Chapter 6: Experiments on NE Alignment

This chapter describes the experimental setup and performance evaluation of the proposed approach to bilingual NE alignment in parallel corpora.

6.1 Experimental Setup

Several corpora were collected to estimate the parameters of the proposed models. Noun phrases from the BDC Electronic Chinese-English Dictionary (BDC, 1992) were used to train a general SPTM. A bilingual organization name corpus and a bilingual location name corpus compiled by Central News Agency (CNA, 2003) and Britannica Concise Encyclopedia (BCE, 2003), respectively, were used to train an NE-specific SPTM. A list of NE pairs collected from *Sinorama* Magazine was used to train a domain-relevant SPTM. To train TM, 2,430 pairs of English names together with their Chinese transliterations (Huai, 1989) and Chinese romanization tables were used. To train CPNR, a Chinese person name corpus containing one million Chinese person names was used. *Sinorama* covers a wide range of topics, including

personalities, places, and events in Taiwan. In Table 6.1, we report the statistics of the selected corpus derived from *Sinorama*.

Table 6.1 Statistics of the *Sinorama* corpus.

Dates	Aligned Sentences	English Words	Chinese Characters
1995-2002	50,000	2,420,000	2,534,000

The NE alignment performance was evaluated according to the precision rate at the NE phrase level:

$$Phrase\ Precision = \frac{number\ of\ correctly\ aligned\ NE\ pairs}{number\ of\ correct\ NE\ pairs}. \quad (6.1)$$

To analyze the performance of the proposed methods for NE alignment, we randomly selected 500 aligned sentences from *Sinorama* and manually labeled the answer keys. Each chosen aligned sentence contains at least one NE pair. Currently, we restrict the lengths of English NEs to be less than 6 words. In total, 1432 pairs of NEs were labeled. The numbers of NE pairs for types PER, LOC, and ORG were 380, 522, and 530, respectively. Table 6.2 shows the statistics of these bilingual NE pairs.

Table 6.2 Occurrence statistics for bilingual NE pairs in the *Sinorama* test set.

NE Type	PER	LOC	ORG
Total Occurrence	380 (26.54%)	522 (36.45%)	530 (37.01%)
Unique Occurrence	352 (31.18%)	327 (28.96%)	450 (39.86%)

6.2 Experimental Results and Discussion

Several experiments were subsequently conducted to analyze the performance enhancement achieved with the proposed methods. Moreover, for the purpose of comparison with established baselines, we evaluated NE alignment against IBM Model 4 (Brown et al., 1993) using the toolkit Giza++[3] (Och and Ney, 2003), which is a publicly available implementation of the IBM models. Experimental results presented here are based on the evaluation criterion mentioned above, as shown in Table 6.3.

Table 6.3 Performance in bilingual NE alignment with the *Sinorama* test set.

Method	PER	LOC	ORG	Average
SPTM+TM (baseline)	85.79%	93.87%	75.66%	84.99%
SPTM+TM+AE	85.79%	93.87%	78.30%	85.96%
SPTM+TM+AE+AH	85.49%	96.17%	84.34%	88.97%
SPTM+TM+AE+CPNR	93.42%	93.87%	78.30%	87.99%
SPTM+TM+AE+AH+CPNR	93.42%	95.79%	84.91%	91.13%
IBM Model 4	29.47%	56.51%	37.92%	42.46%

[3] http://www.fjoch.com/GIZA++.html.

In Table 6.3, the results reveal that each added method has different contributions to different NE types. From the above data, we can make the following observations:

1. AH and AE contributes to ORG remarkably due to the fact that AH and AE help to deal with ORG-type abbreviations occurring in both Chinese and English.
2. CPNR contributes to PER significantly since CPNR attempts to solve the problem of mapping a foreign name to its corresponding Chinese.
3. The performance of ORG is the worst of all. The major reasons are its highly complex structure and great variety. ORG-type NEs are also longer than other types of both English NEs and Chinese NEs, which is also a risk factor for transforming ORG-type NEs. Table 6.4 shows the average lengths of the NE types for the answer set.
4. Each individual method consistently helps to improve the baseline for all types of NEs. Moreover, the approach with all knowledge sources achieves much better results than any other system with partial knowledge sources.
5. The proposed approach significantly outperforms the traditional approach of IBM Model 4.

Table 6.4 Average lengths of the NE types for the answer set in the *Sinorama* test set.

