# The Technical Analyses of Named Entity Translation

Ying Liu
*Chinese Language and Literature Department, Tsinghua University, Beijing, China, 100084*

## Abstract

There are three methods: rule-based method, statistical method and web mining method for named entity translation. The rule-based method did not achieve satisfactory results. High-quality translation equivalents can be obtained from parallel corpora for statistical method, and a prerequisite is the availability of a large scale of annotated corpora. The comparable corpora are easier to obtain than parallel corpora. But translation extraction from comparable corpora achieves lower accuracy than that of parallel corpora. Web mining method can acquire the translation of high-frequency named entities and it is difficult to translate the low-frequency named entities.

*Key words: named entity translation, transliteration similarity, statistical method, web mining method*

## 1 Introduction

The studied named entity types are person, location and organization names in this paper. The task of named entity (NE) translation is to translate a named entity of the source language into that of the target language. Named entity translation is an important topic in the field of computational linguistics, which is significant for statistical machine translation, cross language information retrieval, cross language information extraction and cross language question answering.

Many researchers have tried to solve NE translation using rule-based method, statistical method and web mining method. The corpus or linguistic resources may be different for different method. A NE dictionary or a list of NE pairs is a base for rule-based translation and statistical transliteration method. A large scale of bilingual corpus is an important resource to align bilingual NEs. Web corpus is an additional resource to acquire more NE translations.

The translation of different NE is highly type-dependent. Chen Yu-feng made an analyses for Chinese-English named entity corpus LDC 2005T34, and found the transliterated person names take up 100 percent of all translated person names, transliterated location names account for 89.4 percent of all translated location names, and transliterated organization names are 12.6 percent of all translated organization names[1]. Most person and location equivalences can be transformed primarily through transliteration, some location and most organization equivalences are transformed by combining both semantic

 

translation and phonetic transliteration[2].

The paper is organized as follows. Section 2 describes the methods of named entity translation, overviews the related work and makes detailed analyses. Section 3 reviews the base of NE translation. Section 4 presents the linguistic granularities for NE translation. Section 5 reports evaluation results, and makes a comparison. Finally we draw the conclusion in Section 6.

## 2 The methods and related work

There are three methods for named entity translation, which are rule-based transliteration method, statistical method and web mining method.

Transliteration is the process of replacing words in the source language with their approximate phonetic or spelling equivalents in the target language. Transliteration between languages that use similar alphabets and sound systems is very simple. However, Transliteration between languages that use different alphabets and sound systems is non-trivial task[3].

### 2.1 Rule-based method

The process of rule-based method from a source named entity A to a target named entity B is as follows: ①Map named entity A (grapheme) to a phonemic representation of A. ②Map each phoneme composing the word A to a corresponding character of B.

The rule-based method uses character-syllable table of source language, sub-syllable table of bilingual transliteration and syllable-character table of target language to translate named entities from one language to another language.

**(1) Related work of rule-based method**

Stephen made use of the rule-based method to transliterate English country names into Chinese names by means of five stages: semantic abstraction, syllabification, sub-syllable divisions, mapping to Pinyin, and mapping to Han characters[4].

**(2) Analyses of rule-based method**

It is challenging to translate named entities across language with different alphabets and sound inventories. Linguistic rules are adopted for determining the translation of NEs. It is hard to select the best translation for the NEs which do not have similar pronunciations with its translation or which translations are ambiguous. So the rule-based method did not achieve satisfactory results.

### 2.2 Statistical method

The statistical method includes statistical transliteration method, parallel corpora-based method, comparable corpora-based method. The current dominant technique for translating or aligning NEs is the statistical method in recent years.

The transliteration knowledge or regularities can be obtained by rule-based method and statistical method, and so forth there are rule-based transliteration method and statistical transliteration method.

