A Joint Model to Identify and Align Bilingual Named Entities

Yufeng Chen*
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Chengqing Zong**
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Keh-Yih Su†
Behavior Design Corporation

In this article, an integrated model is derived that jointly identifies and aligns bilingual named entities (NEs) between Chinese and English. The model is motivated by the following observations: (1) whether an NE is translated semantically or phonetically depends greatly on its entity type, (2) entities within an aligned pair should share the same type, and (3) the initially detected NEs can act as anchors and provide further information while selecting NE candidates. Based on these observations, this article proposes a translation mode ratio feature (defined as the proportion of NE internal tokens that are semantically translated), enforces an entity type consistency constraint, and utilizes additional new NE likelihoods (based on the initially detected NE anchors).

Experiments show that this novel method significantly outperforms the baseline. The type-insensitive F-score of identified NE pairs increases from 78.4% to 88.0% (12.2% relative improvement) in our Chinese–English NE alignment task, and the type-sensitive F-score increases from 68.4% to 83.0% (21.3% relative improvement). Furthermore, the proposed model demonstrates its robustness when it is tested across different domains. Finally, when semi-supervised learning is conducted to train the adopted English NE recognition model, the proposed model also significantly boosts the English NE recognition type-sensitive F-score.

* No. 95, Zhongguancun East Road, Haidian District, Beijing 100190, China. E-mail: chenyf@nlpr.ia.ac.cn.
** No. 95, Zhongguancun East Road, Haidian District, Beijing 100190, China. E-mail: cqzong@nlpr.ia.ac.cn.
† Hsinchu, Taiwan. E-mail: bdc.kysu@gmail.com.

Submission received: 9 October 2010; revised submission received: 15 February 2012; accepted for publication: 27 March 2012.

doi:10.1162/COLI_a_00122

1. Introduction

Named entities (NEs), especially person names (PER), location names (LOC), and organization names (ORG), deliver essential context and meaning in human languages. Therefore, NE translation plays a critical role in trans-lingual language processing tasks, such as machine translation (MT) and cross-lingual information retrieval. To learn NE translation knowledge, bilingual NE alignment (which links source NEs and target NEs to generate desired NE pairs) is the first step in producing the NE translation table (which can then be used to train the NE translation model). Furthermore, with additional alignment constraints from the other language, the alignment module can also refine those initially recognized NEs, and thus can be adopted to conduct semi-supervised learning to learn monolingual NE recognition models from a large untagged bilingual corpus.

Because NE alignment can only be conducted after its associated NEs have been identified, the NE recognition errors propagate into the alignment stage. The type-insensitive inclusion rate[1] of the initial recognition stage thus significantly limits the final alignment performance. One way to alleviate this error propagation problem is to jointly perform NE recognition and alignment. Such a combined approach is usually infeasible, however, due to the high computational cost of evaluating alignment scores for a large number[2] of NE pair candidates.

In order to make the problem computationally tractable, a sequential approach is usually used to first identify NEs and then align them. Two such kinds of sequential strategies that alleviate the error propagation problem have been proposed. The first strategy, named **asymmetry alignment** (Al-Onaizan and Knight 2002; Moore 2003; Feng, Lv, and Zhou 2004; Lee, Chang, and Jang 2006), identifies NEs only on the source side and then finds their corresponding NEs on the target side. Although this approach avoids the NE recognition errors resulting from the target side, which would otherwise be brought into the alignment process, the NE recognition errors from the source side continue to affect alignment.

To further reduce the errors from the source side, the second strategy, denoted **symmetry alignment** (Huang, Vogel, and Waibel 2003), expands the NE candidate sets in both languages before conducting the alignment. This is achieved by using the original results as anchors, and enlarging or shrinking the boundaries of the anchors to generate new candidates. This strategy fails to work if the NE anchor has already been missed in the initial NE recognition stage, however. In our data set (1,000 Chinese–English sentence pairs randomly selected from the Chinese News Translation Text corpus [LDC2005T06]), this strategy significantly improves the type-insensitive NE pair inclusion rate from 83.9% to 96.1%;[3] in the meantime, the type-insensitive Chinese NE (CNE) recognition inclusion rate rises from 88.7% to 95.9%, and that of English NE (ENE) from 92.8% to 97.2%. This strategy is thus adopted in this article.

Although the symmetric expansion strategy has substantially alleviated the problem of error propagation, the final alignment accuracy, in terms of type-sensitive F-score

1 This is the percentage of desired NE pairs that are included within the given candidate set, and is the upper bound for NE alignment performance (type-insensitive means disregarding NE types).

2 This number will dramatically increase if the combined approach is adopted, as every possible string will become a NE candidate.

3 This figure is based on the expanded candidates that are constructed without any range limitation. In this case, Inclusion rate = 1 – [Missing Rate of Initial NEs]. In addition, because not every Chinese NE is linked in the given sentence pair, the Inclusion rate of Chinese NEs is even lower than that of NE pairs (95.9% vs. 96.1%).

(achieved by the approach proposed by Huang, Vogel, and Waibel [2003]) continues to be as low as 68.4% (see in Table 3 in Section 4.3). After having examined the data, we found the following: (1) How a given NE is translated, either semantically (called **translation**) or phonetically (called **transliteration**), depends greatly on its associated entity type.[4] The **translation mode ratio**, which is the percentage of NE internal tokens that are translated semantically, thus can help to identify the NE type. (2) Entities within an aligned pair should share the same type, and this restriction should be integrated into the NE alignment model as a constraint. (3) In prior work, the initially identified monolingual NEs were used only to construct the candidate set without playing any role in final NE identification. Indeed, these monolingual NEs do carry other useful information and can act as anchors to give **NE likelihoods**, which can provide additional scope preference information to those regenerated candidates.

Based on these observations, we propose a novel joint model that adopts the translation mode ratio, enforces the entity type consistency constraint, and also utilizes the NE likelihoods. This proposed approach jointly identifies and aligns bilingual NEs under an integrated framework, which consists of three stages: Initial NE Recognition, NE-Candidate Set Expansion, and NE Re-identification & Alignment. The Initial NE Recognition stage identifies the initial NEs and their associated NE types in both the source and target. In the next stage, NE Candidate Set Expansion regenerates the candidate sets in both languages in order to remedy the initial NE recognition errors. In the final stage, NE Re-identification & Alignment *jointly* recognizes and aligns bilingual NEs via the proposed joint model. The experimental results validate our proposed three-step method.

The integrated model that jointly identifies and aligns bilingual named entities between Chinese and English was originally introduced in Chen, Zong, and Su (2010). In this article, the problem has been re-formulated and derived. The new derivation starts from *two given NE sequences*, whereas the original derivation only begins with *one given NE pair*. We also give more details of the problem study, model analysis, and experiments. Moreover, we report additional experiments, which include those that study the effect of adopting different initial NE recognizers and the effectiveness of the proposed model across different domains. Finally, a complete error analysis is given in the current version.

The remainder of this article is organized as follows: Section 2 motivates the proposed method. Afterwards, the proposed model is formally introduced in Section 3. Section 4 describes experiments conducted on various configurations of the method. The associated error analysis and discussion of results are presented in Section 5. Section 6 gives applications of the proposed model. We review related work in Section 7. Finally, conclusions are drawn in Section 8.

2. Motivation

By examining the NEs initially recognized in aligned sentence pairs, we have the following two observations: (1) Alignment can help fix those NEs that are initially incorrectly recognized when they are not the correct counterparts of each other. Therefore,

4 The proportions of semantic translation (which denote the ratios of semantically translated words among all the associated words within NEs) for PER, LOC, and ORG are approximately 0%, 28.6%, and 74.8%, respectively, in the Chinese–English name entity list (2005T34) released by the Linguistic Data Consortium (LDC). Because titles, such as "sir" and "chairman," are not considered part of person names in this corpus, all PERs are transliterated.

alignment and recognition should be jointly optimized. (2) Alignment cannot help in determining the appropriate scope when each word within an NE (or within its larger context window covering the NE) is correctly matched to its counterpart. Therefore, the information of those initial NEs should be utilized to decide the appropriate NE scope. The following two sections further elaborate on these two observations.

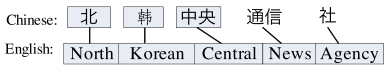
2.1 Alignment Helps NE Recognition

In NE recognition, both boundary identification and type classification are required. The complexity of these tasks varies with different languages, however. For example, Chinese NE boundaries are not obvious because adjacent words are not separated by spaces. In contrast, English NE boundaries are easier to identify with explicit words and capitalization clues. On the other hand, classification of English NE type is considered more challenging (Ji and Grishman 2006).

Because alignment would force the NEs in the linked NE pair to share the same semantic meaning, the NE that is more reliably identified in one language can be used to identify its less reliable counterpart in the other language. This benefit, which is observed in both NE boundary identification and type classification, indicates that alignment can be used to locate those NEs that are initially incorrectly recognized. For example, once the correct boundaries are drawn in one language, word equivalences inside an aligned NE pair can help identify NE boundaries in the language that does not have explicit clues (e.g., Chinese). As shown in Example (1), even though the desired Chinese NE "北韩中央通信社" is only partially recognized as "北韩中央" in the initial recognition stage, it can be recovered if the English counterpart *North Korean [no's] Central News Agency* is given. The reason for this is that *News Agency* is better aligned to "通信社", rather than be deleted, which would occur if "北韩中央" is chosen as the corresponding Chinese NE.

Example (1). Alignment helps recognize NE boundaries

(I) The initial NE detection in a Chinese sentence:
官方的 <ORG>北韩中央</ORG> 通信社引述海军...
(II) The initial NE detection of its English counterpart:
Official <ORG>North Korean Central News Agency </ORG> quoted the navy's statement…
(III) The word alignment between two NEs:



(VI) The re-identified Chinese NE boundary after alignment:
官方的 <ORG>北韩中央通信社</ORG> 引述海军声明...

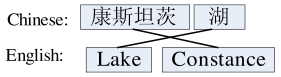
On the other hand, type consistency constraints can help correct the NE type that is less reliably identified. Moreover, in identifying the NE type, it helps if we know whether a word is translated or transliterated. As illustrated in Example (2), the word *lake* in the English NE is linked to the Chinese character "湖", and this mapping is found to be a translation, not a transliteration. Because translation rarely occurs for personal names (Chen, Yang, and Lin 2003), the desired NE type "LOC" should be shared between the English NE *Lake Constance* and its corresponding Chinese NE "康斯坦茨湖." As a result, the original incorrect type "PER" of the given English NE is fixed; it thus corroborates the need for using the translation mode ratio and NE type consistency constraint.

2.2 Initial NEs Carry NE Scope Information

In Huang, Vogel, and Waibel (2003), initial NE information was discarded once the new candidate set was generated. Alignment scores alone, however, are incapable of selecting the appropriate NE scope in some cases. For instance, when each word in an NE is correctly matched to its counterpart, the alignment score might still prefer an incorrect NE pair with smaller scope, even after the normalization of alignment terms has been considered (explained in Section 4.3). On the other hand, when words surrounding the desired NE are also correctly matched to their counterpart, the incorrect NE pair with larger scope might be chosen. As illustrated in Example (3), the desired NE pair {⟨德国⟩::[Germany]} is initially correctly recognized, although the wrong NE pair {⟨德国经济⟩::[Germany economy]} is finally selected from the regenerated candidate set

Example (2). Alignment helps recognize NE types

(I) The initial NE detection result in a Chinese sentence:
在 <LOC>康斯坦茨湖</LOC> 工作的 一艘渡船船长…
(II) The initial NE detection of its English counterpart:
The captain of a ferry boat who works on <PER>Lake Constance </PER>…
(III) The word alignment between two NEs:

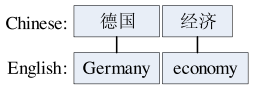


(VI) The re-identified English NE type after alignment:
The captain of a ferry boat who works on <LOC>Lake Constance</LOC>…

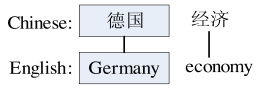
if only alignment scores are used in the final selection process. This is because the extra words "经济" and *economy* are a perfect translation of each other, thus resulting in the incorrect pair {⟨德国经济⟩::[Germany economy]}, which receives a higher alignment score than does the correct NE pair {⟨德国⟩::[Germany]}.

Example (3). Alignment score alone fails to select correct NE scope

(I) The initial NE detection result in a Chinese sentence:
第二季<LOC>德国</LOC> 经济估计成长在…
(II) The initial NE detection of its English counterpart:
<LOC>Germany </LOC> economy in 2nd quarter this year expected to grow…
(III) The NE alignment result without referring initial NEs:



(VI) The final NE alignment result with referring initial NEs:



It must also be noted that an NE alignment model usually ignores a fair amount of potentially useful information that is commonly used in monolingual NE recognition models. For example, the bigrams of text surrounding NEs are frequently adopted in a monolingual NE recognition model, but not in a bilingual alignment model. Other

examples include case information, part-of-speech (POS) triggers, gazetteer features, and external macro context features mentioned in Zhou and Su (2006). The alignment model fails to consider the monolingual context surrounding NEs while determining NE scope, resulting in the error mentioned in Example 3. By ignoring the initial NEs after their corresponding candidate sets have been generated, we lose the information provided by these initial NEs that is otherwise available to the alignment model. Therefore, though the initially detected NEs might be unreliable by themselves, they should act as anchors to provide scope preference information, even after the expanded candidate set has been generated from these NEs.

3. The Proposed Joint Model

Given a Chinese–English sentence pair (Sc, Se), with its *initial* (denoted by lower case letter b, for "beginning") Chinese NEs $[Cb_i, Tc_i]_{i=1}^{Nc}, Nc \geq 1$ and English NEs $[Eb_j, Te_j]_{j=1}^{Ne}, Ne \geq 1$, where Nc and Ne are the numbers of initially recognized NEs of Chinese and English, respectively; Tc_i and Te_j are the original NE types assigned to Cb_i and Eb_j , respectively. (For the reader's convenience, all adopted notations are listed in Table 1 for quick reference.) We first regenerate two NE candidate sets (by enlarging and shrinking the boundaries of those initial NEs) to include, we hope, the correct corresponding candidates that failed to be recognized in the first stage. Let C_1^{Kc} and E_1^{Ke} denote the two sets that include those regenerated candidates for Chinese and English NEs, respectively (Kc and Ke are their set-sizes), and $K = \min(Nc, Ne)$. Then, the total K pairs of the final Chinese and English NEs will be extracted from the Cartesian product of C_1^{Kc} and E_1^{Ke}. Here, only NE pairs in one-to-one mappings will be extracted, as most applications are only interested in this kind of correspondence. Therefore, we will let C_1^K and E_1^K denote the two extracted candidate sets to be linked, and these sets will consist of non-overlapping NE candidates from C_1^{Kc} and E_1^{Ke}, respectively (for conciseness, we will not explicitly distinguish between the indices C_1^K and C_1^{Kc}, or between E_1^K and E_1^{Ke}).

