On Jointly Recognizing and Aligning Bilingual Named Entities

Yufeng Chen, Chengqing Zong
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{chenyf,cqzong}@nlpr.ia.ac.cn

Keh-Yih Su
Behavior Design Corporation
Hsinchu, Taiwan, R.O.C.
bdc.kysu@gmail.com

Abstract

We observe that (1) how a given named entity (NE) is translated (i.e., either semantically or phonetically) depends greatly on its associated entity type, and (2) entities within an aligned pair should share the same type. Also, (3) those initially detected NEs are anchors, whose information should be used to give certainty scores when selecting candidates. From this basis, an integrated model is thus proposed in this paper to jointly identify and align bilingual named entities between Chinese and English. It adopts a new mapping type ratio feature (which is the proportion of NE internal tokens that are semantically translated), enforces an entity type consistency constraint, and utilizes additional monolingual candidate certainty factors (based on those NE anchors). The experiments show that this novel approach has substantially raised the type-sensitive F-score of identified NE-pairs from 68.4% to 81.7% (42.1% F-score imperfection reduction) in our Chinese-English NE alignment task.

1 Introduction

In trans-lingual language processing tasks, such as machine translation and cross-lingual information retrieval, *named entity* (NE) translation is essential. Bilingual NE alignment, which links source NEs and target NEs, is the first step to train the NE translation model.

Since NE alignment can only be conducted after its associated NEs have first been identified, the including-rate of the first recognition stage significantly limits the final alignment performance. To alleviate the above error accumulation problem, two strategies have been proposed in the literature. The first strategy (Al-Onaizan and Knight, 2002; Moore, 2003; Feng et al., 2004; Lee et al., 2006) identifies NEs only on the source side and then finds their corresponding NEs on the target side. In this way, it avoids the NE recognition errors which would otherwise be brought into the alignment stage from the target side; however, the NE errors from the source side still remain.

To further reduce the errors from the source side, the second strategy (Huang et al., 2003) expands the NE candidate-sets in both languages before conducting the alignment, which is done by treating the original results as anchors, and then re-generating further candidates by enlarging or shrinking those anchors' boundaries. Of course, this strategy will be in vain if the NE anchor is missed in the initial detection stage. In our data-set, this strategy significantly raises the NE-pair type-insensitive including-rate [1] from 83.9% to 96.1%, and is thus adopted in this paper.

Although the above expansion strategy has substantially alleviated the error accumulation problem, the final alignment accuracy is still not good (type-sensitive F-score only 68.4%, as indicated in Table 2 in Section 4.2). After having examined the data, we found that: (1) How a given NE is translated, either semantically (called *translation*) or phonetically (called *transliteration*), depends greatly on its associated entity type[2]. The *mapping type ratio*, which is the percentage of NE internal tokens which are translated semantically, can help with the recognition of the associated NE type; (2) Entities within an aligned pair should share the same type, and this restriction should be integrated into NE alignment as a constraint; (3) Those initially identified monolingual NEs can act as anchors to give *monolingual candidate certainty scores*

[1] Which is the percentage of desired NE-pairs that are included in the expanded set, and is the upper bound on NE alignment performance (regardless of NE types).

[2] The proportions of semantic translation, which denote the ratios of semantically translated words among all the associated NE words, for person names (PER), location names (LOC), and organization names (ORG) approximates 0%, 28.6%, and 74.8% respectively in Chinese-English name entity list (2005T34) released by the Linguistic Data Consortium (LDC). Since the title, such as "sir" and "chairman", is not considered as a part of person names in this corpus, PERs are all transliterated there.

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 631–639,
Uppsala, Sweden, 11-16 July 2010.

(preference weightings) for the re-generated candidates.