#### 2.2.1 Statistical transliteration method

For statistical transliteration method, syllable alignment probabilities are learned based on phonetic equivalents of bilingual named entities and the probabilities are used to translate new NEs. There are two kinds of

transliteration-based statistical method. One is phoneme-based transliteration method[3, 5,6]. Another is grapheme-based transliteration method[3, 6]. There are three conversion steps for phoneme-based transliteration method: grapheme-to-phoneme conversion, phoneme-to-phoneme conversion and phoneme-to-grapheme. While grapheme-based transliteration method directly maps source characters into target characters, which is also called as direct orthographical mapping.

**(1) Related work of Statistical transliteration method**

Some researches focused on the statistical transliteration method [3,5,7]. Asif transliterated person names from Bengali to English using modified source-channel model, which made use of the linguistic knowledge of possible conjuncts and diphthongs in Bengali and their equivalents in English[7]. The phrase-based model and the N-gram mode regard person name translation as machine translation. Yaser[3] combined statistical transliteration method, parallel corpora-based method and web-mining method. For person names, Yaser used statistical transliteration method. The transliteration score was a linear combination of the phonetic-based and the spelling-based transliteration scores. The phonetic-based transliteration score was associated with the following three probabilities: the probability of generating an English word, the probability the English word is pronounced, and the probability English phoneme sequence is converted into Arabic writing, which is called as phoneme-based transliteration method. The spelling-based transliteration score was associated with the following two probabilities: the probability of mapping English letter sequences into Arabic letter sequences and the probability of generating an English word, which is grapheme-based transliteration method.

**(2) Analyses of Statistical transliteration method**

It is difficult to translate the NEs which are not translated according to phonetic equivalents. There is significant flexibility in transliteration generation of foreign names in real world, and transliteration selection is somewhat subjective. Hence, translation only relying on the statistical transliteration probabilities may not work well.

**2.2.2 Parallel corpora-based method**

Parallel corpora-based method is to translate source NEs into target NEs using parallel corpora. There are two methods to translate NEs of parallel corpora. The first method involves recognizing NEs of bilingual corpora and then aligning NEs of bilingual corpora[1,2,8]. The second method involves recognizing NEs of source corpora and then translating source NEs to target NEs[3].

Multi-features are calculated from the parallel corpora and used for translating new named entities for the parallel corpora-based method. The features mainly include the transliteration similarity, mutual information, alignment probability and semantic similarity and so on. The statistical models include maximum entropy model, hidden Markov model, conditional random fields and joint source-channel model and so on.

**(1) Related work of parallel corpora-based method**

Fei Huang[2,8], Chen Hua-xing[9],Zhang min[6], Li[10] and Chen YuFeng focused on finding NE correspondence in parallel corpora[1]. Fei Huang's named entity alignment model incorporated named entity transliteration cost, word-based named entity translation cost and named entity tagging cost to align bilingual named entity equivalences. He combined phonetic similarities with semantic similarities to translate named entities. Chen YuFeng utilized paraphrase, transliteration and co-occurrence feature to make basic alignment[1], while named entity boundaries and type could be corrected by means of corrective alignment module. Basic alignment combined semantic feature, transliteration feature and co-occurrence feature. Corrective alignment realized the joint of NE recognition and alignment, which combined the probabilities of Chinese sequence's classification and the probabilities of English sequence's classification. Li Tingting used the conditional random field to recognize Japanese names, and kana names were translated using Moses machine translation system[10]. Zhang Min proposed a phrase-based context-dependent joint probability model for named entity translation [6]. Named entity translation is similar to phrase-based statistical machine translation: target phrases were generated by the lexical mapping model and word reordering was performed by the permutation model at the phrase level.

**(2) Analyses of parallel corpora-based method**

High-quality translation equivalents can be obtained from parallel corpora, and a prerequisite is the availability of a large scale of annotated corpora. So it is fit for building a reliable NE dictionary according to a large scale of parallel corpora. Such corpora are available from the evaluation forums but remain rare and limited in domain and language coverage. It is not a trivial task to obtain large-scale parallel corpora, especially for uncommon language pairs. The quantity of the translation results depends heavily on the scale and coverage of the corpora.