Let $\{C_{a(k)}, E_k\}$ denote a specific NE-linking-pair (where $a(k)$ and k are the associated indices of those regenerated Chinese and English NEs within C_1^K and E_1^K, respectively); the subscript $a(k)$ denotes that $C_{a(k)}$ is aligned to E_k. Let T_k be the NE type to be re-assigned and shared by $C_{a(k)}$ and E_k (as they should denote the same entity). Assuming that only one-to-one mappings of NE pairs will be extracted, the problem of getting

Table 1
Adopted notations (keywords have been *italicized*).

Symbols		Symbols	
Chinese *Sentence*	Sc	English *Sentence*	Se
Initial/Beginning *Chinese* NE	Cb	Initial/Beginning *English* NE	Eb
Regenerated *Chinese* NE	C	Regenerated *English* NE	E
Initial Chinese NE *Type*	Tc	Initial English NE *Type*	Te
Re-assigned NE *Type*	T	NE Sequence *Alignment*	As
Internal Component *Alignment*	A	Internal Translation *Mode*	M
Chinese *Component*	cp	English *Word*	ew
Translation Mode *Ratio*	δ	Chinese *Character*	cc
Left *Distance*	d_L	Right *Distance*	d_R
Initial/Beginning NE *Length*	Lb	Regenerated NE *Length*	L

the final desired aligned NE pairs $\left\{C^*_{a(k)}, E^*_k, T^*_k\right\}_{k=1}^{K}$ is that of finding the most likely allowable combination of NE pairs (and their re-assigned NE types), given all the initial NEs (i.e., $[Cb_i, Tc_i]_{i=1}^{Nc}$ and $[Eb_j, Te_j]_{j=1}^{Ne}$) and the sentences that include them (i.e., *Sc* and *Se*). This can be formulated as follows:

$$\begin{aligned}&\left\{C^*_{a(k)}, E^*_k, T^*_k\right\}_{k=1}^{K}\\ &= \underset{\left\{C_{a(k)},E_k\right\}_{k=1}^{K}}{\arg\max}\left[\max_{T_1^K} P\left(\left\{C_{a(k)}, E_k, T_k\right\}_{k=1}^{K} |[Cb_i, Tc_i]_{i=1}^{Nc}, Sc, [Eb_j, Te_j]_{j=1}^{Ne}, Se\right)\right]\end{aligned} \tag{1}$$

For any given NE–pair sequence $\left\{C_{a(k)}, E_k\right\}_{k=1}^{K}$, the internal max operator will first operate over each re-assigned type-sequences T_1^K (where T belongs to PER, LOC, ORG for each given NE pair included in the NE–pair sequence). The outer argmax operator will cross every admissible NE–pair sequence.

This formulation implies that recognition and alignment are executed jointly with respect to C_1^K and E_1^K, without making any independence assumptions among those NE pairs included in the associated NE–pair sequence. This equation is thus computationally infeasible due to a large search space. Therefore, it is further simplified and derived as follows by first explicitly denoting the link between $C_{a(k)}$ and E_k as $A_{S,k} = \left\{C_{a(k)}, E_k\right\}$. Let $\{A_{S,k}\}_{k=1}^{K}$ (abbreviated as $A_{S,1}^K$) denote one possible alignment between C_1^K and E_1^K. We will then have $K!$ different possible alignments between them (i.e., total factorial of K different $A_{S,1}^K$). Finally, let $[Cb_{a(k)}, Tc_{a(k)}]_{k=1}^{K}$ and $[Eb_k,Te_k]_{k=1}^{K}$ denote those initially recognized corresponding NEs that generate C_1^K and E_1^K (i.e., their associated anchors), respectively. Then we can replace $\left\{C_{a(k)},E_k,\ T_k\right\}_{k=1}^{K}$ in Equation (1) by $[A_{S,k}, C_{a(k)}, E_k, T_k]_{k=1}^{K}$, and derive the original probability $P\left(\left\{C_{a(k)}, E_k, T_k\right\}_{k=1}^{K} |[Cb_i, Tc_i]_{i=1}^{Nc}, Sc, [Eb_j, Te_j]_{j=1}^{Ne}, Se\right)$ as follows.

$$\begin{aligned}&P\left(\left\{C_{a(k)}, E_k, T_k\right\}_{k=1}^{K} |[Cb_i, Tc_i]_{i=1}^{Nc}, Sc, [Eb_j, Te_j]_{j=1}^{Ne}, Se\right)\\ &\approx P\left([A_{S,k}, C_{a(k)}, E_k, T_k]_{k=1}^{K}|[Cb_{a(k)}, Tc_{a(k)}]_{k=1}^{K}, Sc, [Eb_k, Te_k]_{k=1}^{K}, Se\right)\\ &\approx \prod_{k=1}^{K}\left[P\left([A_{S,k}, C_{a(k)}, E_k, T_k]|[Cb_{a(k)}, Tc_{a(k)}], Sc, [Eb_k, Te_k], Se\right)\right]\end{aligned} \tag{2}$$

where $P\left([A_{S,k}, C_{a(k)}, E_k, T_k]|[Cb_{a(k)}, Tc_{a(k)}], Sc, [Eb_k, Te_k], Se\right)$ can be further decomposed as follows.

$$\begin{aligned}&P\left([A_{S,k}, C_{a(k)}, E_k, T_k]|[Cb_{a(k)}, Tc_{a(k)}], Sc, [Eb_k, Te_k], Se\right)\\ &\approx P(A_{S,k}|C_{a(k)}, E_k, T_k) \times P\left(T_k|Tc_{a(k)}, Te_k, Sc, Se\right)\\ &\times P(C_{a(k)}|Cb_{a(k)}, Tc_{a(k)}, T_k, Sc) \times P(E_k|Eb_k, Te_k, T_k, Se)\end{aligned} \tag{3}$$

In Equation (3), $P(A_{S,k}|C_{a(k)}, E_k, T_k)$ and $P\left(T_k|Tc_{a(k)}, Te_k, Sc, Se\right)$ are defined as NE Alignment Probability and NE Type Re-assignment Probability, respectively, for finding the final alignment $A_{S,1}^K$ among C_1^K and E_1^K that have been selected. Both $P(C_{a(k)}|Cb_{a(k)}, Tc_{a(k)}, T_k, Sc)$ and $P(E_k|Eb_k, Te_k, T_k, Se)$ are called NE likelihoods, and are used to assign preference to each selected C_1^K and E_1^K, based on the initial NEs (which act as anchors). For brevity, we drop the associated subscripts hereafter, if there is

no confusion. These probabilities will be further described in Sections 3.1 and 3.2. Finally, the joint identification and alignment framework that incorporates the initial NE recognition process, candidate set construction, and the associated search process are given in Section 3.3.

3.1 Bilingual Related Probabilities

The NE alignment probability represents the likelihood of a specific alignment $A_{S,k}$, given C and E and their associated T. Because Chinese word segmentation introduces errors, especially for transliterated words, the NE alignment probability $P(A_{S,k}|C_{a(k)}, E_k, T_k)$ in Equation (3) is derived from E (i.e., starting from the English part). In addition, because **internal component alignment** (denoted as A, to be defined later) within a given NE pair carries important information (as illustrated in Section 2), the internal component alignment will be introduced as follows.

$$\begin{aligned} P(A_{S,k}|C_{a(k)}, E_k, T_k) &= \sum\nolimits_A P(A|C, E, T) \\ &\approx \max_A P(A|C, E, T) \\ &= \max_A [\frac{1}{R} \times P(A|E, T)] \end{aligned} \tag{4}$$

where $R = \sum_A P(A|E, T)$ is a normalization value,[5] which will be ignored for simplicity, leaving only the probability $P(A|E, T)$ to be derived.

Let A be configured as $A \equiv \langle [cp_{a(n)}, ew_n, M_n]_{n=1}^{N}, \delta \rangle$, where $[cp_{a(n)}, ew_n, M_n]$ denotes a linked pair of a Chinese *component* $cp_{a(n)}$ (which might contain several Chinese characters) and an English word ew_n within C and E, respectively, with their **translation mode** M_n to be either *translation* (abbreviated as *TS*) or *transliteration* (abbreviated as *TL*). We assume that there are N component transformations in total, including N_{TS} translation transformations $[cp_{a(n)}, ew_n, TS]_{n=1}^{N_{TS}}$ and N_{TL} transliteration transformations $[cp_{a(n)}, ew_n, TL]_{n=1}^{N_{TL}}$, such that $N = N_{TS} + N_{TL}$. Moreover, because the statistical distribution of internal translation mode varies greatly across various NE types (as illustrated in footnote [4] of this article), the associated *translation mode ratio* $\delta = (N_{TS}/N)$ is an important feature and is included in the internal component alignment specified previously. For example, if the A between "康斯坦茨湖" and *Constance Lake* is [康斯坦茨, *Constance*, *TL*] and [湖, *Lake*, *TS*] ($N_{TS} = N_{TL} = 1$), then its associated translation mode ratio will be 0.5 (i.e., $\delta = 1/2$).

Therefore, the internal alignment probability $P(A|E, T)$ will be further deduced by introducing the translation mode M_n and the translation mode ratio δ as follows:

$$\begin{aligned} P(A|E, T) &\equiv P([cp_{a(n)}, ew_n, M_n]_{n=1}^{N}, \delta|E, T) \\ &\approx \prod_{n=1}^{N} [P(cp_{a(n)}|M_n, ew_n, T) \times P(M_n|ew_n, T)] \times P(\delta|T) \end{aligned} \tag{5}$$

5 The summation will be taken over various A that can generate C.

Combining Equations (4) and (5), the NE alignment probability $P(A_{S,k}|C_{a(k)}, E_k, T_k)$, which integrates internal component alignment information such as translation mode ratio and NE type constraint, is finally obtained as follows.

$$\begin{aligned} &P(A_{S,k}|C_{a(k)}, E_k, T_k) \\ &\approx \tfrac{1}{R} \times \max_A \left[\prod_{n=1}^{N_A} [P(cp_{A,a(n)}|M_{A,n}, ew_{A,n}, T) \times P(M_{A,n}|ew_{A,n}, T)] \times P(\delta_A|T) \right] \end{aligned} \tag{6}$$

where R is the normalization factor defined by Equation (4), and will be ignored in the final selection. In Equation (6), the mappings between internal elements is trained from the syllable/word alignment of NE pairs of different NE types. For transliteration, the model adopted in Huang, Vogel, and Waibel (2003), which first romanizes Chinese characters and then transliterates them into English characters, is used in estimating $P(cp_{a(n)}|TL, ew_n, T)$. For translation, conditional probability is directly used for $P(cp_{a(n)}|TS, ew_n, T)$.

On the other hand, the *NE type re-assignment probability* $P(T|Tc, Te, Sc, Se)$, proposed in Equation (3), is derived as follows.

$$P(T|Tc, Te, Sc, Se) \approx P(T|Tc, Te) \tag{7}$$

As Equation (7) shows, both the initially assigned Chinese NE type *Tc* and the initially assigned English NE type *Te* are adopted to jointly identify their shared NE type *T*.

3.2 Monolingual NE Likelihoods

The monolingual related probabilities in Equation (3) represent the likelihood that a regenerated NE candidate is the true NE, given its originally detected NE. For Chinese, we derive the likelihood as follows.

$$\begin{aligned} &P(C|Cb, Tc, T, Sc) \\ &\equiv P\left(d_L, d_R, String[C]|Lb, Tc, T\right) \\ &\approx P(d_L|Lb, Tc, T) \times P(d_R|Lb, Tc, T) \times \textstyle\prod_{l=1}^{L} P(cc_l|cc_{l-1}, T) \end{aligned} \tag{8}$$

Here *Lb* is the *length* (in characters) of the original recognized Chinese NE *Cb*. Let d_L and d_R denote the **left** and **right distance**, respectively (which are based on the numbers of Chinese characters), that *C* shrinks/enlarges from the left and the right boundaries of its anchor *Cb*. In Example (1) in Section 2.1, in the case where the given *Cb* and *C* are "北韩中央" and "韩中央通信社", respectively, then d_L and d_R are -1 and +3, respectively. Let *String*[*C*] denote the associated Chinese string of *C*, cc_l denote the *l-th* Chinese character within that string, and *L* denote the total number of Chinese characters within *C*. Then we will have a range of bigram probabilities for candidates with different lengths. Therefore, it is systematically biased[6] (in probability value) towards candidates with shorter lengths. On the English side, following Equation (8), $P(E|Eb, Te, T, Se)$ can be derived similarly; the unit is a word, however, rather than a character.

6 This bias is introduced by the conditional independence assumption made while decomposing $P\left(String[C]|T\right)$ into $\prod_{l=1}^{L} P(cc_l|cc_{l-1}, T)$.

In summary, with factors d_L and d_R, the proposed NE likelihood is able to assign scope preference to each regenerated NE based on its associated initial NE. The initial NE therefore still plays a role in the final selection process, even after its related candidate set has been generated, which is important when all words involved are correctly matched to their counterparts, as explained in Section 2.2. In contrast, Huang, Vogel, and Waibel (2003) adopt only type-dependent bigrams as the NE likelihood. The initial NE thus will not play any role in the final selection process after its related candidate set has been generated. The scope preference information carried by the initial NE is therefore not utilized in their model.

Having integrated all related probabilities (Equations (6), (7), and (8)) together, we now have the final desired model. For simplicity, all the probabilities involved are estimated by the Good-Turing smoothing technique (Chen and Goodman 1998) unless otherwise specified.