NE Type	PER	LOC	ORG
Avg. Length in words (English NE)	1.81	1.58	2.99
Avg. Length in characters (Chinese NE)	2.79	2.70	4.74

More specifically, Table 6.5 shows some examples that demonstrate the performance enhancement achieved by adding more knowledge sources. For simplicity, we only focus on certain NEs that are underlined. The improvements of the work are explained as follows:

1. In example (1), after aligning the anchor point (Liu, 劉 "Liu"), CPNR is activated to successfully detect the first name (Hamilton, 國芊 "Kou Chien"), even though "國芊" cannot be directly transformed from "Hamilton" via transliteration or translation.
2. In example (2), similarly, the surname (Lee, 李 "Li") is located first, then the Chinese given name "遠哲 (Yuan Tse)" is detected by CPNR, though we don't have the full name information from "Yuan T. Lee."

3. In example (3), the set of possible translations of “South Africa” can be {“南非共合國,” “南方非洲,” “南邊非洲,”…}. By applying AH, the abbreviation “南非” can be extracted from the aligned Chinese text.

4. In example (4), the set of high ranking translations of “Taipei First Girls' High School” via SPTM is {“台北第一女孩高學校,” “台北第一女子高學校,” “台北第一女子高級學校,” …}. The correct complete translation is “台北第一女子高級中學.” Even though the complete translation is not generated due to the lack of lexicon coverage, the Chinese abbreviation “北一女” can still be well approximated via AH.

5. In examples (5) and (6), these two examples demonstrate the alignments between acronyms and translations. In example (5), instead of finding “WTO” in the bilingual dictionary, we expand “WTO” to “World Trade Organization” via the acronym-expansion list (in Table 3.3) and then subsequently translate “World Trade Organization” into “世界貿易組織.” In example (6), “Council of Labor Affairs” is an expansion of “CLA,” and one of its translations, “勞工事務委員會,” can be aligned with “勞委會” via the proposed AH.

Table 6.5 Examples of possible Chinese NEs extracted by the proposed approach.

Example	Bilingual Aligned Texts	Results by the SPTM+TM Method	Improved Results	Method(s) Used
(1)	After numerous passenger protests, Hamilton Liu, Deputy Director of the China Airlines Public Relations Office, says,… 經歷了多次旅客抗議事件，華航公關室副主任劉國芊也說，…	劉	劉國芊	Improved by invoking CPNR
(2)	Yuan T. Lee and Chang-lin Tien, two renowned overseas Chinese associated with the University of California at Berkeley, are both outstanding figures in academia yet are very different individuals. 李遠哲與田長霖，這兩位柏克萊加州大學最受矚目的華人，同樣傑出，卻是很不一樣的人。	李	李遠哲	Improved by invoking CPNR
(3)	The writer was a woman named San Mao who had married a Spaniard and moved to Africa. Some 20-odd travel memoirs followed describing the primeval forests of South Africa, the plains of western Africa, … 作者是遠嫁非洲的女作家三毛，從此之後，從南非原始森林、美洲亞馬遜的熱帶雨林、西非大草原…	非洲	南非	Improved by invoking AH
(4)	"When the travel agency notified us that we'd been accepted by the 'Royal Canadian College,' I was really proud…" says Kung Hsi, who passed up 11th grade at the Taipei First Girls' High School evening class division to come to Vancouver to study with her younger brother. 「當初旅行社通知說申請到『皇家學院』，我好得意哦，還到學校吹噓；…」，原本就讀北一女夜間部，高二時和弟弟一起到溫哥華當小留學生的孔曦吐吐舌頭表示。	學校	北一女	Improved by invoking AH
(5)	As for direct trade, after entry into the WTO, we will have to face liberalization anyway, so the government should face up to the problem directly. 而通商在未來加入世界貿易組織後，也必須面臨開放，因此政府亦應予正視。		世界貿易組織	Improved by invoking AE
(6)	According to statistics of the CLA, nearly 30,000 local households have hired housekeepers, of which foreign nationals constitute two-thirds. That means that 20,000 households more or less now rely on foreign housemaids to look after the children and the home. 根據勞委會統計，目前國人家中雇有女傭的家庭將近三萬戶，其中外籍約佔三分之二，亦即二萬戶左右的家庭已在仰賴外籍女傭照顧幼兒及管家。		勞委會	Improved by invoking AH and AE

To investigate whether the proposed approach is sensitive to the change of domain, we also conducted experiments on the LDC[4] parallel corpus of the *Hong Kong News Parallel Text* (abbreviated as *HKNPT*). First of all, to analyze the performance of the proposed methods for NE alignment, we randomly selected 400 aligned sentences from *HKNPT* and manually labeled the answer keys. In total, 893 pairs of NEs were labeled. The numbers of NE pairs for types PER, LOC, and ORG were 168, 348, and 377, respectively. Tables 6.6 and 6.7 show the relevant statistics of these bilingual NE pairs.