**2.2.3 Comparable corpora-based method**

Comparable corpora-based method is to translate source NEs into target NEs using comparable corpora. Entity similarity, entity context similarity, relationship similarity and relationship context similarity may be computed through comparable corpora and employed to translate NEs.

**(1) Related work of Comparable corpora-based method**

Jinhan Kim[11], Taesung Lee[12] and You Gae-won[13] tried to find person name correspondence from comparable corpora. Comparable corpora refer to those texts that are not translations of each other but talk about the same or related topics. Shao combined entity similarity and entity context similarity. Shao computed entity similarity using a probabilistic pronunciation model and computed entity context similarity using a language model[14]. You Gae-won combined entity similarity and relationship similarity[13]. He used edit distance between Chinese Pinyins and English strings for computing entity similarity and monolingual entity co-occurrences for computing relationship similarity. JinHan

Kim combined entity similarity, entity context similarity, relationship similarity and relationship context similarity. JinHan Kim used edit distance to compute the entity similarity between named entities. He also used the cosine similarity between context vectors to compute the entity context similarity. The relationship similarity combined entity similarity and entity context similarity and the relationship context similarity was computed as the cosine similarity between two context association vectors[11].

**(2)Analyses of comparable corpora-based method**

The comparable corpora are easier to obtain than parallel corpora. But translation extraction from comparable corpora achieves lower accuracy than that of parallel corpora.

**2.3 Web mining method**

In order to make full use of large scale of web corpora, web mining method is proposed for named entity translation. There are three basic steps for web mining method. ①Obtain relevant web pages containing the input word. In order to obtain the bilingual web pages containing both the input and its translation, the input and clue words need to be sent to the search engine. A character of the input's translation or the target words co-occurring with the input might be clue words. For example, Sarah Brightman is an English singing star, so Sarah Brightman often co-occurs with star or singer. The translations of star and singer are the clue words for Sarah Brightman. If a English named entity is sent to the search engine, monolingual web pages which don't contain any target word will be always obtained, so the translation of the named entity cannot be obtained. ②Extract translation candidates according to statistical measures. ③Rank the candidates using the statistical model.

**(1) Related work of web mining method**

Zhang Yong-chen[15], Jiang Long[16], Guo Ji[17], Fei Huang[18], Jian-Cheng Wu[19], Fan Yang[20] and Zhao Mingming[21] translated person names using web mining method. Guo Ji proposed a statistical discriminative model to extract translation pairs from Chinese web corpora[17]. Zhang Yong-chen extracted the bilingual dictionary for the special domain based on web corpora with the word relation matrix[15]. Jiang Long acquired translation candidates by combining mutual information, symmetric conditional probability, context dependency and anchor word[16]. He made use of maximum entropy model to rank translation candidates by combining transliteration similarity, the Chi-Square, and bilingual context co-occurrence feature. Fei Huang presented a new framework to mine key phrase translations from web corpora[18]. Fei Huang expanded queries by adding the translations of topic-relevant hint words, retrieved mixed language web pages and extracted the key phrase translation with phonetic, semantic and frequency distance features. Jian-Cheng Wu presented a method to learn source-target surface patterns for web-based terminology translation[19]. The method involves submitting a given term to a search engine, extracting the candidate translations from the returned summaries and subsequently ranking the candidate translations based on the surface patterns,

occurrence counts, and transliteration knowledge. Fan Yang translated Chinese organization names into English equivalence using heuristic query and asymmetric alignment [20]. Zhao Mingming made use of a transliteration model to generate top n translation candidates and used weighted frequency algorithm to extract expansion words from the top n translation candidates, then transliteration feature and co-occurrence feature to rank the translation candidates[21]. Yaser[3] scored organization and location names using a modified IBM model 1, and re-scored candidates by combining straight web count, co-reference and contextual web count.