3.3 Framework for Jointly Identifying and Aligning Bilingual NEs

In jointly identifying and aligning bilingual NEs, a three-stage framework is adopted: (A) Initial NE Recognition, generating the initial NE anchors with off-the-shelf packages, (B) NE Candidate Set Expansion, expanding the associated NE Candidate set to remedy the errors made in the previous stage, and (C) NE Re-identification & Alignment, extracting the final NE pairs from the Cartesian product of source and target candidate sets (created in the second stage) via a search process. Figure 1 presents the detailed procedure of this framework.

1 For each given bilingual sentence pair:
2 (A) **Initial NE Recognition:** The initial Chinese NEs and English NEs are first identified by their corresponding NE recognition toolkits, respectively.
3 (B) **NE Candidate Set Expansion:** To rescue those NEs whose boundaries are incorrectly identified in the previous stage—for each initially detected NE, several NE candidates will be regenerated from the original NE by allowing its boundaries to be shrunk or enlarged within a pre-specified range.
4 (B.1) Create both *C* and *E* candidate sets, which are expanded from those initial NEs recognized in the previous stage.
5 (B.2) Construct a NE pair candidate set (named **NE-Pair-Candidate set**) by generating a Cartesian product of *C* and *E* candidate sets created in the above step.
6 (C) **NE Re-identification & Alignment:** Rank each candidate in the NE-Pair-Candidate set constructed above with the score specified by the proposed model. Let *Nc* and *Ne* be the numbers of those initial Chinese and English NEs in the first stage, respectively, and set $K = min(Nc, Ne)$. Extract top *K* final NE pairs (with their re-assigned NE types) with the highest scores from the NE-Pair-Candidate set.
7 (C.1) FOR each NE pair in the NE-Pair-Candidate set created above:
FOR each re-assigned NE type within {PER, LOC, ORG}
Evaluate the score for the given candidate pair and the given NE type according to the proposed model.
END_FOR
Find the re-assigned NE type with the highest score, then attach it and its corresponding score to the given NE pair
END_FOR (C.1)
8 (C.2) Conduct a beam search process to select the top *K* non-overlapping NE pairs from the NE-Pair-Candidate set with the scores assigned above. The searching process will keep removing those overlapping NEs from the candidate list before each state is branched.

Figure 1
A framework for jointly identifying and aligning bilingual NEs.

Example 4 illustrates how the framework works. For simplicity, we will allow the Chinese NE to enlarge/shrink its boundaries to four characters on each side, and only allow two words for English.

Example 4. An example of candidate set construction

Each NE and its type in this example is separated by "/". Only partial and relevant information is shown here.

(A.1) A Chinese tagged sentence: "据报道⟨加拉巴戈斯/PER⟩ 国家公园以及当地渔民，正合...。";
(A.2) Initial Chinese NE (*Cb*): ⟨加拉巴戈斯/PER⟩, $Nc = 1$;
(A.3) An English tagged sentence: "The report said the [Galapagos/PER] [National Park/ORG] and local fishermen were working together ...";
(A.4) Initial English NEs (*Eb*): [Galapagos/PER], [National Park / ORG], $Ne = 2$;
(B.1.1) Regenerated Chinese candidate set (C_1^K): {⟨加拉巴戈斯⟩, ⟨加拉巴戈斯国家公园⟩, ⟨加⟩, ⟨拉巴戈斯国家公园⟩, ⟨拉巴⟩, ⟨拉⟩, ⟨巴⟩, ⟨戈⟩, ⟨斯⟩, ⟨据报道加拉巴戈斯国家公园⟩, ⟨据报道加拉巴戈⟩, ...}. Total 62 *C* candidates will be generated.
(B.1.2) Regenerated English candidate set (E_1^K): {[Galapagos], [National Park], [Galapagos National Park], [said the Galapagos National Park], [National Park and local], [National], [Park], ...}. Total 24 *E* candidates will be generated.
(B.2) NE-Pair-Candidate set: {⟨加拉巴戈斯⟩::[Galapagos], ⟨加拉巴戈斯国家公园⟩::[Galapagos National Park], ⟨加拉巴戈斯⟩::[National Park], ⟨加拉巴戈斯国家公园⟩::[Park], ...}. Total 1,488 (62×24) NE pairs will be generated.
(C) $K = min(Nc, Ne) = 1$. Therefore, only {⟨加拉巴戈斯国家公园⟩::[Galapagos National Park], ORG} will be extracted.

This example shows that the desired Chinese NE ⟨加拉巴戈斯国家公园⟩ is partially recognized as ⟨加拉巴戈斯⟩ initially with an incorrect NE type PER. In addition, the desired English NE [Galapagos National Park] is split into [Galapagos] and [National Park], initially two NEs. After each NE Candidate set has been expanded, the desired NEs ⟨加拉巴戈斯国家公园⟩ and [Galapagos National Park] are included in *C* and *E* candidate sets, respectively. The desired NE pair {⟨加拉巴戈斯国家公园⟩::[Galapagos National Park], ORG} can thus be located.

If we allow the boundaries to be enlarged/shrunk without any limitation during the expansion step, all NEs that are initially incorrectly recognized can then be included. The generated search space, however, would be too large to be tractable. In our observation, four Chinese characters for both shrinking and enlarging, and two English words for shrinking and three for enlarging, are found to be adequate in most cases. (Note that only the candidate that contains at least one original character/word is allowed.) Under this condition, the inclusion rates for NEs with correct boundaries can be increased to 94.6% (from 88.7%) for Chinese, and 96.3% (from 92.8%) for English, respectively; the NE pair inclusion rate can even be increased to 95.3% from 83.9%. Because the inclusion rate achieved by this strategy (with limited range) is only 0.8% lower than that obtained without any range limitation (which is 96.1%, as some NEs might have been completely missed in the first stage), this setting is adopted in this article to reduce the search space. Even with this expansion strategy, however, those missing and spurious (false positive) errors still cannot be remedied, because we will neither create additional anchors nor delete any existing anchor.

4. Experiments on Various Configurations

To evaluate the proposed approach, prior work (Huang, Vogel, and Waibel 2003) is re-implemented as our baseline (see Section 4.2). This is because the work not only adopts the same candidate set expansion strategy mentioned previously, but also utilizes monolingual information when selecting NE pairs (only a simple bigram model is used, however). This is in contrast to other works (Feng, Lv, and Zhou 2004; Lee, Chang, and Jang 2006), which only used alignment scores.

The same training and test sets are used for the various experiment configurations. The adopted training set includes two parts. The first part consists of 110,874 aligned sentence pairs from newswire data in the Foreign Broadcast Information Service (LDC2003E14[7]) corpus, which is denoted as Training Set I. The average length of the Chinese sentences in this data set is 74.6 characters, and the average length of the English sentences is 30.2 words. Training Set I is initially tagged by Chinese/English NE taggers, and then reference NE boundaries and types are manually labeled. The second part of the training set is the LDC2005T34[8] bilingual NE pair list with a total of 218,772 NE pairs, which is denoted as Training Set II. The required features (e.g., NE type and translation-mode) are then manually labeled throughout the two training sets. Because Training Set II only contains isolated NE pairs that are not associated with their surrounding context, Training Set I is thus required to train those context-related parameters.

In the baseline system, translation cost and transliteration cost models are trained on Training Set II, and tagging cost is trained on Training Set I. For the proposed approach, the NE likelihoods are trained on Training Set I, and Training Set II is used to train the parameters relating to the NE alignment probability.

For the test set, 300 sentence pairs are randomly selected from the Linguistic Data Consortium (LDC) Chinese–English News Text (LDC2005T06) corpus, which contains at least one NE pair in each sentence. The average length of Chinese sentences is 59.4 characters, and the average length of English sentences is 24.8 words. The answer keys to NE recognition and alignment are annotated manually, and used as the gold standard to calculate the metrics of precision (P), recall (R), and F-score (F) for both NE recognition and alignment. A total of 765 Chinese NEs and 747 English NEs are manually identified in the test set, in which there are 718 reference NE pairs (including 214 PER pairs, 371 LOC pairs, and 133 ORG pairs). NE alignment result is a subset of NE recognition results, because not all those recognized NEs can be aligned.

The development set for feature selection and weight training is composed of 200 sentence pairs selected from the LDC2005T06 corpus, which includes 482 manually tagged NE pairs. The average length of Chinese sentences is 56.4 characters, and the average length of English sentences is 23.2 words. There is no overlap between the training, development, and test sets.

These data sets will be adopted in a series of experiments that investigate the proposed model. Among them, the results of initial NE recognition are given in Section 4.1, and those related to the baseline system are given in Section 4.2. In Section 4.3, a series of experiments are conducted to examine the effect of various features adopted in the proposed model. The weighted version of the proposed model is also tested.

7 FBIS multilingual text (`http://projects.ldc.upenn.edu/TIDES/mt2003.html`).

8 The LDC2005T34 data set consists of proofread bilingual entries: 73,352 person names, 76,460 location names, and 68,960 organization names.
(`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T34`).

Table 2
Initial type-sensitive Chinese/English NER performance.

NE type	P (%)	R (%)	F (%)
PER	80.2/79.2	**87.7/85.3**	83.8/82.1
LOC	**89.8/85.9**	87.3/81.5	**88.5/83.6**
ORG	78.6/82.9	82.8/79.6	80.6/81.2
ALL	83.4/82.1	86.0/82.6	84.7/82.3

Furthermore, the effectiveness of adopting different initial NE recognizers is shown in Section 4.4, and the effectiveness of the proposed model across different domains is illustrated in Section 4.5. Finally, the result of directly using all available features under a Maximum Entropy framework without developing a principled model is given in Section 4.6.

4.1 Initial NE Recognizers

Both the baseline alignment system and the proposed model share the same Initial NE Recognition subtask. The systems adopt the Chinese NE recognizer reported in Wu, Zhao, and Xu (2005), which is a hybrid statistical model incorporating multi-knowledge sources, and the English NE recognizer included in the publicly available Mallet toolkit[9] (McCallum 2002) to generate initial NEs. These two initial NE recognizers are adopted because their performance is comparable to other state-of-the-art systems (Gao, Li, Wu, and Huang 2005; Zhou and Su 2006). The NE recognition baseline performances reported subsequently are provided by these two packages. A total of 789 Chinese NEs and 752 English NEs are recognized.

Table 2 shows the initial NE recognition (**NER**) performance for both Chinese and English (the highest performance in each column is in bold). It is observed that the F-score of ORG type is the lowest among all NE types for both English and Chinese. This is because many organization names are only partially recognized or missed altogether. In addition, the precision rate of PER type is lowest among all English NE types because many location names or abbreviated organization names tend to be incorrectly recognized as person names in English. In general, the initial Chinese NER outperforms the initial English NER, as the NE type classification turns out to be a more difficult problem for this English NER system.

4.2 The Baseline System

The model of Huang, Vogel, and Waibel (2003) is re-implemented in our environment as the baseline system, and is briefly sketched here for the reader's convenience. There are three cost features in Huang's alignment model: (1) transliteration cost, which measures the phonetic similarity of the aligned NEs; (2) translation cost, which is similar to IBM model-1 (Brown et al. 1993); and (3) tagging cost, which evaluates bigram probabilities of the aligned NEs based on the same NE type.

9 http://mallet.cs.umass.edu/index.php/Main_Page.

Table 3
NEA type-insensitive (type-sensitive) performance on the test set.

Model	P (%)	R (%)	F (%)
ExpB (*Baseline*)	77.1 (67.1)	79.7 (69.8)	78.4 (68.4)
Exp1 (*B-Probabilities*)	76.2 **(72.3)**	78.5 (**74.6**)	77.3 (**73.4**)
Exp2 (*B-Probabilities_N-Alignment*)	**77.7 (73.5)**	79.9 (**75.7**)	78.8 (**74.6**)
Exp3 (*N-Full_Model*)	**83.7 (78.1)**	**86.2 (80.7)**	**84.9 (79.4)**
Exp4 (*MERT-W*)	**85.9 (80.5)**	**88.4 (83.0)**	**87.1 (81.7)**

In our experiments, the translation cost of the baseline system is trained on Training Set I (with 110,874 aligned sentence pairs) by the GIZA++ toolkit (Och and Ney 2003), the transliteration cost is trained on all person names (all are transliterated) and transliterated location and organization names included in Training Set II. The tagging cost is trained on the tagging result of Training Set I by the initial NE detection system.

When those initially identified NEs are directly used for alignment, only a 64.1% F-score (regarding their NE types) is obtained from this baseline system. This relatively poor performance is mainly due to errors in the initial NE recognition stage that are brought into the subsequent alignment stage. To diminish the accumulative effect of errors, the same expansion strategy described in Section 3.3 is then adopted to enlarge the possible NE candidate set. However, only a slight improvement is obtained (from 64.1% to 68.4% for type-sensitive F-score), as shown in Table 3 in Section 4.3. Therefore, it is conjectured that the baseline alignment model is unable to perform well if the features proposed in this article are not adopted.

4.3 The Re-identification and Alignment Joint Model

To examine the individual effect of features adopted in the model, a series of experiments are first conducted on the development set. All features mentioned in Section 3 are verified by their contributions and are then adopted for further experiments on the test set. Table 3 lists only the representative performance of NE alignment (NEA) on the test set, and gives two performance measures for the experiments. The first one (named **type-insensitive**) only checks the scope of each NE without taking its associated NE type into account (which is the approach adopted in most of the literature on NE recognition), and is reported as the main metric in Table 3. The second one (named **type-sensitive**) also evaluates the associated NE type of each NE. To evaluate the type-sensitive performance for NE pairs with correct boundaries, we give one point to any NE pair that also possesses the correct type-tags on both sides, and give 0.5% if only one side is correct. Of course, zero points are given if both types are incorrect or the boundary of any NE is incorrect. With the rules specified herein, the type-sensitive results are also given within the parentheses in Table 3, and a large degradation is observed. The configurations of various experiments are listed as follows.

ExpB: This is the *baseline system* (Huang, Vogel, and Waibel 2003), which is re-implemented in our environment for comparison.

Exp1: Exp1 (named *B-Probabilities*) adopts *all bilingual related probabilities* involved in Equations (6) and (7), to show the full power of bilingual probabilities.

