Chapter 5: Integrated Approach to NE Alignment

This chapter introduces a unified framework that aims at aligning bilingual NEs in parallel corpora. The overall process of the framework combining the proposed models is also described with illustrated examples.

5.1 Framework of the Proposed Approach

This section describes a framework to achieve the goal of aligning source NEs in the source texts with corresponding NEs in the target texts. A two-stage approach is undertaken to achieve the above goal. The first stage is for data preprocessing, which aims to label English NEs in the English side. A sentence alignment procedure is applied first to align parallel texts at the sentence level. Then, an English NE identifier is introduced to label NEs. In the second stage for the main process, for each English NE in an English sentence, we employ an NE alignment procedure to identify the corresponding Chinese NEs in the aligned sentence.

The alignment procedure combines SPTM, TM and other knowledge functions. These functions include:

(1) AH: abbreviation handling in Chinese NEs.

(2) CPNR: Chinese person name recognition for mapping an English name to Chinese.

(3) AE: acronym expansion for expanding English acronyms to their original forms for better alignment.

Figure 5.1 shows the proposed approach that is designed by integrating the above functions to align bilingual NEs in parallel corpora.

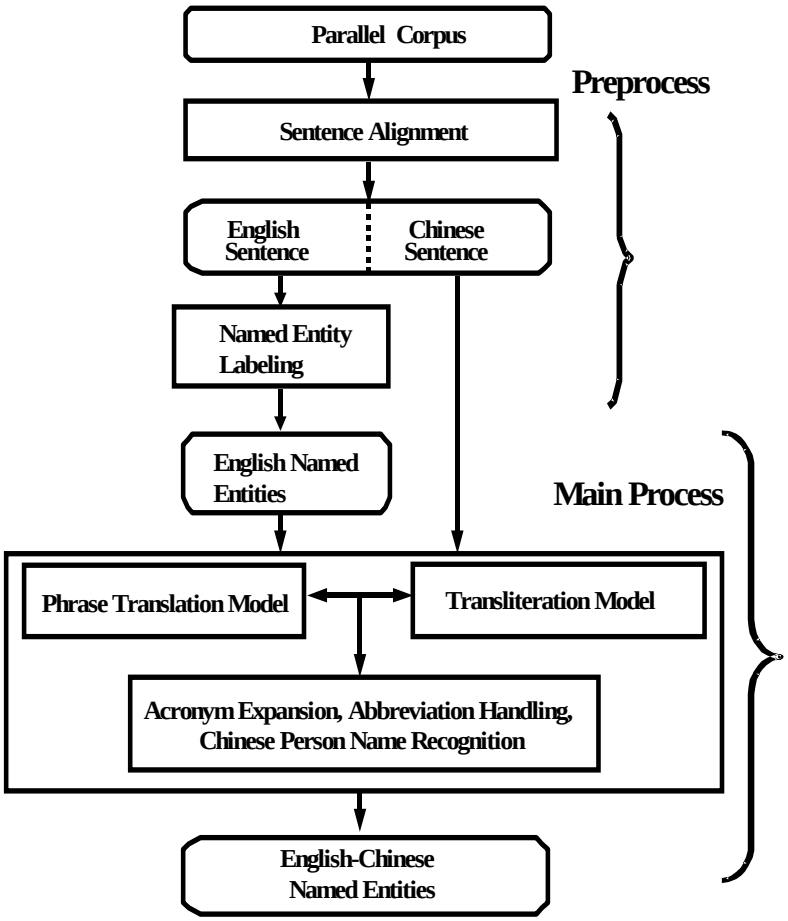

Figure 5.1 The framework for aligning bilingual NEs in parallel corpora.

5.2 Process of Aligning Bilingual NEs in Parallel Corpora

This section describes the detail step-by-step process of aligning bilingual NEs in parallel corpora. Figure 5.2 summarizes the framework of the overall process implemented in a top-down way by integrating SPTM, TM, AH, CPNR, and AE. To illustrate the flexibility of the process, some examples of bilingual NE pairs extracted from aligned sentences in *Sinorama* are shown in Table 5.1. For the sake of clarity, NEs are underlined in Table 5.1.

I. Data preprocess:

(I.1). Perform sentence alignment.

(I.2). Label English named entities.