**(2) Analyses of web mining method**

Web mining method might help to find the translations of more named entities and multi-translations of some names entities. Web mining method can acquire the translation of high-frequency NEs and it is difficult to translate the low-frequency NEs because of the restriction of returned web page and statistic measures for selecting and sorting of translation candidates. So the recall rate of web-mining method may not high. Translation candidates of low-frequency NEs might be obtained via the transliteration similarity[16].

## 3 Base of NE translation

Dictionaries, parallel corpora, comparable corpora and web corpora are base of NE translation. Different dictionaries or corpora may be used for different method or different model.

**(1) NE Dictionary and NE equivalence pairs**

NE bilingual dictionary and NE equivalence pairs are used for phoneme-to-phoneme conversion and character-to-character conversion for transliteration method. NE bilingual dictionary or NE equivalence pairs may be utilized by rule-based transliteration method and statistical transliteration method. Yaser used English-Arabic name list to map English letter sequences into Arabic letter sequences[3]. Yu Heng utilized 40000 Chinese-English name entity lists from LDC2005T34[22].

NE dictionaries with pronunciation are also used for grapheme-to-phoneme conversion and phoneme-to-grapheme conversion for rule-based transliteration method and phoneme-based transliteration method. Kevin utilized on-line CMU pronunciation dictionary consisting of pronunciations of 110,000 words and 8,000 pairs of English/Japanese sound sequences for phoneme-based transliteration method[5]. Yaser used an English pronunciation dictionary to produce the probability of the English word’s pronunciation[3].

Besides, NE dictionaries or NE pairs are used for computing the transliteration similarity. Fei Huang trained Chinese-English surface string transliteration model using the Chinese-English dictionary version 3.1 released by LDC[8]. Jiang used 24,718 person names from LDC2003E01 and CMU pronunciation dictionary to compute pronouncing similarity[16].

**(2) Parallel corpora**

NEs of Parallel corpora need to be tagged if we hope to make full use of the parallel corpora to translate source NEs into target NEs. If NEs of parallel

corpora are all tagged, the translation of NEs is to align NEs of parallel corpora [2,8] . If NEs of source corpora are tagged and NEs of target corpora are not tagged, the translations of source NEs may be found by means of statistical model[9].

The bilingual corpus of Fei Huang[8] contains 152,391 sentence pairs from Xinhua News Agency and the Foreign Broadcast Information Service. NEs in the bilingual corpus are first annotated and then aligned according to the multi-feature cost minimization framework. Chen YuFeng used Chinese-English NE pairs corpus LDC2005T34 and Chinese-English news corpus LDC2005T06 for maximum entropy model of basic alignment[1]. LDC2005T34 was used for the corrective alignment. Zhang Min[6] used LDC Chinese-English NE translation corpus. A bilingual corpus aligned in the source language order is used to train lexical mapping model, and a target language corpus with phrase segmentation in their original word order is used to train permutation model. Chen Huaixing collected smaller scale of Chinese-English corpora containing 1500 sentence pairs and larger scale of Chinese-English corpora containing 202174 sentence pairs[9]. Named entities are tagged only for Chinese corpora. He extracted equivalence of named entities from the smaller corpus and the larger corpus.

**(3) Comparable corpora**

Taesung Lee[12] and Jinhan Kim[11] used the English Gigaword Corpus(LDC2009T13) containing 100,746 news documents from January 2008 to December 2008. The Chinese corpora containing 88,029 news documents are during the same period. Shao used the same dataset[14](but with different time period) as Jinhan Kim[11].

**(4) Web corpus**

Web mining method introduces the web resources into NE translation. Yaser[3] and Guo[17] used the web knowledge to assist NE translation and Zhang [15], Jiang[16], Fei[18], Fan[20] and Zhao[21] extracted the translation equivalents from web pages directly.