Exp2: Furthermore, because the NE alignment probability would favor the candidates with fewer components,[10] it is further *normalized* by converting Equation (6) into the following form:

$$\max_{A} \left\{ \left[\prod_{n=1}^{N_A} P(cp_{A,a(n)}|M_{A,n}, ew_{A,n}, T) \times P(M_{A,n}|ew_{A,n}, T) \right]^{\frac{1}{N_A}} \times P(\delta_A|T) \right\}$$

The experiment covering complete bilingual probabilities, with the normalized versions of Equations (6) and (7), is denoted as Exp2 (named *B-Probabilities_N-Alignment*).

Table 3 indicates that Exp2 achieves the best performance (both type-insensitive and type-sensitive) among different combinations of the bilingual-related probabilities. Due to its effectiveness, the *normalized* bilingual probabilities are hence adopted in all subsequent experiments (i.e., all are based on Exp2).

Exp3: Exp3 (named *N-Full_Model*) manifests the full power of the proposed recognition and alignment joint model, by integrating all monolingual options into Exp2. Note that we use the SRI Language Modeling Toolkit[11] (Stolcke 2002) to train various character-/word-based bigram models on different NE types. They are trained with modified Kneser-Ney smoothing (Kneser and Ney 1995, Chen and Goodman 1998). Note that monolingual bigrams are also normalized with their numbers.

As Exp3 shows, the best configuration is to take advantage of all features proposed in this article and to normalize all feature probabilities. This configuration will thus be taken for further improvement in the following sections.

So far the proposed model weighs all features equally. It is reasonable to expect that features should be weighted differently according to their contribution, however. Those weighting coefficients can be learned from the development set via the well-known **Minimum Error Rate Training** approach (Schlüter and Ney 2001; Och 2003) (commonly abbreviated as MERT). To save computational cost, we only re-evaluate the scores of candidate pairs in a pre-generated pool, instead of regenerating new candidate pairs each time when W_t (the vector of weighting coefficients at i-th iteration) is updated to W_{t+1} (of the next iteration). For each sentence pair, its corresponding pool is first created by using W_0 (i.e., Exp3) to generate the top 50 NE pairs[12] resulting from the beam search process. Subsequently, when we switch W_t to W_{t+1}, we only re-score (and then re-rank) those candidate pairs inside the pool according to W_{t+1}.

Exp 4 presents the weighted version of the proposed joint model obtained from MERT training (*MERT-W, N-Full_Model*, abbreviated as MERT-W). The result demonstrates that MERT is effective and useful. Entries in bold indicate that the model significantly outperforms the baseline system. (All statistical significance tests in this article are measured with 95% confidence level on 1,000 re-sampling batches [Zhang, Vogel, and Waibel 2004]).

10 It is biased by the E word-count due to the *sufficiency* (Freedman 2005; Liese and Miescke 2008) assumption made during decomposition.

11 `http://www.speech.sri.com/projects/srilm/`.

12 Because most sentence pairs possess less than four NE pairs under Exp3, 50 NE pairs should be sufficient.

Table 4
NER type-insensitive (type-sensitive) performance of different English NE recognizers.

English NE recognizers	P (%)	R (%)	F (%)
Mallet Toolkit	91.8 (82.1)	92.4 (82.6)	92.1 (82.3)
Stanford NE recognizer	93.7 (84.7)	91.4 (82.3)	92.5 (83.5)
Minor Third	90.8 (82.1)	89.5 (80.7)	90.1 (81.4)

Table 5
NEA type-insensitive (type-sensitive) performance with the same Chinese NE recognizer (Wu's system) and different English NE recognizers.

NE alignment on different recognizers	P (%)	R (%)	F (%)	Upper bound (%)
Mallet Toolkit	85.9 (80.5)	88.4 (83.0)	87.1 (81.7)	95.3
Stanford NE recognizer	85.9 (80.2)	88.4 (82.7)	87.1 (81.4)	95.0
Minor Third	85.7 (80.2)	88.1 (82.7)	86.9 (81.4)	94.2

Compared to the baseline system, the MERT-W version has substantially raised the test set type-insensitive F-score of identified NE pairs from 78.4% to 87.1% (11.1% relative improvement), and the type-sensitive F-score from 68.4% to 81.7% (19.4% relative improvement). Therefore, this MERT-W version is adopted in all further experiments.

4.4 Effect of Adopting Different Initial NE Recognizers

To study whether the final performance of NE alignment is sensitive to the choice of initial NE recognizers, we investigate the final alignment performance across different Chinese and English NE recognizers.

First, we test the NE alignment performance with the same Chinese NE recognizer (Wu's system, adopted earlier) but with different English NE recognizers that include the Mallet toolkit (used before), the Stanford NE recognizer (Finkel, Grenager, and Manning 2005), and Minor Third (Cohen 2004). Table 4 shows the type-insensitive and type-sensitive (within parentheses) results. Table 5 shows the effect on NE alignment performance. From Tables 4 and 5, we find that NE alignment performance is actually not sensitive to the NE recognition result. Although the performance of different NE recognizers are various (type-insensitive[13] F-scores are 90.1%, 92.1%, and 92.5%, respectively), the gaps among their corresponding NE alignment results are negligible (type-sensitive F-scores of weighted versions are 81.7%, 81.4%, and 81.4%, respectively), as their candidate sets are enlarged based on initially recognized NEs. It is also noteworthy that although the F-score of the Stanford NE recognizer is higher than that of the Mallet toolkit, its corresponding NE alignment performance is lower than the model based on the Mallet toolkit. We conjecture that the lower recall of Stanford NE

13 Because the initial NE recognizer mainly provides NE anchors, NE type is less relevant to the following alignment.

Table 6
NER type-insensitive (type-sensitive) performance of different Chinese NE recognizers.

Chinese NE recognizers	P (%)	R (%)	F (%)
Wu's System	86.2 (83.4)	88.7 (86.0)	87.4 (84.7)
BaseNER	88.3 (85.9)	84.9 (82.5)	86.6 (84.2)
S-MSRSeg	86.8 (84.7)	81.5 (79.5)	84.1 (82.1)

Table 7
NEA type-insensitive (type-sensitive) performance with the same English NE recognizer (Mallet system) and different Chinese NE recognizers.

NE alignment on different recognizers	P (%)	R (%)	F (%)	Upper bound (%)
Wu's System	85.9 (80.5)	88.4 (83.0)	87.1 (81.7)	95.3
BaseNER	85.6 (79.9)	88.1 (82.4)	86.8 (81.1)	94.2
S-MSRSeg	84.5 (78.9)	87.1 (81.4)	85.8 (80.1)	93.3

recognizer leads to lower NE alignment performance because the recall of NER is closely related to the NE pair inclusion rate (please refer to footnote 1), which is the upper bound of its corresponding NE alignment performance.

Similarly, we test the NE alignment performance with the same English NE recognizer (Mallet) but with different Chinese NE recognizers, including Wu's system (as before), BaseNER (Zhao and Kit 2008), and S-MSRSeg (Gao, Li, Wu, and Huang 2005). The comparisons are given in Tables 6 and 7. From these tables we also see that the NE alignment result is not sensitive to the NE recognition result (84.1% to 87.4% type-insensitive for NER vs. 80.1% to 81.7% type-sensitive in F-score for NEA), although the performance of NE alignment is related to the recall of the Chinese NE recognizer (the weaker side). We also note that the type-insensitive F-score performance gap among various English NE recognizers in Table 5 is less than that of the Chinese NE recognizers in Table 7 (0.2% vs. 1.3%), which is mainly due to the different gaps among their original performances (2.4% vs. 3.3%, shown by Tables 4 and 6).

Furthermore, NE alignment based on the worst Chinese NE recognizer (S-MSRSeg) and the worst English NE recognizer (Minor Third) is conducted in Table 8. From Table 8 we see that the final NE alignment performance is primarily determined by the weaker side, which is the one that gives the lower recognition recall rate. In this particular case, the performance of the combination of S-MSRSeg and Minor Third (79.9% type-sensitive F-score) is mainly driven by the performance of the Chinese S-MSRSeg (80.1%

Table 8
NEA type-insensitive (type-sensitive) performance with a different English NE recognizer and another Chinese NE recognizer.

NE alignment on different recognizers	P (%)	R (%)	F (%)	Upper bound (%)
Wu & Mallet	85.9 (80.5)	88.4 (83.0)	87.1 (81.7)	95.3
S-MSRSeg & Minor	84.7 (78.6)	87.3 (81.3)	86.0 (79.9)	93.9

Table 9
Initial NE recognition type-insensitive (type-sensitive) performance across various domains.

Different domains	Language	P (%)	R (%)	F (%)
News (Table 2)	Chinese (Wu)	86.2 (83.4)	88.7 (86.0)	87.4 (84.7)
	English (Mallet)	91.8 (82.1)	92.4 (82.6)	92.1 (82.3)
HK Hansards	Chinese (Wu)	91.4 (88.5)	89.1 (87.3)	90.2 (87.9)
	English (Mallet)	93.5 (90.4)	94.3 (91.2)	93.9 (90.8)
Computer	Chinese (Wu)	82.7 (81.4)	87.9 (86.5)	85.2 (83.9)
	English (Mallet)	76.6 (72.9)	88.9 (85.2)	82.3 (78.6)

Table 10
The superiority of our joint model on three different domains indicated by type-insensitive (type-sensitive) performance (those significant entries are marked in comparison with baseline).

Different domains	Model	P (%)	R (%)	F (%)
News (Table 3)	Baseline	77.1 (67.1)	79.7 (69.8)	78.4 (68.4)
	Proposed Model	**85.9 (80.5)**	**88.4 (83.0)**	**87.1 (81.7)**
HK Hansards	Baseline	86.3 (83.3)	87.1 (84.1)	86.7 (83.7)
	Proposed Model	88.2 (**86.5**)	89.1 (**87.3**)	88.6 (**86.9**)
Computer	Baseline	69.4 (66.1)	80.3 (77.1)	74.5 (70.3)
	Proposed Model	**75.5 (72.4)**	**86.2 (83.1)**	**79.6 (76.5)**

in Table 7). This is because the NE pair inclusion rate is usually dominated by the weaker side.

4.5 Effectiveness of the Proposed Model Across Different Domains

To test the effectiveness of the joint model across domains, we compare the baseline and our joint model on three different domains (News, HK Hansards, and Computer Technology). To do this, two other test sets are selected from HK Hansards (LDC2004T08) and from the computer domain (training data in CWMT08),[14] respectively (the test set used in the previous sections is from the News domain). Each of these new test sets also includes 300 randomly selected sentence pairs.

Table 9 shows the initial NE recognition performance across those three different domains. Also, it is clear from Table 10 that our joint model outperforms the baseline in all three domains, which indicates that the advantage of our joint model holds over various domains. On the other hand, the smaller improvement observed in the HK Hansards domain might be due to the possibly easier task of initial NE recognition and NE alignment.[15] (Note that the baseline performance in this domain is much higher than others—with an NE alignment type-sensitive F-score of 83.7% compared with

14 `http://nlpr-web.ia.ac.cn/cwmt-2008`.

15 Note that 40.3% sentence pairs in the HK Hansards corpus contains only one NE pair (alignment would be trivial in this case); this ratio is 15.7% and 27.0% for News and Computer domains, respectively.

Table 11
Comparison between a ME framework and the derived model on the same test set.

Model Data set–Size	400	4,000	40,000	90,412
ME Framework (Maxent)	38.9 (0%)	51.6 (0%)	63.8 (0%)	69.5 (0%)
ME Framework (YASMET)	−2.4 (−6.2%)	−1.2 (−2.3%)	−1.2 (−1.9%)	−1.6 (−2.3%)
Weighted-Joint-Model	**+2.6** (**+6.9%**)	**+3.5** (**+6.9%**)	**+3.4** (**+5.3%**)	**+2.9** (**+4.2%**)

68.4% and 70.3% in News and Computer domains, respectively). Therefore, those novel features of our joint model are not crucial in easy cases.

4.6 Maximum Entropy Framework with Primitive Features

We propose and derive the model described previously in a principled manner. One might wonder, however, whether it is worthwhile to derive such a model after all related features have been proposed, as all proposed features can also be directly integrated into the well-known maximum entropy (ME) framework (Berger, Della Pietra, and Della Pietra 1996) without making any assumptions. To show that not only features, but also the adopted model contributes to performance improvement, we build an ME model that directly adopts all primitive features mentioned previously as its input (including the internal component alignment-pair, initial and final NE type, NE bigram-based string, and left/right distance), without involving any related probabilities derived in the proposed model.

Because an ME approach can be trained only on linked NE pairs, those sentence pairs that include at least one NE pair are first extracted from Training Set I. A total of 90,412 sentence pairs are obtained, as some sentence pairs only have either Chinese or English NEs, and 298,302 NE pairs are identified. This ME method is implemented with the YASMET[16] package, and is tested under various training-set sizes (400, 4,000, 40,000, and 90,412 sentence pairs). Because the NEs of the bilingual NE pair list (Training Set II) do not contain their corresponding sentences, the ME approach lacks the necessary context to extract specific ME features and hence this list is left out of our training data for both the baseline ME model and our joint model.

In order to compare different ME approaches, we also try Zhang's Maxent package[17] with five classes (i.e., PER, LOC, ORG, Incorrect-Boundaries, Correct-Boundaries-Incorrect-Type). A five-class approach outperforms a three-class approach (YASMET) in this case (it has many more features as well). Table 11 shows only the type-sensitive F-scores evaluated on the same test set to save space. The data within the parentheses are relative improvements, and entries in bold indicate that the performance of the derived model is statistically better than that of the ME models.

The improvement indicated in Table 11 clearly illustrates the benefit of deriving the model. Because a reasonably derived model not only shares the same training set with the primitive ME approach, but also enjoys the additional knowledge introduced by the human researcher (i.e., the assumptions/constraints implied by the model), it is not surprising that a good model does perform well, and the relative improvement becomes more noticeable when the training set becomes smaller.

16 `http://www.fjoch.com/YASMET.html`.
17 `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.

When we take a closer look at the 18 instances where our model correctly identifies the NE alignments and Maxent fails to do so on the test set, we find that the internal component alignment-pair feature in the Maxent approach to be dominant in causing 56% of the errors (10 out of 18). In contrast, our corresponding internal mapping probability $P(cp_{a(n)}|M_n, ew_n, T)$ in Equation (5) makes the correct decisions 90% of the time (9 out of 10). In fact, even $P(cp_{a(n)}|ew_n, T)$ (the simpler form) makes correct predictions 80% of the time (8 out of 10).