Based on the above observation, a new joint model which adopts the *mapping type ratio*, enforces the *entity type consistency* constraint, and also utilizes the *monolingual candidate certainty factors* is proposed in this paper to jointly identify and align bilingual NEs under an integrated framework. This framework is decomposed into three subtasks: *Initial Detection*, *Expansion*, and *Alignment&Re-identification*. The *Initial Detection* subtask first locates the initial NEs and their associated NE types inside both the Chinese and English sides. Afterwards, the *Expansion* subtask re-generates the candidate-sets in both languages to recover those initial NE recognition errors. Finally, the *Alignment&Re-identification* subtask *jointly* recognizes and aligns bilingual NEs via the proposed joint model presented in Section 3. With this new approach, 41.8% imperfection reduction in type-sensitive F-score, from 68.4% to 81.6%, has been observed in our Chinese-English NE alignment task.

2 Motivation

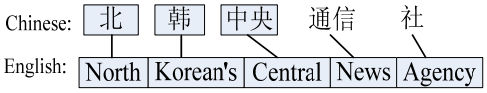
The problem of NE recognition requires both boundary identification and type classification. However, the complexity of these tasks varies with different languages. For example, Chinese NE boundaries are especially difficult to identify because Chinese is not a tokenized language. In contrast, English NE boundaries are easier to identify due to capitalization clues. On the other hand, classification of English NE types can be more challenging (Ji et al., 2006). Since alignment would force the linked NE pair to share the same semantic meaning, the NE that is more reliably identified in one language can be used to ensure its counterpart in another language. This benefits both the NE boundary identification and type classification processes, and it hints that alignment can help to re-identify those initially recognized NEs which had been less reliable.

As shown in the following example, although the desired NE "北韩中央通信社" is recognized partially as "北韩中央" in the initial recognition stage, it would be more preferred if its English counterpart "North Korean's Central News Agency" is given. The reason for this is that "News Agency" would prefer to be linked to "通信社", rather than to be deleted (which would happen if "北韩中央" is chosen as the corresponding Chinese NE).

(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's Central News Agency </ORG> quoted the navy's statement…

(III) The word alignment between two NEs:



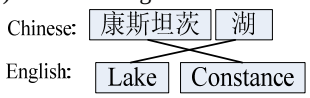
(VI) The re-identified Chinese NE boundary after alignment:
官方的 <ORG>北韩中央通信社</ORG> 引述海军声明...

As another example, the word "lake" in the English NE is linked to the Chinese character "湖" as illustrated below, and this mapping is found to be a translation and not a transliteration. Since translation rarely occurs for personal names (Chen et al., 2003), the desired NE type "LOC" would be preferred to 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, and the necessity of using mapping type ratio and NE type consistency constraint becomes evident.

(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>…

3 The Proposed Model

As mentioned in the introduction section, given a Chinese-English sentence-pair (CS, ES), with its initially recognized Chinese NEs $< CNE_i, CType_i >_{i=1}^{S}, S \geq 1$ and English NEs $[ENE_j, EType_j]_{j=1}^{T}, T \geq 1$ ($CType_i$ and $Etype_j$ are original NE types assigned to CNE_i and ENE_j, respectively), we will first re-generate two NE candidate-sets from them by enlarging and shrinking the boundaries of those initially recognized NEs. Let $RCNE_1^{K_C}$ and $RENE_1^{K_E}$ denote these two *re-generated* candidate sets for Chinese and English NEs respectively (K_C and K_E are their set-sizes), and $K = \min(S, T)$, then a total K pairs of final Chinese and English NEs will be picked up from the Cartesian product of

$RCNE_1^{K_C}$ and $RENE_1^{K_E}$, according to their associated *linking score*, which is defined as follows.

Let $Score(RCNE_{<k>}, RENE_{[k]})$ denote the associated *linking score* for a given candidate-pair $RCNE_{<k>}$ and $RENE_{[k]}$, where $<k>$ and $[k]$ are the associated indexes of the re-generated Chinese and English NE candidates, respectively. Furthermore, let $RType_k$ be the NE type to be *re-assigned* and shared by $RCNE_{<k>}$ and $RENE_{[k]}$ (as they possess the same meaning). Assume that $RCNE_{<k>}$ and $RENE_{[k]}$ are derived from initially recognized CNE_i and ENE_j, respectively, and M_{IC} denotes their *internal component mapping*, to be defined in Section 3.1, then $Score(RCNE_{<k>}, RENE_{[k]})$ is defined as follows:

$$Score(RCNE_{<k>}, RENE_{[k]}) = \max_{M_{IC}, RType_k} P\begin{pmatrix} M_{IC}, RType_k, RCNE_{<k>}, RENE_{[k]} \mid \\ <CNE_i, CType_i>, CS, [ENE_j, EType_j], ES \end{pmatrix} \quad (1)$$

Here, the "max" operator varies over each possible internal component mapping M_{IC} and re-assigned type (PER, LOC, and ORG). For brevity, we will drop those associated subscripts from now on, if there is no confusion.

