. Firstly, utilize the characteristic of an entry, which is correlated to different languages on Wikipedia to acquire the correlation between the entry in source language and that in target language, whose entry is the same with that in the source language. Then decode the character according to the correlation with the target language to acquire the Chinese- Khmer named entity equivalences. In order to compensate for the dependency of this internal web link-based extraction method on the link information and to obtain more Chinese- Khmer named entity equivalences, this paper utilizes the characteristic of Wikipedia that different language descriptions on the same entity are comparable corpora to extract the Chinese- Khmer named entity equivalences by computing the feature similarity in transliteration and in translation with the combination of the characteristic that personal name and place name is quite similar in pronunciation and that the organization name is featured with translation characteristics.

## 2 Analysis on the Wikipedia Page Structure

The analysis on the page structure reveals that the entry information in source language is always covered in the web title of the page that contains this entry. Since all of the entries on Wikipedia have been provided with links to the other languages, the link analysis shows that the hyperlink of an entry to the other language always contains the information of the other language for this entry. In this way, both of the Chinese information and the Khmer information of an entry might be found on the same Wikipedia page. Then through the further analysis on the Chinese content and the Khmer content of an entry on the same Wikipedia page, it reveals that the Chinese content and the Khmer content are featured with the following characteristics such as entity co-occurrence and similar description content etc.

## 3 Extraction of Chinese- Khmer Named Entity Equivalences Based on the Internal Links on Wikipedia

The Chinese entries with links to Khmer can be obtained on Wikipedia pages when Chinese is the source language and Khmer is the target language. Take the Chinese entry of “China” as an instance. The web title for this entry on the Chinese web page is “China- Wikipedia, the free encyclopedia” with the header information having been stored between the html tags of <title> and </title> in the page source. Actually the header information for the Entry “China” is “<title> China- Wikipedia, the free encyclopedia</title>” in the page source. However on Wikipedia, the page source of each

entry only consists of an html tag <title></title>. Then the extraction of the information between "<title>" and "- Wikipedia, the free encyclopedia</title>" also indicates the entry information in Chinese.

Since Wikipedia provides different language links to the same entry, then the link of the entries containing Khmer link information will be started with "km" in a webpage source file. For example, the Khmer link to the entry of "China" is indicated as "href=//km.wikipedia.org/wiki/%E1%9E%85%E1%9E%B7%E1%9E%93" in the webpage source file, where the link contains a UTF-8 encoded character string besides the domain name address that is linked to. Followed with the decoded character string, it's the relevant entry in Khmer corresponding to the Chinese entry.

A bilingual Chinese- Khmer entry will be obtained after html tag filtering is conducted on the extracted bilingual Chinese-Khmer information. However such a bilingual Chinese- Khmer word pair always contains some entries that are not named entities. Since numerous researches have been made on the Chinese named entity recognition by now, this paper then adopts the Chinese named entity recognition method to extract the Chinese-Khmer named entity equivalence from the bilingual Chinese-Khmer word pairs. The process for the extraction of Chinese-Khmer named entity equivalences based on links is shown in Figure 1 below.

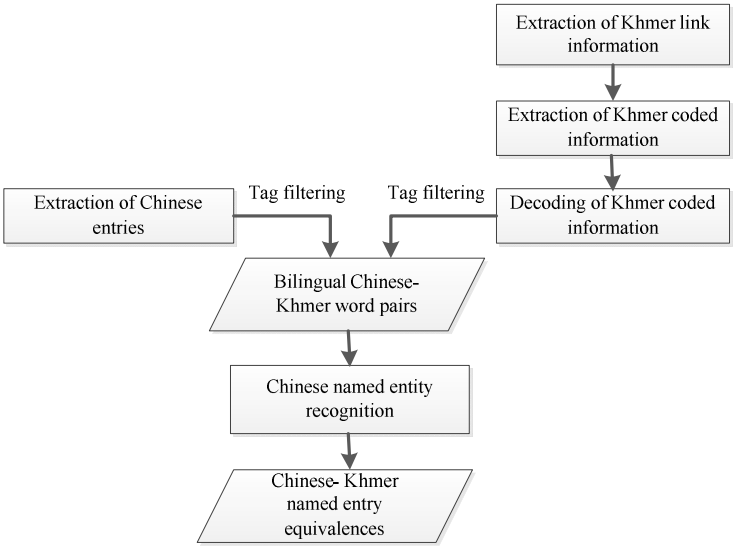


**Fig. 1.** Process for the extraction of Chinese-Khmer named entity trans-lingual equivalence based on links

## 4 Extraction of Chinese- Khmer Named Entity Equivalence Based on Feature Similarity

On Wikipedia, the link-based extraction of named entity equivalence is subject to the link information of this entry in the other languages. However since numerous named