One example of the ten errors made by Maxent is "玻利维亚举行" (Bolivia holds), which is incorrectly linked to *Bolivia holds* as a NE pair by the ME approach, whereas our model correctly aligns "玻利维亚" with *Bolivia*. This is because "玻利维亚" is transliterated into *Bolivia*, but "举行" is translated into *holds*. Given T = LOC, the internal mapping probability $P(cp_{a(n)}|ew_n, T)$ thus disfavors the translation mode between "举行" and *hold* within an LOC NE pair, and prefers the correct result. This example illustrates the utility of the explicit dependency constraint imposed by the model, which is not possible in the ME approach.

5. Discussion and Error Analysis

Although the proposed model substantially improves alignment performance, errors do remain. Therefore, we would like to know what the limitations of the proposed model are, and what kinds of problems still remain—an essential component in finding future directions for further improvements. In the test set, a total of 718 NE reference-pairs and 739 aligned NE pairs are generated from the proposed joint model (MERT-W version). Among the generated NE pairs, there are 104 (out of 739) boundary errors (regardless of their re-assigned types), or 14.1%. Also, among the remaining 635 NE pairs with correct boundaries, 41 (6.5%) are re-assigned to the incorrect NE type. Boundary identification, therefore, is still a crucial problem.

Before investigating the errors made by MERT-W, we would like to understand its limits. As mentioned in Section 3.3, the inclusion rate of those desired NE pairs (within the Cartesian product of expanded candidate sets C_1^K and E_1^K) is the upper bound for the system adopted in the final selection stage, which in the current setting is 95.3%. In comparison, the type-insensitive F-score of MERT-W is 87.1%, indicating that there is still a significant 8.2% scope for improvement, even though a great improvement has already been made over the baseline system. We examine this gap and propose solutions to address the errors in the following section.

5.1 Classification of Type-Insensitive NE Pair Errors

There are 111 type-insensitive NE pair errors in the test set (104 boundary errors plus 7 others not included in the output list due to missing anchors), and these can be classified into the following six main categories.

(I) *Reference Inconsistency* (11%): The NE references from Chinese and English are not correctly matched, which rules out the possibility of generating the correct NE pair.

(II) *Missing Anchor* (14%): Although the NE reference is consistent, not all their associated NE anchors are generated in the initial recognition stage, which cannot be remedied by the expansion strategy adopted in this article.

(III) *Over-generating Anchors* (10%): Similarly, additional spurious NE anchors are also generated in the initial recognition stage, and result in incorrect NE pairs.

The remaining cases with correct corresponding anchors are further classified as follows.

(IV) *Inconsistent Components* (12%): Although their corresponding anchors are correctly generated, the internal components of those reference pairs are originally unmatched due to deletions or insertions occurring in the NE translation. For example, in the reference {⟨乌伯林更镇⟩::[Berlingen]}, the Chinese component "镇" (*town*) is originally unmatched because its English correspondent *town* does not exist in the given English sentence. Therefore, only one incorrect pair {⟨乌伯林更⟩::[Berlingen]} is generated. Other cases with all matched components are further classified in the following.

(V) *Expansion Limitation* (5%): Even though all the internal components are matched, the desired candidate (i.e., the *reference*) is still not covered by the candidate set after its expansion.

(VI) *Others* (48%): Even if all internal components are matched and their references are also included in the candidate set, some errors still remain, mainly due to the limitations of the current model. These cases account for the majority of the errors that will be further analyzed subsequently in the article.

Table 12 shows the distribution of the six defined categories. {⟨CNE⟩::[ENE]} is a specified NE pair, and the unmatched components are underlined. The numbers in

Table 12
Distribution of various error categories (type-insensitive).

Consistency Problem	Anchor Problem	Error Categories	Reference NE pair	Initially Recognized NEs	Final Output	Percentage
Inconsistent References (11%)	NA	(I) *Reference Inconsistency*	⟨佛塞特⟩::[ ]; ⟨⟩::[Northam]	CNE: ⟨塞特⟩ ENE: [Northam]	⟨塞特⟩::[Northam]	11% (12)
Consistent References (89%)	Incorrect Anchors (24%)	(II) *Missing Anchor*	⟨东协⟩::[ASEAN]	CNE: ⟨⟩ ENE: [ASEAN]	No such alignment	14% (16)
		(III) *Over-generating Anchors*	⟨李南信⟩::[Lee Nam-shin]	CNE: ⟨李南信⟩ ENE: [Lee Nam-shin]; [South]	⟨南⟩::[South]	10% (11)
	Correct Anchors (65%)	(IV) *Inconsistent Components*	⟨乌伯林更镇⟩::[Berlingen]	CNE: ⟨乌伯林⟩ ENE: [Berlingen]	⟨乌伯林⟩::[Berlingen]	12% (13)
		(V) *Expansion Limitation* (*Matched Components, Excluded Reference*)	⟨英国葛兰素史克美占药厂⟩::[British Pharmaceutical Firm GlaxoSmithKline]	CNE: ⟨英国⟩; ⟨葛兰素⟩; ⟨史克美⟩ ENE: [British Pharmaceutical]; [Firm GlaxoSmithKline]	⟨英国⟩::[British]	5% (6)
		(VI) *Others* (*Matched Components, Included Reference*)	⟨南北韩⟩::[South and North Koreas]	CNE: ⟨韩⟩ ENE: [North Koreas]	⟨北韩⟩::[North Koreas]	48% (53)

parentheses in the last column denote the number of NE pairs of the corresponding category. Among all categories, Category (I) errors (Reference Inconsistency, 11%) are irrelevant to the alignment model, and are attributed to the asymmetrical distribution of bilingual NEs (corresponding NEs might sometimes be missed or replaced by the pronoun *it*). As illustrated in Table 12, CNE "佛塞特" (*Fossett*) is initially recognized as "塞特" and finally linked to an irrelevant ENE *Northam* (诺森). This occurs because their corresponding counterparts *Fossett* and "诺森" do not appear in the original sentence pair, and our alignment model assumes that the linking between NEs is a one-to-one mapping. One possible simple solution would be to set a minimal threshold on alignment scores that filters out such spurious linking. This may introduce the risk that some correct NE pairs might be pruned away at the same time, however.

Category (II) errors (Missing Anchor, 14%) are due to the absence of the associated NE anchors in the initial recognition stage. As an example in Table 12, the corresponding anchor of the Chinese NE "东协" (*ASEAN*) is not initially identified. Because each candidate set is generated from the given anchor, a missing anchor implies that its associated candidate set will not exist, thereby making it impossible to generate the corresponding NE pair. Although increasing the number of output anchors generated from the initial recognition stage can relieve this problem, doing so makes the subsequent alignment task harder. Additionally, the spurious anchors generated might introduce even more errors.

The errors in Category (III) (Over-generating Anchors, 10%) are due to spurious anchors generated in the initial recognition stage. For instance, the CNE "李南信" is originally aligned with the ENE *Lee Nam-shin* by transliteration. A spurious ENE *South* is also identified in the initial stage, however. This spurious ENE *South* is then incorrectly linked to a virtual CNE with the highest score, "南," which is a sub-string of the desired CNE "李南信," and also a Chinese translation for *south*. This prevents the correct NE pair from being generated. Both missing and over-generating anchor problems are largely dependent on the NE recognition toolkits adopted in the initial stage. Using the current expansion strategy, the initial NE recognizers with lower recall (or precision) tend to result in worse NE-pair recall (or precision) in the final alignment stage.

Category (IV) errors (Inconsistent Components, 12%) are caused by internal components within NEs that were not originally matched. Because words in NEs are not always translated literally, there are insertions and deletions during NE translation. As an example, shown in Table 12 and illustrated previously, the incorrect result {⟨乌伯林更⟩::[Berlingen]} is generated for its reference {⟨乌伯林更镇⟩::[Berlingen]}, as its Chinese component "镇" (*town*) is originally unmatched. In the worse case, those unmatched components could interleave with matched components within the NE pair, and thus prevent some matched components from being included. For example, in the reference {⟨欧盟执行委员会⟩::[European Commission]}, both the Chinese components "盟" (*alliance*) and "执行" (*execution*) have no counterparts, and they would have prevented the matched portion {⟨委员会⟩::[Commission]} from being included in the final output. As a result, only {⟨欧⟩::[European]} is eventually extracted. To tackle this problem of component insertions/deletions that sometimes occur in English–Chinese translation, the alignment model should be further enhanced to allow the component to be linked to an empty element, NULL. Introducing this freedom, however, might have the side effect of including additional spurious Chinese characters (or English words). Further study is required to justify this idea; given that this category accounts for only 12% of the errors, we propose to defer this for later studies.

Furthermore, Category (V) errors (Expansion Limitation, 5%) are caused by the problem that the desired candidate (i.e., reference) is excluded during the candidate

set expansion stage. Table 12 shows that the final output of the reference {⟨英国葛兰素史克美占药厂⟩::[British Pharmaceutical Firm GlaxoSmithKline]} is {⟨英国⟩::[British]}. This reference has three Chinese initial anchors: "英国" (*British*), "葛兰素" (*Glaxo*) and "史克美" (*SmithKline*), and it also has two English initial anchors: *British Pharmaceutical* and *Firm GlaxoSmithKline*. Because only four characters are allowed for boundary enlarging/shrinking for Chinese anchors (three words for English anchors), the reference CNE is beyond the scope of any Chinese initial anchor during the expansion stage. Therefore, it could not be included in the candidate set for final selection. In addition, the adjacent Chinese component "葛兰素" could not be recovered due to the translation re-ordering of the Chinese components—its counterpart *Glaxo* is far apart from *British* in the given sentence. Similarly, the adjacent English words *Pharmaceutical Firm* could not be recovered, as its counterpart "药厂" is far apart from "英国." Although loosening the constraint during the expansion stage can increase the reference coverage, it must be weighed against the corresponding lower precision.

Finally, for those NE pairs with aligned components and included references, the proposed model still makes a significant number of mistakes. These kinds of errors, Category (VI) (Others, 48%), account for the largest portion among all errors. Therefore, they are further hierarchically classified in Table 13 according to their associated transformation types and origins.

The incorrect NE pairs with aligned components and included references (i.e., [VI] Other Category in Table 12) are first classified by their corresponding transformation types: (VI.A) *Abnormal Transformation* (27%), whose transformation types are not assumed by the model (i.e., neither normally translated nor normally transliterated); and (VI.B) *Normal Transformation* (21%), whose components are either normally translated or normally transliterated. A detailed explanation is given as follows.

Table 13
Distribution of Category (VI) error classes (type-insensitive).

Transformation Type	Classes	Reference NE pair	Initially Recognized NEs	Final Output	Percentage
(VI.A) Abnormal Transformation (27%)	(VI.A.1) English Acronym	⟨国际刑事法庭::[ICC]	CNE: ⟨国际刑事法庭⟩; ⟨法新社⟩ ENE: [ICC]	⟨法新社⟩::[ICC]	11% (12)
	(VI.A.2) Chinese Abbreviation	⟨维和部队⟩::[Peace-keeping Troop]	CNE: ⟨部队⟩ ENE: [Peace-keeping Troop]	⟨部队⟩::[Troop]	8% (9)
	(VI.A.3) Irregular Translation	⟨明仁⟩::[Akihito]	CNE: ⟨明仁⟩ ENE: [Akihito]	⟨明仁⟩::[Hirohito]	8% (9)
(VI.B) Normal Transformation (21%)	(VI.B.1) Bias from NE likelihoods	*Translation*: ⟨南北韩⟩ ::[South and North Koreas] *Transliteration*: ⟨毕翠克丝⟩ ::[Beatrix]	*Translation*: CNE: ⟨韩⟩ ENE: [North Koreas] *Transliteration*: CNE: ⟨毕翠⟩; ⟨克丝⟩ ENE: [Beatrix]	*Translation*: ⟨北韩⟩ ::[North Koreas] *Transliteration*: ⟨毕翠克⟩ ::[Beatrix]	18% (20)
	(VI.B.2) Bias from Bilingual Probabilities	*Translation*: ⟨世界杯⟩ ::[World Cup] *Transliteration*: ⟨伊斯兰堡⟩ ::[Islamabad]	*Translation*: CNE: ⟨世界杯⟩ ENE: [World Cup] *Transliteration*: CNE: ⟨伊斯兰堡⟩ ENE: [Islamabad]	*Translation*: ⟨世界杯冠军⟩ ::[championship at World Cup] *Transliteration*: ⟨伊斯兰堡地⟩ ::[Islamabad]	3% (3)

(VI.A) **Abnormal Transformation** (27%): This category includes transformation types that are not assumed by our alignment model. It can be further divided into three classes according to their origins: (VI.A.1) *English Acronym* (11%), whose ENE is an acronym; (VI.A.2) *Chinese Abbreviation* (8%), whose CNE is an abbreviation; and (VI.A.3) *Irregular Translation* (8%), whose components are transformed neither semantically nor phonetically. These cases are interesting and are illustrated herein.

In the row (VI.A.1) (English Acronym, 11%) of Table 13, a Chinese NE "国际刑事法庭" (*International Criminal Court*) is tagged as "国际刑事法庭/ORG," whereas its English counterpart is the acronym *ICC*. Linking "国际刑事法庭" to *ICC* is thus beyond the ability of our model. On the other hand, Chinese NEs are also occasionally abbreviated. For example, in Class (VI.A.2) (Chinese Abbreviation, 8%), "维和" is the Chinese abbreviation of "维持和平" (*Peace-keeping*), which is also difficult to align to its English counterpart. Such acronym and abbreviation cases are not rare in NE translation. We believe that an expansion table (or even anaphora analysis) for acronyms and abbreviations can help handle such issues.