II. Main process:

For each English named entity e in an English sentence S_e, align the corresponding Chinese named entity f in the aligned Chinese sentence S_f as follows:

(II.1). Generate all possible Chinese NE candidates by means of the proposed SPTM model using general-purpose and domain-specific lexicons. More specifically, for each labeled e, apply SPTM and AH to find translation equivalents $\{f_1\}$ in S_f.

(II.2). For each content word w in e that does not have a corresponding translation in S_f, apply the proposed TM and CPNR modules to extract the corresponding translation equivalents $\{f_2\}$ in S_f.

(II.3). Merge $\{f_1\}$ with $\{f_2\}$ to form a set of potential translation equivalents $\{f\}$.

(II.4). Rank $\{f\}$ based on the scores. Choose the candidate f with the maximum score as the answer to form the pair (e, f) as the result.

Figure 5.2 The process of aligning bilingual NEs in parallel corpora.

Table 5.1 Examples of NE pairs in aligned sentences.

Example	Bilingual Sentences	Bilingual NE Pairs
(1)	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。	(CLA, 勞委會)
(2)	Christopher Roberts comes from California, and has just turned 40. He has a double doctorate in composition and double bass performance from the Juilliard School, New York. Fifteen years ago he received funding to carry out research in the Star Mountains of central New Guinea … 羅白華來自美國加州，今年剛過四十，擁有美國紐約茱麗亞音樂學院作曲及低音大提琴雙重博士，十五年前得到一筆基金贊助，到新幾內亞中部的星辰山脈，…	(Christopher Roberts, 羅白華), (California, 加州), (Juilliard School, 茱麗亞音樂學院), (New York, 紐約) (Star Mountains, 星辰山脈), (New Guinea, 新幾內亞)
(3)	“Visas are hard to come by,” says Amy Hung, who is currently studying at the Lincoln College Center. Quite a few people who come to Vancouver and find out things aren’t quite right try to switch to the U.S., but they usually wind up coming right back. 「簽證很難拿」，目前就讀「林肯大學中心」的洪瑩芬表示，有不少人來到溫哥華後發現情況不對，想轉往美國，結果都被打了回票，	(Amy Hung, 洪瑩芬), (Lincoln College Center, 林肯大學中心), (Vancouver, 溫哥華), (U.S., 美國)

For the purpose of extracting bilingual NE pairs from parallel corpora, in Step (I.1) in Figure 5.2, a sentence alignment procedure based on length and lexical information (Chuang et al., 2002) is applied to align parallel texts at the sentence level. In Step (I.2), an HMM based English NE identifier, based on case information, POS tags, the words themselves, and previously predicted NE tags, is applied to approximately label NE candidates for each sentence in the English text. Next, the labels are manually corrected. A general overview of NE recognition systems adopted in the NE recognition shared task of CoNLL-2003 can be found in (Sang and Meulder, 2003). Many studies have focused on identifying monolingual NEs, especially in English. In this study, on the other hand, our focus is the alignment of bilingual NEs. We shall further clarify the main process by means of illustrative examples in the following.

In Step (II.1), an acronym can be expanded to its full name by looking up an acronym-expansion table, as shown in Table 3.3. For instance, in example (1) in Table 5.1, "CLA" is expanded into "Council of Labor Affairs." In this step, a set of potential Chinese NE candidates $\{f_l\}$ for each e is generated via the proposed phrase translation model. For instance, in example (1) in Table 5.2, possible translations of "Council," "Labor," and "Affairs" are {協會, 委員會, …}, {勞工, 人工, …}, and {事務, 行政, …}, respectively. Therefore, after applying SPTM and AH, we find that

the set $\{f_1\}$ of "CLA" is {勞委會，勞委，勞會，委會，…}, as shown in Table 5.2. In this case, $\{f_2\}$ is empty, since neither TM nor CPNR is activated in Step (II.2). Thus, $\{f\}$ is the same as $\{f_1\}$. Finally, in Step (II.4), we can extract the NE pair (CLA, 勞委會) by choosing the top-1 ranking of the candidates in $\{f\}$. Note that "勞委會" is an abbreviation of "勞工事務委員會," which is a translation equivalent of "Council of Labor Affairs."