## 4 Linguistic granularities

Stephen's rules were based on syllables[4]. The statistical translation model can be built on the granularity of phoneme[3,5], syllable[16, 22] , grapheme [3, 6, 22], character[23], word [2,8,9, 16], phrase [6] , structure[9], POS[23] Semantic prediction[22] and NE annotation[2,3,8].

Researchers combined many granularities in their models. Kevin combined phonemes and words[5]. Stephen used English-Chinese unitary consonant correspondences, consonant pairs, double consonant correspondences, English phoneme-Chinese Pinyin mapping table, and Chinese Pinyin to Han table[4]. Zou used character and its tagging, transition of character tagging, character string and its tagging, and transition of character string tagging for maximum entropy model and conditional random fields[23]. Chen combined the annotation of Chinese and English NEs , context word, Character and Chinese and English word sequence's classifications[1]. Li combined word, POS,

sentence constituent, tagging categories of Japanese characters and suffix of Japanese person names[10]. Min[6] used the phrase pairs in the lexical mapping model and the target phrases in the permutation model. Yu utilized graphemes, syllables, and syllable annotations for English-Chinese person transliteration[22].

## 5 The evaluation of named entities translation

Precision(P), recall(R), F measure(F) and accuracy(A) are often used for the evaluating named entities translation. Table 1 shows the evaluation results of four methods.

From table 1, we find that F of parallel corpus method is higher than other methods. The accuracy of web mining method[17,18] are higher. Except person names of Tasung, NEs' accuracy and precision of transliteration method and comparable corpus method are lower.

Table 1 the evaluation comparison of four methods

| | | A% | P% | R% | F% | | | A% | P% | R% | F% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Parallel Corpus method | [9] | 83.5 | | | | Compa-rable corpus method | [14] | | 52.5 | 30.5 | 38.6 |
| | [8] | | 73.8 | 90.5 | 81 | | [13] | | 69.9 | 53.7 | 60.7 |
| | [6] | 51.5 | | | | | [12] person | | 81 | 79 | 80 |
| | [10] | 85.8 | | | | | [12] | | 56 | 52 | 54 |
| | [1] | | 79.1 | 84.3 | 81.2 | | [11] | | 67 | 62 | 64 |
| Web mining method | [19] | | 89 | 53 | 66 | Transli-teration method | [5] | 64 | | | |
| | [18] | 80 | | | | | [3] | 65.2 | | | |
| | [17] | 82.1 | | | | | [22] | 64.3 | | | |
| | [20] | | 48.7 | | | | | | | | |

In addition to the above evaluation metrics, the word error rate, BLEU score and NIST score[6], top-1 inclusion[18,21] and coverage[16] are also used for evaluating named entities translation. Zou reported that the performance of the N-gram model was the best of four models[23]. Min reported the accuracy of the exact for E2C open test and closed test were 51.5% and 90.9%[6]. The accuracy of the exact for C2E open test and closed test were 36.1% and 81.3%. Yaser reported the accuracies of the top 1 named entities for Development Test Set and Blind Test Set are 65.20% and 72.57%[3].

## 6 Conclusions

The methods of named entity translation are mainly discussed and compared in this paper. There are three methods: rule-based method, statistical method and web mining method for named entity translation. The dictionary, parallel corpus, comparable corpus and web are the base of named entity translation. Named entities may be translated using different granularities: phoneme, syllable, grapheme, character, word, phrase, and structure. The translation accuracies or precisions of parallel corpus-based

method and web mining method are higher than those of statistical transliteration method and comparable corpus-based method.

## 7 Acknowledgments

This work was financially supported by national natural science foundation(61171114) and Independent scientific research plan from the Ministry of Education(20111081010).

## References:

[1] Chen Yu-Feng, Zong Cheng-Qing and Su Keh-Yih: Joint Chinese-English named entity recognition and alignment. Chinese Journal of Computers. 34(9):1688-1696(2011).

[2] Fei Huang, Stephan Vogel and Alex Waibel:Improving Named Entity Translation Combining Phonetic and Semantic Similaritie. Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics. 281-288(2004).