It is also known that some loanwords or out-of-vocabulary terms are translated neither semantically nor phonetically. As an example for Class (VI.A.3) (Irregular Translation, 8%), CNE "明仁" (which is the name of a Japanese emperor, and consists of Japanese kanji characters) is incorrectly linked to an English word *Hirohito* (whose Chinese translation should be "昭和"), although it should be linked to ENE *Akihito*. In this example, the Japanese kanji "明仁" is directly adopted as the corresponding Chinese characters (as those characters are originally borrowed from Chinese), which would be pronounced as *ming-ren* in Chinese and thus deviates significantly from the English pronunciation of *Akihito*. Therefore, it is translated neither semantically nor phonetically. This phenomenon mainly occurs in loanwords or out-of-vocabulary terms and the model would have to be extended to cover those new conversion types. Such an extension is very likely to be language-pair dependent (e.g., with an additional Japanese phonetic table for cases such as the given example), however.

(VI.B) **Normal Transformation** (21%): Components of this category are translated normally. It can be further divided into two classes according to their sources: (VI.B.1) *Bias from NE likelihoods* (18%), which prefers the incorrect NE pair scope due to its associated monolingual likelihood scores, and (VI.B.2) *Bias from Bilingual Probabilities* (3%), which introduces extra non-NE words in the output due to high alignment scores of words that are adjacent to the NE. Further illustration is given as follows.

As Class (VI.B.1) (*Bias from NE likelihoods*, 18%) shows in Table 13, the Chinese NE "南北韩" and the English NE *South and North Koreas* are initially recognized as "韩" (*Korea*) and *North Koreas*, respectively; the model finally chooses a partial alignment result {⟨北韩⟩::[North Koreas]}. In this case, every component in either the CNE or the ENE is well matched to its counterpart. Therefore, there is no significant difference among the alignment scores of various NE pair candidates with different scopes (such as {⟨韩⟩::[Koreas]}, {⟨北韩⟩::[North Koreas]}, and {⟨南北韩⟩::[South and North Koreas]}, etc.)

The same situation also appears in the transliteration case. For example, the final output of the reference {⟨毕翠克丝⟩::[Beatrix]} is {⟨毕翠克⟩::[Beatrix]}, and it has two Chinese initial anchors: "毕翠" (pronounced as *bi-cui* in Chinese) and "克丝" (*ke-si*). According to the training data, both "克" (*ke*) and "克丝" (*ke-si*) could be aligned to the English letter *x*. There is no significant difference, therefore, in the alignment scores between {⟨毕翠克⟩::[Beatrix]} and {⟨毕翠克丝⟩::[Beatrix]}.

Because the NE alignment feature has only negligible discrimination power in these cases (as described in Section 2.2), monolingual likelihood scores dominate the scope

preference. Addressing this shortcoming is beyond the capability of the alignment model, and the adjacent contextual (non-NE) bigrams, to be proposed later, can only correct 5 of the 20 errors in this class. No easy and effective solution for this kind of problem can currently be found.

Contrary to this case, Class (VI.B.2) (*Bias from Bilingual Probabilities*, 3%) accounts for those cases where incorrect NE pairs are selected due to the bilingual alignment score. For example, the final output of the reference {⟨世界杯⟩::[World Cup]} is {⟨世界杯冠军⟩::[championship at World Cup]}. Although both desired CNE and ENE have already been correctly recognized in the initial stage, the bilingual NE alignment feature prefers to include the additional Chinese common noun "冠军" and an extra English word *championship*, because they are a perfect mutual translation. At first glance, it seems that we could use the lower casing of *championship* as a feature. Other references with lower case words could also be found (e.g., {⟨世贸组织部长会议⟩::[WTO ministerial meeting]}), however. Therefore, additional features such as their relative positions are also required.

The same situation also occurs in transliteration. For example, the Chinese NE "伊斯兰堡" (*yi-si-lan-bao*, in Chinese pronunciation) and the English NE *Islamabad* are both correctly recognized in the initial stage. The model chooses a longer alignment result {⟨伊斯兰堡地⟩::[Islamabad]} in the final stage, however. In this case, the Chinese character "地" (land, pronounced as *di* in Chinese) could also be phonetically aligned to syllable "d" with high probability. Therefore, there is no significant difference in the alignment scores between ⟨伊斯兰堡⟩::[Islamabad] and ⟨伊斯兰堡地⟩::[Islamabad]. We may need to resort to using a richer bilingual context (i.e., "在伊斯兰堡地图上…"; *In the map of Islamabad* . . .) as features to resolve this issue. If its Chinese adjacent contextual word "地图" (*map*, pronounced as *di-tu*) could be aligned to the corresponding English word *map*, and given that "地图" and *map* are common nouns in their respective languages, it is possible to determine that this extra Chinese character "地" should not be linked to the ENE *Islamabad*.

Addressing these problems requires that both translation and transliteration models be more complex and must use additional features (possibly knowledge-rich features). Because this class accounts for only 3% of the errors, we leave the problem for future work.

5.1.1 Features Contributing to Boundary Errors. Among the 111 alignment errors analyzed, 76 of them[18] have their references covered by the expanded candidate set. The scores of their associated features are then further inspected to determine which features contribute to the errors. This is assessed by counting the number of times (denoted by *#Worst*) that a specific weighted feature-score gets the worst difference when those incorrect NE pairs are compared to their corresponding references. A large #Worst would imply that this feature should get more attention in pursuing further performance improvement.

The top four related statistics are given in Table 14, which indicates that F1 (Normalized TS/TL Transformation) is the most dominant feature in making those errors. Following this, F2 (Normalized Translation Mode), F7 (Normalized Chinese Bigram), and F10 (Normalized English Bigram) are on the second tier. Both F1 and F2 are related to alignment, which coincides with our observation that alignment-related Categories

18 This is the number of those entries under Categories (III), (IV), and (VI) in Table 12 after subtracting one (76 = 11 + 13 + 53 − 1), as one reference in Category (III) cannot be found after expansion.

Table 14
Top four worst-case statistics of features for NE boundary errors.

Features	#Worst
F1: $[\prod_{n=1}^{N} P(cp_{a(n)}\|M_n, ew_n, T)]^{1/N}$ (Normalized TS/TL Transformation)	17
F2: $[\prod_{n=1}^{N} P(M_n\|ew_n, T)]^{1/N}$ (Normalized Translation Mode)	10
F7: $[\prod_{l=1}^{L} P(cc_l\|cc_{l-1}, T)]^{1/L}$ (Normalized Chinese Bigram)	10
F10: $[\prod_{n=1}^{N} P(ew_n\|ew_{n-1}, T)]^{1/N}$(Normalized English Bigram)	9

(i.e., IV, VI.A, and VI.B.2) occupy the largest portion of errors (62% of 76 inspected errors). The errors dominated by F7 and F10 are further discussed as follows.

Among the ten errors dominated by F7, except for three (in which one is due to a spurious anchor, and two are due to abnormal transformations), all others selected the sub-strings of their corresponding CNE references. Furthermore, each selected sub-string included the Chinese bigrams that appear more frequently than those within the remaining sub-string (of its reference CNE). In other words, F7 tends to select only the portion with high frequency bigrams when all related components are aligned. For example, for the reference {⟨儿童拒菸运动中心⟩::[National Center for Tobacco-Free Kids]}, only {⟨中心⟩::[Center]} is extracted, as this Chinese bigram is frequently used in various organization names. Further inspection of those seven cases (which prefer the sub-string) reveals that three of them are with unmatched components (Category IV); therefore only four of them are, in fact, due to the problem of F7.

On the other hand, among the nine errors caused mainly by F10, only three of them chose the sub-strings of their corresponding ENE reference, and the remaining six errors selected the strings unrelated to the reference due to spurious anchors and an acronym. It therefore seems that different languages possess different error patterns.

The problem of preferring a more frequent sub-string cannot be solved by normalizing related bigrams. This is because the current bigram model does not consider the implied restriction on the context surrounding the given CNE. In other words, a given CNE also implies that its left and right adjacent characters should not be a part of CNE (or its left and right adjacent characters must be in a non-NE region). In our data set, two adjacent non-NE Chinese characters (or words for English) are found to be sufficient for both left and right contexts. Therefore, the following additional terms are further proposed to take care of this issue: $P(cc_{-1}|cc_0, T) \times P(cc_0|cc_1, T) \times P(cc_{L+2}|cc_{L+1}, T) \times P(cc_{L+1}|cc_L, T)$, where cc_1^L is the given CNE, cc_0 and cc_{L+1} are its left and right adjacent non-NE characters, respectively. This formula can be easily derived from $P(C|Cb, Tc, T, Sc)$, similar to Equation (8). The derivation also applies to English. To test this supposition, the related experiment (Exp4 [*MERT-W, N+Full_Model*] specified in Table 3) is updated as Experiment 5 with the probability features shown here.

Exp5: This experiment (named *MERT-W, N-Full_Model+Contextual-Bigram*, and denoted by *MERT-W-CB*) replaces $[\prod_{l=1}^{L} P(cc_l|cc_{l-1}, T)]^{\frac{1}{L}}$ in the original Exp4 with

$$P(cc_{-1}|cc_0, T) \cdot P(cc_0|cc_1, T) \cdot P(cc_{L+2}|cc_{L+1}, T) \cdot P(cc_{L+1}|cc_L, T) \cdot [\prod_{l=1}^{L} P(cc_l|cc_{l-1}, T)]^{\frac{1}{L}}$$

The same is done for English.

Table 15 shows the performance on the test set (data from Exp4 are also listed for comparison). The entries in bold indicate statistically significant improvements over

Table 15
Effect of adjacent contextual (non-NE) bigrams on the test set.

Model	P(%)	R(%)	F(%)
Exp4 (*MERT-W*)	85.9 (80.5)	88.4 (83.0)	87.1 (81.7)
Exp5 (*MERT-W-CB*): Add Contextual-Bigram to Original N-Bigram	**86.7 (81.7)**	**89.3 (84.1)**	**88.0 (83.0)**

their counterparts. Results show that the performance has indeed improved, and six of the targeted seven cases (four CNE and three ENE errors, as mentioned previously) have been corrected (the remaining error is due to data sparseness and cannot be corrected). According to coefficient weighting by MERT process, $P(cc_0|cc_1, T)$ is more important than $P(cc_{-1}|cc_0, T)$, indicating that the closer a non-NE character is to the given NE, the more influential it is. This observation confirms our intuition about the context effect. A similar trend is also observed for other contextual non-NE characters.

5.2 NE Type Errors

In addition to the 111 boundary errors just analyzed, there are also NE type errors. Among those 635 NE pairs with correct boundaries in the test set, there are 41 (6.5%) NE type errors in Exp4 (MERT-W). Among them, 175 PER, 248 LOC, and 212 ORG NE types are assigned. The associated confusion matrix of various NE types is shown in Table 16 (the numbers within the parentheses are the relative ratios of their output types). All except 5 of the 41 NE type errors originated from transliterated NE pairs (not shown in the table). This is consistent with our observation that even a human annotator finds it challenging to identify correct types for transliterated NE pairs in the absence of context.

Table 16 shows that PER has the highest error rate, LOC follows as the second, and ORG is a distant third. In addition, PER and LOC are the types that are most often confused with one another. These observations match the distribution of transliterations in each type (the transliteration mode ratios for PER, LOC, and ORG are 100%, 71.4%, and 25.2%, respectively), as it is very difficult to determine the type when a NE is transliterated without context.

To solve NE type errors originating from transliteration, the adjacent contextual non-NE characters are also helpful. For example, {〈梅尔斯〉::[Myers]} is incorrectly identified as LOC, when in fact it should have been PER in the context "主席梅尔斯" (*president Myers*). The left adjacent contextual bigrams "主席" (*president*) should indicate that the following NE is likely to be PER. Table 15 shows that an additional four type errors are also corrected apart from the six boundary errors. Therefore, in comparison with the original MERT-W, this new version (MERT-W-CB) gains more in type-sensitive

Table 16
Distribution of the NE type errors (MERT-W).

NE type	Reference PER Type	Reference LOC Type	Reference ORG Type
Output Type = PER	153 (87.4%)	16 (9.1%)	6 (3.4%)
Output Type = LOC	14 (5.6%)	232 (93.5%)	2 (0.8%)
Output Type = ORG	0 (0%)	3 (1.4%)	209 (98.6%)

F-score (an increase of 1.5%, from 81.5% to 83.0%) than in type-insensitive F-score (an increase of 0.9%, from 87.1% to 88.0%).

The adjacent contextual non-NE characters have only limited power in disambiguating NE types, however. For instance, the reference {⟨喀布尔⟩::[Kabul]} is incorrectly identified as PER, which should be LOC in the context "在喀布尔" (*at Kabul*). Because "喀布尔" follows a preposition "在" (*at*) in the associated context, it indicates that "喀布尔" is a location name. We can also easily find counterexamples, however, such as "在布希周围" (*around Bush*), in which "布希" (*Bush*) is PER, not LOC. In fact, both PER and LOC can be freely exchanged in this situation in either a Chinese or English context. Therefore, more complicated syntactic or semantic information is required in some cases involving transliteration.

Another interesting statistic not shown in Table 16 is the distribution of errors versus initial NE types (i.e., *Tc* and *Te*) assigned in the first stage. Among 41 errors, 22 (54%) have both incorrect *Tc* and *Te*, 16 (39%) have correct *Tc* but wrong *Te*, and only 3 (7%) have correct *Te* but incorrect *Tc*. This distribution shows that *Tc* is more reliably assigned in the first stage than is *Te*, which confirms the observation in Ji and Grishman (2006) that English NE type assignment is more challenging.

5.2.1 Features Contributing to Type Errors. Similarly, the study for #Worst on weighted feature-scores is also performed under MERT-W-CB (i.e., Exp5). There are 37 type errors with correct boundaries, and we have examined that F4 (NE Type Re-assignment) is the most dominant feature in making these errors. Because this feature $LogP(T|Tc, Te)$ (Equation [7]) always assigns an incorrect type when both *Tc* and *Te* are incorrect (11 out of 13 F4 errors belong to this class), it is not surprising that it is ranked at the top (22 [59%] out of a total of 37 errors have both incorrect *Tc* and *Te*). In addition, the feature $LogP(\delta_A|T)$ in Equation (6) always prefers PER (which are always completely transliterated in the corpus) when all components in an NE are transliterated (i.e., $\delta_A = 0$). Therefore, the feature is ranked second (31 out of 37 errors are complete transliterations, although only 10 of them should have been assigned PER). It is found that all nine cases in this category are not PER (only two of them are not transliterated), which further supports our analysis. Solving these type errors requires that these two features be conditioned on more features, and requires further study. Finally, the top four features in making type errors are related to alignment (Equations [6] and [7]), which indicates that monolingual lexicon information (both English and Chinese bigrams) is more reliable in deciding NE type.