Table 5.2 Sets of Chinese NE candidates.

English NE	Chinese translation/Transliteration	Chinese NE Candidate
Council of Labor Affairs (CLA)	協會勞工事務，委員會勞工事務， 協會人工事務，委員會人工事務， 協會勞工行政，委員會勞工行政， 勞工事務協會，勞工事務委員會， 人工事務協會，人工事務委員會，…	勞委會，勞委， 勞會，委會，…
Juilliard School	茱麗亞學校， 茱麗亞學院， 學校茱麗亞， 學院茱麗亞	茱麗亞音樂學院， 茱麗亞音樂學， 茱麗亞，學院，…

Since NEs are sometimes transformed in a flexible and inconsistent manner, the merging process of $\{f_1\}$ with $\{f_2\}$ in Step (II.3) should also be performed in a flexible manner. More specifically, let us consider the NE "Juilliard School" in example (2) in Table 5.1 "Juilliard" can be transliterated as "茱麗亞 (Ju Li Ya)" by TM and "School" can be translated as "學院" (or "學校") by dictionary lookup, respectively, after Steps (II.1) and (II.2). However, the correct NE pair is (Juilliard School, 茱麗亞

音樂學院), which cannot be identified directly since "音樂" ("music" in English) never appears in the original English NE. In contrast, in Step (II.3), we have taken care of this issue by using tolerant position offsets to construct a single named entity "茱麗亞音樂學院," instead of two separated partial named entities "茱麗亞" and "學院," as shown in Figure 5.3.

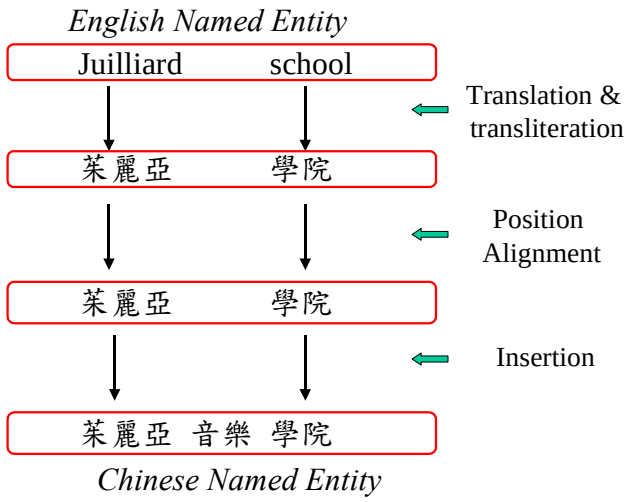

Figure 5.3 Alignment of the NE pair (Juilliard School, 茱麗亞音樂學院).

In the next example, we will demonstrate how CPNR helps to improve the performance of PER-type alignment. Person names are almost always transliterated. However, in example (3) in Table 5.1, "瑩芬 (Ying Fen)" is neither transliterated nor translated from "Amy" in the PER-type NE pair (Amy Hung, 洪瑩芬 "Hung Ying Fen"). In this example, "Hung" is a Chinese last name and can be translated as "洪," which forms $\{f_1\}$. But "Amy" is a foreign name that does not have a corresponding transliteration in the aligned Chinese sentence. The goal of the next step is to find the

association between a foreign name and a Chinese name. Therefore, in Step (II.2), CPNR is activated by “Amy,” and the Chinese given name “瑩芬” is then detected by CPNR, forming $\{f_2\}$. Thus, the pair (Amy Hung, 洪瑩芬) is successfully detected by merging $\{f_1\}$ with $\{f_2\}$.

To extract bilingual NE pairs, the symmetric approach in previous studies requires identifying NEs in both languages. However, developing two NE identifiers instead of one requires a lot of effort. Moreover, two NE identifiers do not always extract NE pairs consistently, especially when one of the NE identifiers is not as capable as the other one. For example, the Chinese NE identifier is currently not well developed due to the fact that there are no spaces between Chinese characters, leading to ambiguity in word segmentation. The proposed framework proposed here, on the other hand, only needs a reliable English NE identifier together with the proposed models associated with multiple knowledge sources to extract NE pairs. Experimental results show that our approach can achieve excellent performance for bilingual NE alignment. More details about the experiments will be reported in Chapter 6.