Table 6.6 Occurrence statistics for bilingual NE pairs in the *HKNPT* test set.

NE Type	PER	LOC	ORG
Total Occurrence	168 (18.81%)	348 (38.97%)	377 (42.22%)
Unique Occurrence	149 (25.17%)	190 (32.09%)	253 (42.74%)

Table 6.7 Average lengths of the NE types for the answer set in the *HKNPT* test set.

NE Type	PER	LOC	ORG
Avg. Length in words (English NE)	1.99	2.17	2.73
Avg. Length in characters (Chinese NE)	2.81	2.95	4.84

[4] http://www.ldc.upenn.edu/.

Since Cantonese transliterations of Chinese characters are quite different from Mandarin transliterations of Chinese characters, two Cantonese romanization systems[5] were adopted in the experiment. The experimental results are shown in Table 6.8.

Table 6.8 Performance in bilingual NE alignment with the *HKNPT* test set.

Method	PER	LOC	ORG	Average
SPTM+TM (baseline)	51.79%	76.72%	54.91%	62.82%
SPTM+TM+AE	51.79%	80.75%	66.84%	69.43%
SPTM+TM+AE+AH	51.79%	83.62%	72.94%	73.12%
SPTM+TM+AE+CPNR	83.93%	80.46%	66.58%	75.25%
SPTM+TM+AE+AH+CPNR	84.52%	83.91%	74.80%	80.18%
IBM Model 4	40.48%	83.33%	74.01%	71.33%

As shown in Table 6.8, without fine-tuning the phrase translation model, our approach outperforms IBM model 4 on average. The performance of the proposed approach is significantly better than IBM model 4 on PER-type NEs and competes with IBM model 4 on LOC-type and ORG-type NEs. To highlight the performance enhancement achieved with the proposed methods, we also plot the performance data in Figure 6.1.

[5] Details about Cantonese romanization schemes we used can be found at "http://www.info.gov.hk/digital21/eng/structure/jyutping.html" and "http://home.netvigator.com/~spikel/canton.txt."

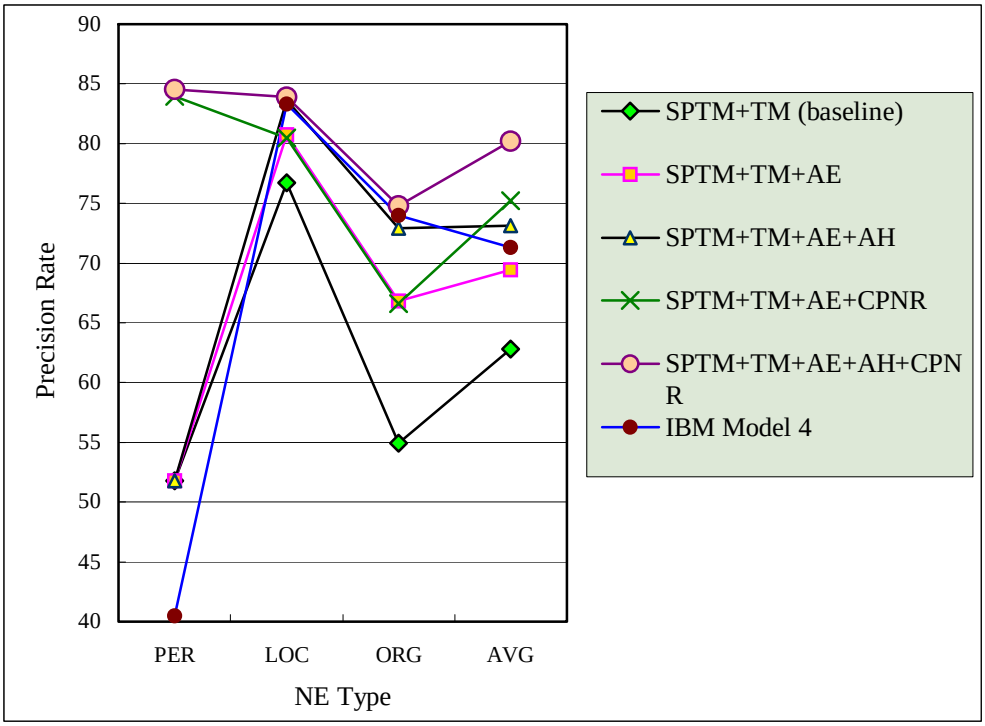

Figure 6.1 Plot of the statistics in Table 6.8.