6. Applications of the Proposed Model

It would be interesting to know how the proposed model performs in real applications. Because MERT-W performs best in our tests, it is adopted in this study on real applications. Section 6.1 presents the effectiveness of improving NE recognition, and Section 6.2 shows how the improved NE recognition can be used in learning a monolingual NE recognition model (also NE translation table/models) in a semi-supervised manner.

6.1 On Improving Monolingual NE Recognition

As explained in Section 2.1, the alignment result can also be used to refine the initially recognized NEs. The improvements that MERT-W made in refining the boundaries and NE types of those initially recognized Chinese/English NEs are shown in Tables 17 and 18, respectively. For comparison, the rows associated with the initial recognizers and the

Table 17
Type-insensitive improvement for Chinese/English NER.

NE type	Model	P (%)	R (%)	F (%)
PER	Initial	85.9/91.3	89.3/91.1	87.6/91.2
	Baseline	**88.1 (+2.2)/ 92.9 (+1.6)**	**89.8 (+0.5)/ 92.2 (+1.1)**	**89.0 (+1.4)/ 92.6 (+1.4)**
	MERT-W	**90.3 (+4.4)/ 94.5 (+3.2)**	**90.5 (+1.2)/ 93.0 (+1.9)**	**90.4 (+2.8)/ 93.8 (+2.6)**
LOC	Initial	90.9/93.6	91.4/93.4	91.1/93.5
	Baseline	**91.8 (+0.9)/ 94.7 (+1.1)**	91.8 (+0.4)/ **94.2 (+0.8)**	**91.8 (+0.7)/ 94.5 (+1.0)**
	MERT-W	**94.2 (+3.3)/ 95.8 (+2.2)**	**93.2 (+1.8)/ 95.2 (+1.8)**	**93.6 (+2.5)/ 95.5 (+2.0)**
ORG	Initial	81.1/88.9	83.9/87.7	82.5/88.3
	Baseline	**84.4 (+3.3)/ 90.7 (+1.8)**	**86.6 (+2.7)/ 89.2 (+1.5)**	**85.5 (+3.0)/ 90.0 (+1.7)**
	MERT-W	**86.8 (+5.7)/ 92.8 (+3.9)**	**88.7 (+4.8)/ 89.9 (+2.2)**	**87.8 (+5.3)/ 91.4 (+3.1)**
ALL	Initial	86.0/92.2	88.7/92.2	87.3/92.5
	Baseline	**88.8 (+2.8)/ 93.6 (+1.4)**	**89.6 (+0.9)/ 93.9 (+1.1)**	**89.2 (+1.9)/ 93.7 (+1.2)**
	MERT-W	**91.4 (+5.4)/ 95.2 (+3.0)**	**91.1 (+2.4)/ 94.7 (+1.9)**	**91.2 (+3.9)/ 94.9 (+2.4)**

alignment baseline systems are also given in the two tables (as before, the entries in bold indicate that differences are statistically significant). In addition, figures in parentheses indicate the corresponding differences in performance compared to the initial version (shown in Table 2).

Table 17 shows that both the baseline and MERT-W systems have significantly improved the initial NE recognition type-insensitive results for both Chinese and English. It also shows that MERT-W significantly outperforms the baseline. In particular, Chinese ORG is observed to yield the largest improvement among NE types in both Chinese and English, which matches our previous observations that the boundary of a Chinese ORG is difficult to identify using only the information from the Chinese sentence.

The type-sensitive results are given in Table 18, which shows that MERT-W also significantly improves the initial NE recognition results for both Chinese and English.

Table 18
Type-sensitive improvement for Chinese/English NER.

NE type	Model	P (%)	R (%)	F (%)
PER	Initial	80.2/79.2	87.7/85.3	83.8/82.1
	Baseline	79.4 (−0.8)/ 78.6 (−0.6)	88.1 (+0.4)/ **86.6 (+1.3)**	83.5 (−0.3)/ 82.4 (+0.3)
	MERT-W	**85.6 (+5.4)/ 85.6 (+6.4)**	**89.9 (+2.2)/ 87.9 (+2.6)**	**87.7 (+3.9)/ 86.7 (+4.6)**
LOC	Initial	89.8/85.9	87.3/81.5	88.5/83.6
	Baseline	90.5 (+0.7)/ 86.5 (+0.6)	83.6 (−3.7)/ 80.4 (−1.1)	86.9 (−1.6)/ 83.3 (−0.3)
	MERT-W	**93.8 (+4.0)/ 89.3 (+3.4)**	87.1 (−0.2)/ **84.1(+2.7)**	90.3 (+1.8)/ **86.6 (+3.0)**
ORG	Initial	78.6/82.9	82.8/79.6	80.6/81.2
	Baseline	79.5 (−0.9)/ 82.2 (−0.7)	80.9 (−1.9)/ 80.1 (+0.5)	80.2 (−0.4)/ 81.1 (−0.1)
	MERT-W	**85.6 (+7.0)/ 86.8 (+3.9)**	**88.4 (+5.6)/ 88.7 (+9.1)**	**87.0 (+6.4)/ 87.7 (+6.5)**
ALL	Initial	83.4/82.1	86.0/82.6	84.7/82.3
	Baseline	83.2 (−0.2)/ 82.2 (+0.1)	83.9 (−2.1)/ 82.3 (−0.3)	83.5 (−1.2)/ 82.2 (−0.1)
	MERT-W	**88.7 (+5.3)/ 87.3 (+5.2)**	**88.4 (+2.4)/ 86.6 (+4.0)**	**88.6 (+3.9)/ 86.9 (+4.6)**

Note that English ORG yields the largest gain among NE types in both Chinese and English, again supporting our earlier observation that an English ORG cannot be easily identified when only the English sentence is available. It must be noted that the baseline alignment model deteriorates the original NE recognition in overall performance (even though it can correct some NE initial boundary errors as shown in Table 17), because it does not use the features/constraints proposed in the joint model.

6.2 On Learning NE Recognition Models via Semi-Supervised Learning

In many NLP applications, the associated process will be considerably simplified if the included NEs can be identified first. Therefore, it is important to have a good NE recognizer that has been well trained. Various domains frequently have different sets of NEs, however, and new NEs also emerge over time. We thus need to periodically update the NE recognition model (also the NE translation table/model, if it is for MT), which necessitates the need to ensure short training times (including set-up time and human effort). This requirement can be addressed well in a semi-supervised learning set-up where parameters/tables are learned from a large unlabeled corpus with a small (albeit human) annotated seed set.

Under the semi-supervised learning framework, however, maximizing likelihood does not imply a minimizing of error rate at the same time. Without additional constraints, a monolingual NE model is usually unable to converge to the desired point in the parameter space. On the other hand, as shown in the previous section, the alignment module can further refine the initially recognized NEs with additional mapping constraints from the other language. The proposed joint model thus can be used to train the monolingual NE recognition model via semi-supervised learning on a large unlabeled bilingual corpus. In other words, when semi-supervised learning is conducted for learning the NE recognition model, MERT-W is expected to guide the search process for convergence towards the human annotation. This advantage is important for regularly updating the NE recognition and translation table/models.

We now outline our semi-supervised learning procedure as follows.

(1) A small pre-labeled corpus acts as seed data. Based on these seed data, train the Chinese/English NE recognition toolkit (described in Section 4.1) and the adopted NE alignment model.

(2) Perform the Chinese/English NE recognition toolkits and the NE alignment model, trained in Step (1), on a large unlabeled (sentence-aligned) bilingual corpus to report those NE pairs that they identify.

(3) Denote the located NE pairs as correctly labeled and combine them with the seed data as new labeled training data.

(4) Re-train the Chinese/English NE recognition toolkit and also re-train the NE alignment model on the newly labeled training data.

(5) Repeat Steps (2) through (4) until convergence. The final Chinese/English NE recognizer and the NE alignment model are compared to their initial versions.

Because the adopted Chinese NE recognizer (Wu, Zhao, and Xu 2005) cannot be re-trained (because it is not an open source toolkit), only the English NE recognizer and

Table 19
English NE recognition on test data after semi-supervised learning.

Model Seed Size (sentence pairs)	100	400	4,000	40,000
Initial-NER	36.7 (0%)	58.6 (0%)	71.4 (0%)	79.1 (0%)
NER-Only	−2.3 (−6.3%)	−0.5 (−0.8%)	−0.3 (−0.4%)	−0.1 (−0.1%)
NER+Alignment-Baseline	**+4.9** (**+13.4%**)	**+3.4** (**+5.8%**)	**+1.7** (**+2.4%**)	+0.7 (+0.9%)
NER+Alignment-Weighted	**+10.7** (**+29.2%**)	**+8.7** (**+14.8%**)	**+4.8** (**+6.7%**)	**+2.3** (**+2.9%**)

the alignment model are updated during training iterations. In our experiments, 50,412 sentence pairs are first extracted from Training Set I as unlabeled data. Various labeled data sets are then extracted from the remaining data as seed corpora with different sizes (100, 400, 4,000, and 40,000 sentence pairs). The test set is still the same 300 sentence pairs that were adopted previously.

Table 19 shows the type-sensitive F-score of English NER on the test set at convergence.[19] The *Initial-NER* in Table 19 indicates the initial performance of the NER model re-trained from different seed corpora. Three different approaches are tested: (1) English NER only (*NER-Only*), (2) the alignment baseline model (*NER+Alignment-Baseline*), and (3) our weighted joint model MERT-W (*NER+Alignment-Weighted*). The first case performs semi-supervised learning only on English data without involving alignment, whereas the last two cases include alignment, and both the English NER model and the alignment model are re-trained during iterations. The numbers in parentheses show relative improvements over *Initial-NER*. The entries in bold indicate statistically significant improvements over *Initial-NER*.

As Table 19 shows, using the NER model alone, the performance may drop after convergence. This is because maximizing likelihood does not imply minimizing the error rate. With additional mapping constraints from the other language, however, the alignment module can guide the search process to converge to a more desirable point in the parameter space. It must be noted here that the contribution of additional constraints increases with smaller seed corpora, because constraints become more important when the labeled data set is smaller.

Table 20 shows only the type-sensitive F-score of NE pair alignment on the test set before and after convergence due to space constraints. Two different models are tested: (1) the alignment baseline model (*NER+Alignment-Baseline*), and (2) our weighted joint model (*NER+Alignment-Joint*). The data in parentheses indicate relative improvements over the performance before training. The entries in bold indicate statistically significant improvements over the model before training.

As Table 20 demonstrates, the alignment performance can also be improved with semi-supervised learning. Note that the improvement is greater on smaller data sets, which is common in most semi-supervised learning tasks.

19 The iteration process will stop when the last two consecutive iterations share more than 99.9% of their output.

Table 20
NE alignment on test data after semi-supervised learning.

Model Seed Size (sentence pairs)	100	400	4,000	40,000
NER+Alignment-Baseline	33.5 (0%)	50.2 (0%)	62.6 (0%)	67.3 (0%)
NER+Alignment-Baseline (After)	**+4.2 (+12.5%)**	**+3.0 (+6.0%)**	**+1.8 (+2.9%)**	**+1.2 (+1.8%)**
NER+Alignment-Joint	38.1 (0%)	56.3 (0%)	67.7 (0%)	72.5 (0%)
NER+Alignment-Joint (After)	**+10.1 (+26.5%)**	**+8.7 (+15.5%)**	**+6.0 (+8.9%)**	**+3.3 (+4.6%)**

As shown by Tables 19 and 20, the proposed joint model gains more from the learning process in comparison with the alignment-baseline model, because information from the aligned sentence is utilized more effectively. Results demonstrate that the proposed joint model, combined with semi-supervised learning, offers significant improvement for semi-automatically updating the NE recognition model and the NE translation table. Additionally, the impact is greater when less time is available for labeling seed data.

7. Related Work

There is significant work on identifying NEs within monolingual texts across languages, such as English (Chinchor 1998; Mikheev, Grover, and Moens 1998; Borthwick 1999) and Chinese (Chen et al. 1998a; Sun, Zhou, and Gao 2003), to name a few. Various approaches to identifying NEs have also been proposed, such as hidden Markov models (Bikel et al. 1997; Bikel, Schwartz, and Weischedel 1999), conditional random fields (McCallum and Li 2003; Jiao et al. 2006), modified transformation-based learning (Black and Vasilakopoulos 2002), boosting (Collins 2002; Wu et al. 2002), AdaBoost (Carreras, Marquez, and Padro 2002), and adopting semi-supervised learning (Wong and Ng 2007; Liao and Veeramachaneni 2009). Furthermore, features including local information (e.g., token, part-of-speech) and global information (e.g., label consistency, context features) from monolingual resources have been adopted (Krishman and Manning 2006; Zhou and Su 2006). In prior work on the use of bilingual NE alignment for NE recognition, Huang and Vogel (2004) used an iterative process to extract a smaller but cleaner NE translation dictionary and then used the dictionary to improve the monolingual NE annotation quality. Ji and Grishman (2007) adopted several heuristic rules for using bilingual-text information to correct NE recognition errors.

In aligning bilingual NEs from two given NE lists, the NE translation model is usually adopted. Typically, an NE is either transliterated or semantically translated. For transliteration, Knight and Graehl (1998) were pioneers in adopting the probabilistic model to align the components within an NE pair. Since then, similar approaches have been applied to various language pairs such as English/Arabic (Stalls and Knight 1998), English/Chinese (Chen et al. 1998b; Wan and Verspoor 1998; Lin and Chen 2002; Lee and Chang 2003; Lee, Chang, and Jang 2003; Gao, Wong, and Lam 2004;

Pervouchine, Li, and Lin 2009), English/Japanese (Knight and Graehl 1998; Tsuji 2002), and English/Korean (Lee and Choi 1997; Oh and Choi 2002, 2005). Moreover, Li, Zhang, and Su (2004), and Li et al. (2007) presented a joint source channel model for transliteration, and automated the semantic transliteration process, which takes origin and gender into account for personal names.