The associated probability factors in the above linking score can be further derived as follows.

$$\begin{aligned} & P\left(M_{IC}, RType, RCNE, RENE \,\middle|\, \begin{matrix} <CNE, CType>, CS, \\ [ENE, EType], ES \end{matrix} \right) \\ & \cong P(M_{IC} \mid RType, RCNE, RENE) \\ & \quad \times P(RCNE \mid CNE, CType, CS, RType) \\ & \quad \times P(RENE \mid ENE, EType, ES, RType) \\ & \quad \times P(RType \mid CNE, ENE, CType, EType) \end{aligned} \quad (2)$$

In the above equation, $P(M_{IC} \mid RType, RCNE, RENE)$ and $P(RType \mid CNE, ENE, CType, EType)$ are the *Bilingual Alignment Factor* and the *Bilingual Type Re-assignment Factor* respectively, to represent the bilingual related scores (Section 3.1). Also, $P(RCNE \mid CNE, CType, CS, RType)$ and $P(RENE \mid ENE, EType, ES, RType)$ are *Monolingual Candidate Certainty Factors* (Section 3.2) used to assign preference to each selected *RCNE* and *RENE*, based on the initially recognized NEs (which act as anchors).

3.1 Bilingual Related Factors

The *bilingual alignment factor* mainly represents the likelihood value of a specific internal component mapping M_{IC}, given a pair of possible NE configurations *RCNE* and *RENE* and their associated *RType*. Since Chinese word segmentation is problematic, especially for transliterated words, the bilingual alignment factor $P(M_{IC} \mid RType, RCNE, RENE)$ in Eq (2) is derived to be conditioned on *RENE* (i.e., starting from the English part).

We define the internal component mapping M_{IC} to be $M_{IC} \equiv <[cpn_{<n>}, ew_{[n]}, Mtype_n]_{n=1}^{N}, \delta>$, where $[cpn_{<n>}, ew_{[n]}, Mtype_n]$ denotes a linked pair consisting of a *Chinese component* $cpn_{<n>}$ (which might contain several Chinese characters) and an *English word* $ew_{[n]}$ within *RCNE* and *RENE* respectively, with their *internal mapping type* $Mtype_n$ to be either *translation* (abbreviated as *TS*) or *transliteration* (abbreviated as *TL*). In total, there are N component mappings, with N_{TS} translation mappings $[cpn_{<n_1>}, ew_{[n_1]}, TS]_{n_1=1}^{N_{TS}}$ and N_{TL} transliteration mappings $[cpn_{<n_2>}, ew_{[n_2]}, TL]_{n_2=1}^{N_{TL}}$, so that $N = N_{TS} + N_{TL}$.

Moreover, since the mapping type distributions of various NE types deviate greatly from one another, as illustrated in the second footnote, the associated *mapping type ratio* $\delta = (N_{TS} / N)$ is thus an important feature, and is included in the internal component mapping configuration specified above. For example, the M_{IC} between "康斯坦茨湖" and "Constance Lake" is [康斯坦茨, *Constance*, *TL*] and [湖, *Lake*, *TS*], so its associated mapping type ratio will be "0.5" (i.e., 1/2). Therefore, the internal mapping $P(M_{IC} \mid RType, RENE)$ is further deduced by introducing the internal mapping type $Mtype_n$ and the mapping type ratio δ as follows:

$$\begin{aligned} & P(M_{IC} \mid RType, RENE) \\ & \equiv P([cpn_{<n>}, ew_{[n]}, Mtype_n]_{n=1}^{N}, \delta \mid RType, RENE) \\ & \approx \prod_{n=1}^{N} \begin{bmatrix} P(cpn_{<n>} \mid Mtype_n, ew_{[n]}, RType) \\ \times P(Mtype_n \mid ew_{[n]}, RType) \end{bmatrix} \\ & \quad \times P(\delta \mid RType) \end{aligned} \quad (3)$$