Although most NE pairs were extracted correctly from the test corpora, some NE pairs were not, as shown in Table 6.9. For simplicity, only erroneously aligned NE pairs are underlined in the table. These errors are explained as follows:

1. In example (1), the proposed TM fails to extract "阿爾讓特 (A Erh Jang Te)," since the transliteration score of the pair (Argenteuil, 阿爾讓特) is too low to exceed the threshold. This is due to the fact that "Argenteuil" is French, not English.

2. Another type of error occurs due to the fact that some NEs are not transformed literally. For instance, in example (2), (Osteogenesis Imperfecta Association, 玻璃娃娃協會) cannot be translated by translating the individual words (glass, 玻璃), (doll, 娃娃), and (association, 協會).

Obviously, the semantic meanings of “osteogenesis” and “imperfecta” are quite different from meanings of “glass” and “doll,” respectively, even though we have these translations of “osteogenesis” and “imperfecta” in the employed dictionary.

3. In example (3), (“Pan Viet,” 越盛公司) cannot be identified correctly, since “Viet” is much closer to “為 (Wei)” than to “盛 (Sheng),” based on the similarity at the grapheme level. In fact, “Pan Viet” is a foreign language name of “越盛公司 (Yueh Sheng Company).” Certainly, the proposed approach has difficulty solving this case, where the NE mapping is not transformed through the combination of transliteration and translation.

4. In examples (4), (6), and (7), the errors are caused by the limited coverage of the lexicons we used. Currently, the employed bilingual dictionary does not have these translations (Vetting, 評審), (Pratas, 東沙), and (Tainan, 府城). In particular, the most frequent translation or transliteration of “Tainan” is “台南 (Tai Nan)” or “臺南 (Tai Nan),” while “府城 (Fu Cheng)” is an old name of “台南,” which is misaligned to “探訪 (Tan Fan)” based on the similarity at the grapheme level via the proposed TM. Of course, if we can incrementally add vocabulary entries to the employed dictionary, the performance will be further improved.

5. In example (5), (Tsui Hark, 徐克 “Hsu Ko”) fails to be extracted due to the anchor point “Tsui” is mapped to the Chinese last name “崔 (Tsui)” instead of “徐(Hsu),” based on the WG romanization system. And the first name “Hark” is close to “豪氣 (Hao Chi)” instead of “克 (Ko),” based on the transliteration score. Therefore, “Tsui Hark” is aligned to “豪氣,” not to “徐克.” This error is inevitable since “Tsui Hark” is the transliteration of “徐克” in Cantonese instead of Mandarin. One way to solve this problem is to adopt Cantonese romanization tables.

6. In example (8), the correct Chinese NE “尤曾家麗 (Yau Jang Ga Lai)” consists of the surname “尤曾” and the given name “家麗.” (Carrie Yau, 尤曾家麗) was not extracted, because the anchor point “Yau” was first aligned with the Chinese surname name “尤” instead of “尤曾.” Thus, CPNR was applied to detect the subsequent two characters “曾家” instead of “家麗.”

Table 6.9 Examples of alignment errors made using the proposed approach.

Example	Type	Bilingual Sentences	Correct NE Pairs	Miss-aligned Chinese NEs	Corpus
(1)	LOC	At the end of the 19th century, French urbanites were well accustomed to the comforts of life in a modern city, and liked to make outings by train to destinations like the newly popular Seine town of Argenteuil. 在上一個世紀的交替，歐洲的法國市民已經過著工業化與現代城市的生活。人們坐著火車到近郊休閒，像是塞納河沿岸的阿爾讓特港就是新興的休閒市鎮。	(Argenteuil, 阿爾讓特 "A Erh Jang Te")		*Sinorama*
(2)	ORG	Lin Yu-chih, founder of the Osteogenesis Imperfecta Association, says that the law should not only reduce the financial burden on patients' families, but also… 「玻璃娃娃協會」發起人林煜智指出，法案通過，一方面減輕病患家庭的負擔，也…	(Osteogenesis Imperfecta Association, 玻璃娃娃協會)	協會	*Sinorama*
(3)	ORG	…a natural hothouse for banana growing… provides broad spaces to roam, and Pan Viet chairman C.F. Chang moved quickly to seize the investment opportunity. …為香蕉鋪就了成長的溫床…則提供了開闊的馳騁空間，越盛公司董事長張哲發投資的快馬，便趁勢奔騰其中。	(Pan Viet, 越盛公司 "Yueh Sheng Company")	為 "Wei"	*Sinorama*
(4)	ORG	The vetting process was not easy because of the large number and diverse nature of the applications. The main criterion adopted by the Vetting Committee was whether the application would contribute to the further development of the service sectors. 由於申請數量眾多，且來自不同的服務行業，評選工作並不容易，而評審委員會所採納的主要評選標準，是計劃是否有助各個服務行業進一步發展。	(Vetting Committee, 評審委員會)	委員會	*HKNPT*