In contrast, research on automatic NE semantic translation is less common. Zhang et al. (2005) proposed a phrase-based context-dependent joint probability model for semantic translation, which is similar to phrase-level translation models in statistical MT (Zong and Seligman 2005; Hu, Zong, and Xu 2006). Chen, Yang, and Lin (2003) and Chen et al. (2006) studied formulation and transformation rules for English–Chinese NEs. They adopted a frequency-based approach for extracting key words of NEs with or without dictionary assistance and constructed transformation rules from the bilingual NE corpus. Their studies focused on transformation rules with particular attention to distinguishing translated parts from transliterated parts; the performance of rule-application in NE translation was not described, however. Chen and Zong (2008) proposed a chunk-based probabilistic translation model for organization names, although, its application to person and location names has not been studied.

Because new NEs emerge from time to time and can be transformed in various ways (translation, transliteration, or other abnormal types), the NE transliteration/translation models mentioned usually lead to unsatisfactory results, especially for infrequently occurring NEs. Recent studies have therefore focused on extracting new NE pairs from either bilingual corpora or Web resources, so that the corresponding human translation can be directly adopted (or used for training). To do this, however, NE alignment is an essential tool. Due to the relatively poor quality of Web data, such alignment approaches are usually limited unless significant effort is devoted to data cleaning; therefore, we do not discuss these approaches.

To extract the NE pairs from bilingual corpora, all NE alignment approaches we found in the literature are conducted after an initial NE identification stage so that the complexity of the task can be reduced. The associated cost is that those initial NE recognition errors propagate into the following alignment stage. Both symmetric (which first identifies NEs in both languages) and asymmetric (which first identifies NEs in only one language) strategies have been proposed to mitigate this problem (Moore 2003), and are described here.

For the symmetric strategy, Huang and Vogel (2004) proposed to extract the NE translation dictionary from the bilingual corpus, and then used it to improve the NE annotation performance iteratively. Huang, Vogel, and Waibel (2003) described a multi-feature NE alignment model to extract NE equivalences (with translation, transliteration, and tagging features), from which a NE translation dictionary was then constructed. Kumano et al. (2004) proposed a method to extract English–Chinese NE pairs from a content-aligned corpus. This approach tries to find the correspondences between bilingual NE groups based on the similarity in their order of appearance in each document. Additionally, an abridged version of our work has been presented in our ACL-10 paper (Chen, Zong, and Su 2010). Among those symmetric approaches, only Huang, Vogel, and Waibel and Chen, Zong, and Su adopt the expansion strategy, described below.

For the asymmetric strategy, Al-Onaizan and Knight (2002) proposed an algorithm to translate NEs from Arabic to English using monolingual and bilingual resources. Given an Arabic NE, they used transliteration models (including a phonetic-based and a spelling-based model), a bilingual dictionary, and an English news corpus to first generate a list of English candidates, which were then re-scored by a Web resource.

Moore (2003) developed an approach to learning phrase translations from a parallel corpus based on a sequence of cost models. A maximum entropy model for NE alignment was presented in Feng, Lv, and Zhou (2004). Lee, Chang, and Jang (2006) proposed to align bilingual NEs in a bilingual corpus by incorporating a statistical model with multiple sources. Turning to comparable corpora, Shao and Ng (2004) presented a hybrid method to mine new translations from Chinese–English comparable corpora, combining both transliteration and context information. Sproat, Tao, and Zhai (2006) investigated the Chinese–English NE transliteration equivalence within comparable corpora.

Although these asymmetry strategies can prevent NE recognition errors on the target side from affecting alignment, errors on the source side continue to propagate to later stages. To reduce error propagation from both the source and the target, Huang, Vogel, and Waibel (2003) proposed to first identify the NEs in both the source and target, and then enlarge the obtained NE candidate sets for both languages before conducting alignment. Based on the observation that NE boundaries are frequently identified incorrectly, the enlarging procedure is done by treating the original recognition results as anchors and then increasing the number of candidates by expanding or shrinking the boundaries of those originally recognized NEs in both languages.

Our approach also adopts the expansion strategy. It differs from the works of Huang et al. (2003) and others in several ways, however. First, in all the alignment papers mentioned here, the adopted probabilities are directly used as features for log-linear combination or ME training without derivation. In contrast, our work fully derives a probabilistic joint model, for both identification and alignment, in a principled way. Second, unlike previous approaches that discard the information of initially identified NE anchors after the anchors have been expanded, our approach uses this information in the final selection process. Third, we propose new features, such as translation mode and its ratio, boundary shifting distance, and contextual bigrams. Fourth, we introduce a normalization step that removes the systematic bias preferring shorter NEs. Fifth, the effect of each individual feature, the influence of adopting different NE recognizers, the effectiveness across different domains, the effect of using a derived model (compared to ME), and the effect of the alignment model in semi-supervised learning are studied. Finally, the causes of alignment errors and type re-assignment errors are extensively investigated and categorized.

8. Conclusion

This article develops a novel and principled model for jointly conducting NE recognition and alignment. To the best of our knowledge, this is the first work that formally captures the interactions between NE recognition and NE alignment. The joint model not only greatly improves NE alignment performance, but also significantly boosts NE recognition performance.

Our experiments show that the new NE likelihoods are more effective than the bigram model used in the baseline system. Moreover, both the translation mode ratio and the entity type consistency constraint are critical in identifying the associated NE boundaries and types, as evidenced by the 21.3% relative improvement on type-sensitive F-score (from 68.4% to 83.0%) in our Chinese–English NE alignment task. The superiority of the proposed model has been shown to hold over the various domains tested.

Furthermore, the joint alignment model can also be used to refine the initially recognized NEs. This is achieved by utilizing additional mapping information from the

other language. In our experiments, when semi-supervised learning is conducted to train the adopted English NE model (with only 100 seed sentence pairs), the proposed model greatly boosts the English NE recognition type-sensitive F-score from 36.7% to 47.4% (29.2% relative improvement) in the test set.

Finally, the proposed model does not utilize language-dependent features. For example, Chinese characters and English words adopted in the model are visible units in the given languages, and no language-dependent features, such as morpheme/part-of-speech (or prefix/suffix), are used. In addition, the model does not use linguistic rules or tree banks. Therefore, although our experiments are conducted on Chinese–English language pairs, it is expected that the proposed approach can be applied to other language pairs with little adaptation effort.

Acknowledgments

This research has been funded by the Natural Science Foundation of China under grant nos. 61003160 and 60975053 and supported by the Hi-Tech Research and Development Program ("863" Program) of China under grant no. 2011AA01A207. Thanks are also given to the authors' associate, Tao Zhuang, for his great help on the publication version.

References

Al-Onaizan, Yaser and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 400–408, Philadelphia, PA.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bikel, Daniel M., Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC.

Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1–3):211–231.

Black, W. J. and Argyrios Vasilakopoulos. 2002. Language independent named entity classification by modified transformation-based learning and by decision tree induction. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 159–162, Taipei.

Borthwick, A. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Brown, Perer F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Carreras, X., L. Marquez, and L. Padro. 2002. Named entity extraction using adaboost. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 167–170, Taipei.

Chen, Hsin-His, Yung-Wei Ding, Shih-Chung Tsai, and Guo-Wei Bian. 1998a. Description of the NTU system used for met2. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, pages 121–129, Fairfax, VA.

Chen, Hsin-His, S.-J. Huang, Y.-W. Ding, and S.-C. Tsai. 1998b. Proper name translation in cross-language information retrieval. In *Proceedings of the 17th COLING and 36th ACL Conference*, pages 232–236, Montreal.

Chen, Hsin-His, W.-C. Lin, C. Yang, and W.-H. Lin. 2006. Translating/transliterating named entities for multilingual information access. *Journal of the American Society for Information Science and Technology (Special Issue on Multilingual Information Systems)*, 57(5):645–659.

Chen, Hsin-His, Changhua Yang, and Ying Lin. 2003. Learning formulation and transformation rules for multilingual named entities. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 1–8, Sapporo.

Chen, Stanley F. and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA.

Chen, Yufeng and Chengqing Zong. 2008. A structure-based model for Chinese organization name translation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(1):1–30.

Chen, Yufeng, Chengqing Zong, and Keh-Yih Su. 2010. On jointly recognizing and aligning bilingual named entities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 631–639, Uppsala.

Chinchor, Nancy. 1998. Overview of muc-7/met-2. In *Proceedings of Message Understanding Conference MUC-7*, pages 1–4, Fairfax, VA.

Cohen, William W. 2004. *MinorThird: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data*. Available at `http://minorthird.sourceforge.net`.

Collins, Michael. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Philadelphia, PA.

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 372–379, Barcelona.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor, MI.

Freedman, D. A. 2005. *Statistical Models: Theory and Practice*. Cambridge University Press.

Gao, Jianfeng, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574.

Gao, Wei, Kam-Fam Wong, and Wai Lam. 2004. Transliteration of foreign names for OOV problem. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 110–119, Sanya.

Hu, Rile, Chengqing Zong, and Bo Xu. 2006. An approach to automatic acquisition of translation templates based on phrase structure extraction and alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1656–1663.

Huang, Fei and Stephan Vogel. 2004. Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interface*, pages 253–258, Pittsburgh, PA.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of ACL'03, Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 9–16, Sapporo.

Ji, Heng and Ralph Grishman. 2006. Analysis and repair of name tagger errors. In *Proceedings of COLING/ACL 2006*, pages 420–427, Sydney.

Ji, Heng and Ralph Grishman. 2007. Collaborative entity extraction and translation. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 73–84, Borovets.

Jiao, Feng, Shaojun Wang, Chi H. Lee, Russell Greiner, and Dale Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 209–216, Sydney.

Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184, Detroit, MI.

Knight, Kevin and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Krishman, Vijay and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of 44th Annual Meeting of the Association for Computational Linguistics*, pages 1,121–1,128, Sydney.

Kumano, T., H. Kashioka, H. Tanaka, and T. Fukusima. 2004. Acquiring bilingual named entity translations from content-aligned corpora. In *Proceedings of the First International Joint Conference on Natural Language Processing*, pages 177–186, Hainan Island.

Lee, Chun-Jen and Jason S. Chang. 2003. Acquisition of English-Chinese transliterated word pairs from parallel aligned texts using a statistical machine transliteration model. In *Proceedings of HLT-NAACL 2003*

Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, pages 96–103, Edmonton.

Lee, Chun-Jen, Jason S. Chang, and Jyh-Shing R. Jang. 2003. A statistical approach to Chinese-to-English back transliteration. In *Proceedings of the 17th Pacific Asia Conference on Language, Information, and Computation*, pages 310–318, Singapore.

Lee, Chun-Jen, Jason S. Chang, and Jyh-Shing R. Jang. 2006. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):121–145.

Lee, Jae Sung and Key-Sun Choi. 1997. A statistical method to generate various foreign word transliterations in multilingual information retrieval system. In *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages (IRAL)*, pages 123–128, Tsukuba.

Li, Haizhou, Khe Chai Sim, Jin Shea Kuo, and Minghui Dong. 2007. Semantic transliteration of personal names. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 120–127, Prague.

Li, Haizhou, Min Zhang, and Jian Su. 2004. A joint source channel model for machine transliteration. In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 159–166, Barcelona.

Liao, Wenhui and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65, Boulder, CO.

Liese, Friedrich and Klaus-J. Miescke. 2008. *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer, Berlin.

Lin, Wei-Hao and Hsin-Hsi Chen. 2002. Backward transliteration by learning phonetic similarity. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 139–145, Taipei.

McCallum, Andrew and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2003)*, pages 188–191, Edmonton.

McCallum, Andrew Kachites. 2002. *MALLET: A Machine Learning for Language Toolkit*. Available at `http://mallet.cs.umass.edu`.

Mikheev, A., C. Grover, and M. Moens. 1998. Description of the LTG system used for MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, pages 1–12, Fairfax, VA.

Moore, R. C. 2003. Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the 10th Conference of the European Chapter of ACL*, pages 259–266, Budapest.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Oh, Jong-Hoon and Key-Sun Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 758–764, Taipei.

Oh, Jong-Hoon and Key-Sun Choi. 2005. An ensemble of grapheme and phoneme for machine transliteration. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 450–461, Jeju Island.

Pervouchine, Vladimir, Haizhou Li, and Bo Lin. 2009. Transliteration alignment. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 136–144, Singapore.

Schlüter, Ralf and Hermann Ney. 2001. Model-based MCE bound to the true Bayes error. *IEEE Signal Processing Letters*, 8(5):131–133.

Shao, Li and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 618–624, Geneva.

Sproat, Richard, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 73–80, Sydney.

Stalls, B. G. and Kevin Knight. 1998. Translating names and technical

terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41, Montreal.

Stolcke, Andreas. 2002. SRILM—An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Sun, Jian, Ming Zhou, and Jianfeng Gao. 2003. A class-based language model approach to Chinese named entity identification. *Computational Linguistics and Chinese Language Processing*, 8(2):1–28.

Tsuji, Keita. 2002. Automatic extraction of translational Japanese-Katakana and English word pairs from bilingual corpora. *International Journal of Computer Processing of Oriental Languages*, 15(3):261–279.

Wan, S. and C. M. Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proceedings of the 17th COLING and 36th ACL*, pages 1,352–1,356, Montreal.

Wong, Yingchuan and Hwee Tou Ng. 2007. One class per named entity: Exploiting unlabeled text for named entity recognition. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 1,763–1,768, Hyderabad.

Wu, D., G. Ngai, M. Carpuat, J. Larsen, and Y. Yang. 2002. Boosting for named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 195–198, Taipei.

Wu, Youzheng, Jun Zhao, and Bo Xu. 2005. Chinese named entity recognition model based on multiple features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 427–434, Vancouver.

Zhang, Min, Haizhou Li, Jian Su, and Hendra Setiawan. 2005. A phrase-based context-dependent joint probability model for named entity translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 600–611, Jeju Island.

Zhang, Ying, S. Vogel, and A. Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 2,051–2,054, Lisbon.

Zhao, Hai and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, pages 106–111, Hyderabad.

Zhou, GuoDong and Jian Su. 2006. Machine learning-based named entity recognition via effective integration of various evidences. *Natural Language Engineering*, 11(2):189–206.

Zong, Chengqing and Mark Seligman. 2005. Toward practical spoken language translation. *Machine Translation*, 19(2):113–137.