In the above equation, the mappings between internal components are trained from the syllable/word alignment of NE pairs of different NE types. In more detail, for transliteration, the model adopted in (Huang et al., 2003), which first Romanizes Chinese characters and then transliterates them into English characters, is

used for $P(cpn_{<n>} | TL_n, ew_{[n]}, RType)$. For translation, conditional probability is directly used for $P(cpn_{<n>} | TS_n, ew_{[n]}, RType)$.

Lastly, the *bilingual type re-assignment factor* $P(RType | CNE, ENE, CType, EType)$ proposed in Eq (2) is derived as follows:

$$P(RType | RCNE, RENE, CType, EType) \cong P(RType | CType, EType) \quad (4)$$

As Eq (4) shows, both the Chinese initial NE type and English initial NE type are adopted to jointly identify their shared NE type *RType*.

3.2 Monolingual Candidate Certainty Factors

On the other hand, the *monolingual candidate certainty factors* in Eq (2) indicate the likelihood that a re-generated NE candidate is the true NE given its originally detected NE. For Chinese, it is derived as follows:

$$\begin{aligned} & P(RCNE | CNE, CType, CS, RType) \\ & \cong P(LeftD, RightD, Str[RCNE] | Len_C, CType, RType) \\ & \approx P(LeftD | Len_C, CType, RType) \\ & \times P(RightD | Len_C, CType, RType) \\ & \times \prod_{m=1}^{M} P(cc_m | cc_{m-1}, RType) \end{aligned} \quad (5)$$

Where, the subscript *C* denotes Chinese, and Len_C is the *length* of the originally recognized Chinese NE *CNE*. *LeftD* and *RightD* denote the *left and right distance* (which are the numbers of Chinese characters) that *RCNE* shrinks/enlarges from the left and right boundary of its anchor *CNE*, respectively. As in the above example, assume that *CNE* and *RCNE* are "北韩中央" and "韩中央通信社" respectively, *LeftD* and *RightD* will be "-1" and "+3". Also, *Str*[*RCNE*] stands for the associated Chinese string of *RCNE*, cc_m denotes the *m-th Chinese character* within that string, and *M* denotes the total number of Chinese characters within *RCNE*.

On the English side, following Eq (5), $P(RENE | ENE, EType, ES, RType)$ can be derived similarly, except that *LeftD* and *RightD* will be measured in number of English words. For instance, with *ENE* and *RENE* as "*Lake Constance*" and "*on Lake Constance*" respectively, *LeftD* and *RightD* will be "+1" and "0". Also, the bigram unit cc_m of the Chinese NE string is replaced by the English word unit ew_n.

All the bilingual and monolingual factors mentioned above, which are derived from Eq (1), are weighted differently according to their contributions. The corresponding weighting coefficients are obtained using the well-known *Minimum Error Rate Training* (Och, 2003; commonly abbreviated as MERT) algorithm by minimizing the number of associated errors in the development set.

3.3 Framework for the Proposed Model

The above model is implemented with a three-stage framework: (A) Initial NE Recognition; (B) NE-Candidate-Set Expansion; and (C) NE Alignment&Re-identification. The Following Diagram gives the details of this framework:

For each given bilingual sentence-pair:

(A) Initial NE Recognition: generates the initial NE anchors with off-the-self packages.

(B) NE-Candidate-Set Expansion: For each initially detected NE, several NE candidates will be re-generated from the original NE by allowing its boundaries to be shrunk or enlarged within a pre-specified range.

- (B.1) Create both RCNE and RENE candidate-sets, which are expanded from those initial NEs identified in the previous stage.
- (B.2) Construct an NE-pair candidate-set (named *NE-Pair-Candidate-Set*), which is the Cartesian product of the RCNE and RENE candidate-sets created above.