Table 6.9 Examples of alignment errors made using the proposed approach. (Cont.)

Example	Type	Bilingual Sentences	Correct NE Pairs	Miss-aligned Chinese NEs	Corpus
(5)	PER	Ever since Hong Kong's talented director Tsui Hark brought in kungfu genius Jet Li from mainland China, and using dazzling cinematic techniques, …"Proudly facing the myriad pounding waves, our hot blood like the red sunlight..." 自從香港鬼才導演徐克找來大陸武術奇才李連杰，以眩目的電影手法…「豪氣面對萬重浪，熱血像那紅日光…」	(Tsui Hark, 徐克 "Hsu Ko")	豪氣 "Hao Chi"	*Sinorama*
(6)	LOC	To survey the distribution and numbers of green turtles in the area, in July of last year Cheng Yi-chun went to the Pratas Islands,… 為了調查該地區綠蠵龜的分佈及族群狀況，去年七月，程一駿曾前往東沙群島做綠蠵龜分佈調查，…	(Pratas Islands, 東沙群島 "Tung Sha Islands")	群島	*Sinorama*
(7)	LOC	… Monica C. Kuo, of the Department of Landscape Architecture of Chinese Culture University, suggests that cycling around the narrow lanes of historic places like Tainan, Lukang and Sanhsia … …中國文化大學景觀學系系主任郭瓊瑩也建議，騎著單車探訪府城、鹿港、三峽等小鎮，…	(Tainan, 府城 "Fu Cheng")	探訪 "Tan Fan"	*Sinorama*
(8)	PER	Following is a question by the Hon Howard Young and a reply by the Acting Secretary for Security, Mrs Carrie Yau, in the Provisional Legislative Council today (Wednesday) : 以下為今日（星期三）在臨時立法會會議上楊孝華議員的提問和署理保安局局長尤曾家麗的答覆：	(Carrie Yau, 尤曾家麗 "Yau Jang Ga Lai")	尤曾家	*HKNPT*

Table 6.10 shows the comparison between *Sinorama* and *HKNPT* in terms of the performance achieved using language-specific knowledge sources, including SPTM, TM, AE, AH, and CPNR. Detailed statistics on the numbers of translations and transliterations occurring in two test corpora are given in Table 6.11.

Table 6.10 Performance for each language-specific knowledge source in the two corpora.

Corpus	SPTM	TM	AE	AH	CPNR
Sinorama	92.74%	93.37%	88.46%	87.01%	88.24%
HKNPT	83.41%	84.01%	90.41%	87.50%	92.45%

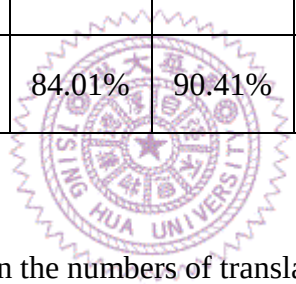

Table 6.11 Detailed statistics on the numbers of translations and transliterations in the two corpora.

Corpus		PER	LOC	ORG
Sinorama	Numbers of Translations	37	469	512
	Numbers of Transliterations	345	118	78
HKNPT	Numbers of Translations	5	263	377
	Numbers of Transliterations	163	145	15