entities on Wikipedia are not expressed by entries, there won't be any link available in the other languages corresponding to them. However they can still be found in the introduction of the entries. Through the analysis on the introduction of an entry in different languages, it reveals that the introduction of the same entry in Chinese and in Khmer is almost the same. Therefore in this paper, such entries are considered as the comparable corpus. Firstly, this paper identifies the named entities in Chinese and in Khmer through the named entity recognition method. Then according to the different types of the named entities, this paper separately utilizes the characteristics of transliteration and translation to compute the similarity in the candidate named entity equivalence for the purpose to obtain finally the Chinese-Khmer named entity equivalence with the specific steps provided as below:

(1) Acquire the textual description of an entry in Chinese on Wikipedia.
(2) Acquire the corresponding textual description of an entry in Khmer according to the hyperlink to this entry in Khmer.
(3) Filter the Chinese and Khmer texts of this entry.
(4) Conduct named entity recognition on the Chinese text.
(5) Conduct named entity recognition on the Khmer text.
(6) Compute the similarity in the candidate named entity equivalences.
(7) Acquire the bilingual Chinese-Khmer named entity equivalences.

### 4.1 Characteristics of Transliteration

Due to the linguistic characteristics of Khmer, there's a high degree of similarity between the production of the personal name and place name in Khmer and the production of them in Chinese. In view of this, this paper would like to acquire the personal name and place name equivalences through the computation on the phonetic similarity in the candidate named entity equivalences. In allusion to the characteristics of transliteration[8-9], this paper utilizes a transliteration probability dictionary obtained through the training of the statistical machine translation model to generate a transliteration probability dictionary.

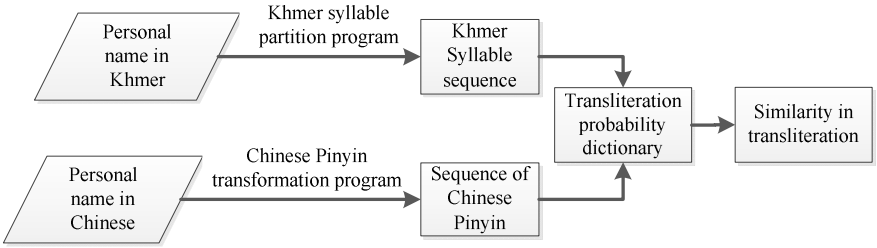


**Fig. 2.** Flow chart for the calculation of similarity in transliteration

As shown in Figure 2, the generated transliteration probability dictionary can be used directly to compute the similarity in the personal name and place name from the Chinese-Khmer named entity equivalences. Three steps are provided as below according to this computing method:

(1) Decompose the candidate personal name or place name equivalence in Khmer into syllable sequences.

(2) Transform the personal name or place name in Chinese into Pinyin sequences.

(3) Use the transliteration probability dictionary to assess the similarity of the candidate Chinese-Khmer personal name or place name equivalence in transliteration.

Regarding the process to calculate the similarity in the candidate Chinese-Khmer personal name or place name equivalence[10], this paper takes the personal name equivalence as an instance. Assume that there's a candidate Chinese-Khmer personal name equivalence, which contains Cp, the personal name entity in Chinese and Kp, the personal name entity in Khmer. Firstly, transform Cp into a Pinyin sequence, $ne_c$= { $c_1,c_2,...c_n$ } . For example, "WangJunJie" can be transformed into "wang/jun/jie". As to Kp, the personal name entity in Khmer, it can be decomposed into kcc syllable sequences, $ne_k$={$k_1,k_2,...k_n$}. For example, "វ៉ាងសង់ជីង (WangJunJie)" can be transformed into "វ៉ាង/សង់/ជីង.Apply the generated nec, the Chinese Pinyin sequence and nek, the Khmer syllable sequence that has been generated to the transliteration probability dictionary so as to calculate the similarity between the candidate Chinese-Khmer name equivalences, which are Cp and Kp in transliteration with the computational formula (1) provided as below:

$$Score(ne_c, ne_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(P(c_i \mid k_j) + P(k_j \mid c_i)\right)}{m+n} \tag{1}$$