[3] Yaser Al-Onaizan and Kevin Knight: Translating named entities using monolingual and bilingual resources. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.400–408(2002).

[4] Stephen Wan and Cornelia Maria Verspoor: Automatic English-Chinese name transliteration for development of multilingual resources. Proceedings of the 17th international conference on Computational linguistics. Vol 2.1352-1356(1998).

[5] Kevin Knight and Jonathan Graehl: Machine transliteration. Computational Linguistics. 24(4):599-612(1998).

[6] Min Zhang, Haizhou Li and Jian Su, et al: A Phrase-Based Context-Dependent Joint Probability Model for Named Entity Translation. IJCNLP, volume 3651 of Lecture Notes in Computer Science. 600-611(2005).

[7] Asif Ekbal, Sudip Kumar Naskar and Sivaji Bandyopadhyay: A Modified Joint Source-Channel Model for Transliteration. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions .191-198(2006).

[8] Fei Huang, Stephan Vogel and Alex Waibel: Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition. 15:9-16(2003).

[9] Chen Hua-xing, Yin Cun-yan and Chen Jia-jun: An Approach to Extract Named Entity Translingual Equivalence. Journal of Chinese Information Processing. 22(4):55-60(2008).

[10] Li Tingting,Zhao Tiejun and Zhang Chunyue: Statistical Japanese Names Recognition and Translation. Intelligent Computer and Applications.2(1):4-7(2012).

[11] Jinhan Kim, Long Jiang and Seung-Won Hwang et al: mining entity

translations from comparable corpora: a holistic graph mapping approach. Proceedings of the 20th ACM international conference on Information and knowledge management.1295-1304(2011).
[12] Taesung Lee and Seung-won Hwang:Bootstrapping Entity Translation on Weakly Comparable Corpora. The 51st Annual Meeting of the Association for Computational Linguistic.4-9(2013).
[13] You Gae-won, Hwang Seung-won and Song Young-in, et al: Efficient Entity Translation Mining-A Parallelized Graph Alignment Approach. ACM Transactions on Information Systems. 30(4):1-23(2012).
[14] Shao L. and H.T.Ng: Mining new word translations from comparable corpora. Proceedings of the 20th international conference on Computational Linguistics.(2004).
[15] Zhang Yongchen, Sun Le and Li Fei, et al: Bilingual Dictionary Extraction for Special Domain Based on Web Data. Journal of Chinese Information Processing.20(2):16-23(2006).
[16] Jiang Long, Zhou Ming and Jian Lifeng: Named Entity Translation with Web Mining and Transliteration. Journal of Chinese Information Processing. 21(1):23-29(2007).
[17] Guo Ji, Lv Ya-juan and Liu Qun: An Effective Method to Extract Translation Pairs from Web Corpora. Journal of Chinese Information Processing. 22(6):103-109(2008).
[18] Fei Huang, Ying Zhang and Stephan Vogel: Mining key phrase translation from web corpora. Proceedings of HLT/EMNLP-2005.483-490(2005).
[19] Jian-Cheng Wu and Jason S. Chang: Learning to Find English to Chinese Transliterations on the Web. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.996–1004(2007).
[20] Fan Yang, Jun Zhao and Kang Liu: A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. 387-395(2009).
[21] Zhao Mingming, Hong Yu and Yao Jianmin, et al: Research on Name Entity Translation Based on Transliteration and Web. Proceedings of the sixth National Conference on Informational Retrieval. 357-366(2010).
[22] Yu Heng, Tu Zhaopeng and Liu Qun,et al: Lattice-based Multi-granularity Name-Entity Machine Transliteration. Journal of Chinese Information Processing.27(4):16-21(2013).
[23] Zou Bo and Zhao Jun: Comparison of Several English-Chinese Name Transliteration Methods. Proceedings of the fourth Student Symposium on Computational Linguistics. (2008).