Notably, on average, the performance of the proposed approach in the *Sinorama* test was better than that in the *HKNPT* test. One major reason is that the employed translation lexicon was trained with a general dictionary, a list of NE pairs, and the *Sinorama* corpus. Thus, the employed lexicon, especially for the proper names of LOC-type NEs and ORG-type NEs, does not cover many word translations in *HKNPT*. For the LOC-type NEs, some examples that were not aligned correctly in *HKNPT* are given as follows: (Lantau, 大嶼山), (Castle Peak, 青山), (Stanley, 赤柱), (Repulse Bay, 淺水灣), (Queensway, 金鐘道), and (Trio, 三星灣). As for the ORG-type NEs, some erroneous examples are given as follows: (Marine and Land Enforcement Command, 海域巡邏組), (Treasury, 庫務局), (Geotechnical Engineering Office, 土力工程處), (Department of Justice, 律政司), (Arch SD, 建築署), and (Correctional Services Department, 懲教署). However, the above problems can be alleviated by adding a small gazetteer of well-known names or by automatically learning word translations from domain-specific corpora. The other reason for the poorer performance in the *HKNPT* test was transliteration failures. Some transliteration error cases, which show the difficulties that were encountered when NEs were transliterated in the *HKNPT* test, are listed as follows: (Mauritian, 毛里求斯 "Mao Li Chiu Ssu"), (Robert Ribeiro, 李義 "Li I"), (Tony Blair, 貝理雅 "Pei Li Ya"), (Burrell, 貝偉 "Pei Wei"), (Derek Roebuck, 羅德立 "Lo Te Li"), and (Felice Lieh Mak, 麥列菲菲

"Mai Lieh Fei Fei"). All the above cases cannot be solved via transliteration or CPNR. We also noticed that the performance of the IBM Model 4 in the *HKNPT* test was much better than that in the *Sinorama* test. One possible reason may be the number of aligned sentences in the corpora used in our experiment, since the IBM Model 4 was designed for word alignment based on sufficient co-occurrence statistics of large parallel corpora. In our experiment, the *Sinorama* corpus consisted of 50,000 aligned sentences, which is much less than that of the *HKNPT* corpus, in which 600,000 aligned sentences were used.

By incorporating the proposed baseline method with extra knowledge functions, we achieved significant improvement in NE alignment in our experiments on different test data. We believe that the proposed framework, achieved by integrating various language functions, can be further improved through the use of more refined models and language-specified functions. The proposed SPTM can be refined by introducing additional linguistic information. More specifically, we can train more fine parameters of SPTM models for each individual NE type instead of tying all types into a single SPTM. Moreover, if we have sufficient training data, we can train SPTM constrained by NE keywords. For example, the translation "處" for "department" almost always appears in the last position of Chinese ORG-type NEs. As for CPNR, we can enhance Chinese surname detection performance by considering composite surnames that are

composed of two one-character surnames, such as “尤曾.” Furthermore, the proposed CPNR can be refined by introducing gender information of PER-typed NEs. For instance, given the source NE “Amy Hung” with two candidate names “洪瑩芬 (Hung, Ying-Fen)” and “洪俊傑 (Hung, Chun-Chieh),” it is obvious that “洪瑩芬” should be selected since it is a female name. Contextual information, such as personal titles and speech-act verbs, can also be used to improve the NE alignment performance by analyzing the contextual structures of aligned sentences.

One major limitation of training statistical translation models using parallel corpora is the lack of large parallel corpora. A potentially effective way to alleviate this problem is to develop algorithms for automatically extracting parallel corpora from the Web (Nie et al., 1999; Kilgarriff and Grefenstette, 2003; Resnik and Smith, 2003; Yang and Li, 2003). We believe that, now that the Web has become a huge repository of resources, many bilingual corpora could be automatically mined from the Web. Automatic mining of parallel corpora from the Web will benefit the approach to efficiently developing bilingual text processing tools using parallel corpora.

As the need for the acquisition of bilingual NE pairs is becoming increasingly essential, we have proposed a unified framework to achieve the goal of automatic bilingual NE alignment in parallel corpora. Several experiments were conducted to

show the better performance of the proposed methods when confronted with various types of NEs. To deal with bilingual NE transformation based on translation/transliteration, a baseline method based on SPTM and TM was proposed. To further improve the performance, several knowledge sources were explored in this study. We have successfully expanded an English acronym via AE and aligned the expansion with its translation, which is a very effective approach to aligning ORG-type NEs. We used AH to measure the similarity between a Chinese NE candidate and its abbreviation. This feature is highly effective for ORG-type NEs, since Chinese ORG-type NEs frequently appear in abbreviated forms. CPNR was applied to enhance the association between pairs of PER-type NEs, especially for foreign names or abbreviated names and their corresponding Chinese names. All of the proposed strategies for AH, CPNR, and AE appeared to substantially improve the performance in the experiments. As a consequence, by combining all the proposed knowledge sources, the average rate could be improved from 84.99% to 91.13% for *Sinorama*, and from 62.82% to 80.18% for *HKNPT*.