Where $ne_c$ is the Pinyin sequence of Cp, the personal name entity in Chinese and n indicates the length. $c_i$ is the $i^{th}$ syllable in the Sequence $ne_c$, $ne_k$ is the syllable sequence of Kp, the personal name entity in Khmer and m is the length. Meanwhile $k_j$ is the $j^{th}$ syllable in Sequence $ne_k$, $P(c_i|k_j)$ is the probability that $c_i$ can be translated into $k_j$, while $P(k_j|c_i)$ is the probability that $k_j$ can be translated into $c_i$.

### 4.2 Translation Features

The organization name from the Chinese-Khmer named entity equivalences isn't featured with the characteristics of transliteration. Therefore this paper utilizes a translation model to determine the similarity between these two organization names. In order to calculate the translation model probability of these two organization names, this paper utilizes the translation model probability applied in the IBM statistical translation model to calculate the similarity between the organization names from the candidate Chinese-Khmer named entity equivalences.

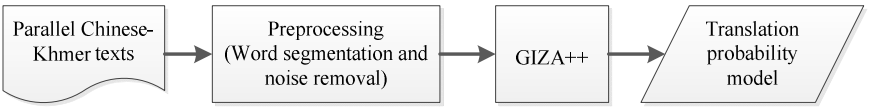


**Fig. 3.** The process for the generation of a translation probability model

Figure 3 shows the flow chart for the generation of a translation probability model with the parallel Chinese-Khmer alignment corpora acting as the input data and the translation probability model acting as the output data.

The steps to train a translation probability model are provided as below:

(1) Pre-process the parallel Chinese-Khmer aligned texts to remove the noise and interference.

(2) Align the words with the help of GIZA++.

(3) Calculate the translation probability.

As shown in Formula (2), use the generated translation probability dictionary to calculate the similarity of the organization names from the candidate Chinese-Khmer named entity equivalences in the translation features:

$$Score_{trans}(ne_c, ne_k) = \frac{2 \times \sum_{i=1}^{n} \sum_{j=1}^{m} \Pr(k_i \mid c_j)}{m+n} \tag{2}$$

Where $ne_c$ represents the named entity in Chinese, n is the number of the words in $ne_c$, the named entity in Chinese, $ne_k$ represents the named entity in Khmer, m is the number of the words in $ne_k$, the named entity in Khmer, $k_i$ is the $i^{th}$ word in $ne_k$ and $c_j$ refers to the $j^{th}$ word in $ne_c$.

## 5 Experiment and Assessment

This paper chooses totally 5000 Chinese entries from Wikipedia including such named entities as personal name, place name and organization name. Then tag manually the named entities from these 5000 Chinese entries, the corresponding Khmer and the entity types, choosing randomly 500 entries, which are also the named entities with Chinese content and with Khmer content as the comparable corpus, where both of the Chinese entities and Khmer entities are tagged.

**Table 1.** Number of Chinese entries in Wikipedia

| Entries | Personal name | Place name | Organization name | Non-entity |
|---|---|---|---|---|
| 5000 | 1448 | 1836 | 1672 | 44 |

**Table 2.** Number of entities in the comparable corpus

| Chinese corpus (500) | | | Khmer corpus (500) | | |
|---|---|---|---|---|---|
| Personal name | Place | Organiza-tion name | Personal name | Place name | Organizati name |
| 361 | 539 | 452 | 212 | 487 | 284 |

**Table 3.** Number of Chinese-Khmer named entity equivalences extracted based on links

| Assessment | Total entities | Personal name | Place name | Organization name |
|---|---|---|---|---|
| Input | 4956 | 1448 | 1836 | 1672 |
| Output | 4238 | 1184 | 1658 | 1396 |
| Recall rate | 85.5% | 81.7% | 90% | 83.4% |
| Rate of accuracy | 96.5% | 89.7% | 99.5% | 98.9% |
| F value | 90.67% | 85.5% | 94.5% | 90.5% |

In the experiment, where the extraction of the Chinese-Khmer named entity equivalences is made based on the links ,as shown in the analytical table 3, the number of the extracted Chinese-Khmer named entity equivalences is 4238, which is less than 4956, the number of the inputted Chinese entities. Through the analysis on the extracted Chinese-Khmer named entity equivalences, it reveals that no equivalences in Khmer have been extracted to correspond to 671 Chinese entities. It's because that no Khmer links have been available on Wikipedia to these 671 Chinese entities. Moreover although 47 pairs of Chinese-Khmer named entity equivalences have been extracted, they fail to be covered in the effective set of named entity equivalence since all of them have been recognized as non-entities during the named entity recognition, including 28 pairs of organization names, indicating that the recognition on the organization name is yet to be improved. Regarding the rate of accuracy, high accuracy is achieved in the place name and organization name compared with the personal name, whose accuracy is quite low. The reason is that family name is not covered in the Chinese name, while it has been attached to the Khmer name after the extraction.

In the experiment, where the Chinese-Khmer named entity equivalence is extracted based on the feature similarity, this paper has extracted separately the personal name, the place name and the organization name based on the characteristics of transliteration and the translation features.

**Table 4.** Performance for the extraction based on feature similarity

| Category | Characteristics | Recall rate | Rate of accuracy | F value |
|---|---|---|---|---|
| Personal name | Transliteration | 65% | 81% | 75% |
| | Translation | 87% | 45% | 59% |
| Place name | Transliteration | 50% | 71% | 59% |
| | Translation | 79% | 75% | 77% |
| Organization name | Transliteration | 7% | 4% | 5% |
| | Translation | 42% | 24% | 31% |

As shown in the analytical table 4, the result is not satisfactory when the extraction of organization name equivalence is made based on the characteristics of transliteration and the translation features. The main reason is that most of the organization names are involved with both of the transliteration and translation. Therefore the pure utilization of the characteristics of transliteration or the translation features won't bring about good effect. Moreover some of the organization names are shown in the abbreviated form, which will also bring some interference to the extraction of equivalences.

## 6 Conclusions

Since the link-based extraction of Chinese-Khmer named entity equivalence is made based on the multi-language characteristic of Wikipedia and the structural features of its web page.Also since the feature similarity-based extraction of Chinese-Khmer named entity equivalence is able to adopt different features based on the different types of named entities, good effect will be achieved in the extraction of personal name equivalence and the place name equivalence. In this way, it will makes up for the deficiency in the link-based approach that depends too much on the link information. However as to the extraction of the organization name, the extraction performance will be improved significantly if both of the characteristics of transliteration and the translation features can be integrated in the extraction.

## References

1. Ru, K., Xu, J., Zhang, Y., Wu, P.: A method to construct chinese-japanese named entity translation equivalents using monolingual corpora. In: Zhou, G., Li, J., Zhao, D., Feng, Y. (eds.) NLPCC 2013. CCIS, vol. 400, pp. 164–175. Springer, Heidelberg (2013)
2. Chen, H.X., Yin, C.Y., Chen, J.J.: An Approach to Extract Nam ed Entity Translingual Equivalence. Journal of Chinese information **22**(4), 55–60 (2008)
3. Meng, H., Lo, W.K., Chen, B., et al.: Generating phonetic cognates to handle named entities in English-Chinese cross- language spoken doeument retrieval. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop, Trento, pp. 311–314 (2001)
4. Huang, F., Vogel, S., Waibel, A.: Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In: Proceedings of Association of Computational linguistics, Sapporo, pp. 9–16 (2003)
5. Cao, G.H., Gao, J.F., Nie, J.Y.: A system to mine large-scale bilingual dictionaries from monolingual web pages. In: Proceedings of MT Summit XI, Copenhagen, Denmark, pp. 57–64 (2007)
6. Lee, L., Aw, A., Zhang, M., et al.: Em-based hybrid model for bilingual terminology extraction from comparable corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 639–646. Association for Computational Linguistics (2010)
7. Yu, K., Tsujii, J.: Bilingual dictionary extraction from wikipedia. In: Machine Translation Summit XII, Ottawa, Canada (2009)
8. Kim, J., Hwang, S., Jiang, L., et al.: Entity Translation Mining from Comparable Corpora: Combining Graph Mapping with Corpus Latent Features (2012)
9. Udupa, R., Saravanan, K., Kumaran, A., et al.: Mint: a method for effective and scalable mining of named entity transliterations from large comparable corpora. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 799–807. Association for Computational Linguistics (2009)
10. Li, L., Wang, P., Huang, D., et al.: Minning English-Chinese Named Entity Pairs from Comparable Corpora. ACM Transactions on Asian Language Information Processing **10**(4) (2011)