(C) NE Alignment&Re-identification: Rank each candidate in the NE-Pair-Candidate-Set constructed above with the linking score specified in Eq (1). Afterwards, conduct a beam search process to select the top *K* non-overlapping NE-pairs from this set.

Diagram 1. Steps to Generate the Final NE-Pairs

It is our observation that, four Chinese characters for both shrinking and enlarging, two English words for shrinking and three for enlarging are enough in most cases. Under these conditions, the including-rates for NEs with correct boundaries are raised to 95.8% for Chinese and 97.4% for English; and even the NE-pair including rate is raised to 95.3%. Since the above range limitation setting has an including-rate only 0.8% lower than that can be obtained without any range limitation (which is 96.1%), it is adopted in this paper to greatly reduce the number of NE-pair-candidates.

4 Experiments

To evaluate the proposed joint approach, a prior work (Huang et al., 2003) is re-implemented in our environment as the baseline, in which the translation cost, transliteration cost and tagging cost are used. This model is selected for comparison because it not only adopts the same candidate-set expansion strategy as mentioned above, but also utilizes the monolingual information when selecting NE-pairs (however, only a simple bi-gram model is used as the tagging cost in their paper). Note that it enforces the same NE type only when the tagging cost is evaluated:

$$C_{tag} = \min_{RType}[-\log(\prod_{m=1}^{M} P(cc_m \mid cc_{m-1}, RType)) - \log(\prod_{n=1}^{N} P(ew_n \mid ew_{n-1}, RType))]$$

To give a fairer comparison, the same training-set and testing-set are adopted. The training-set includes two parts. The first part consists of 90,412 aligned sentence-pairs newswire data from the Foreign Broadcast Information Service (FBIS), which is denoted as Training-Set-I. The second Part of the training set is the LDC2005T34 bilingual NE dictionary[3], which is denoted as Training-Set-II. The required feature information is then manually labeled throughout the two training sets.

In our experiments, for the baseline system, the translation cost and the transliteration cost are trained on Training-Set-II, while the tagging cost is trained on Training-Set-I. For the proposed approach, the monolingual candidate certainty factors are trained on Training-Set-I, and Training-Set-II is used to train the parameters relating to bilingual alignment factors.

For the testing-set, 300 sentence pairs are randomly selected from the LDC Chinese-English News Text (LDC2005T06). The average length of the Chinese sentences is 59.4 characters, while the average length of the English sentences is 24.8 words. Afterwards, the answer keys for NE recognition and alignment were annotated manually, and used as the gold standard to calculate metrics of precision (P), recall (R), and F-score (F) for both *NE recognition* (NER) and *NE alignment* (NEA). In Total 765 Chinese NEs and 747 English NEs were manually labeled in the testing-set, within which there are only 718 NE pairs, including 214 PER, 371 LOC and 133 ORG NE-pairs. The number of NE pairs is less than that of NEs, because not all those recognized NEs can be aligned.

Besides, the development-set for MERT weight training is composed of 200 sentence pairs selected from the LDC2005T06 corpus, which includes 482 manually tagged NE pairs. There is no overlap between the training-sets, the development-set and the testing-set.

4.1 Baseline System

Both the baseline and the proposed models share the same initial detection subtask, which adopts the Chinese NE recognizer reported by Wu et al. (2005), which is a hybrid statistical model incorporating multi-knowledge sources, and the English NE recognizer included in the publicly available Mallet toolkit[4] to generate initial NEs. Initial Chinese NEs and English NEs are recognized by these two available packages respectively.

NE-type	P (%): C/E	R (%): C/E	F (%): C/E
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

Table 1. Initial Chinese/English NER

Table 1 shows the initial NE recognition performances for both Chinese and English (the largest entry in each column is highlighted for visibility). From Table 1, 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 partially recognized or missed. Besides, not shown in the table, the location names or abbreviated organization names tend to be incorrectly recognized as person names. 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.

When those initially identified NEs are directly used for baseline alignment, only 64.1% F score (regard of their name types) is obtained. Such a low performance is mainly due to those NE recognition errors which have been brought into the alignment stage.

To diminish the effect of errors accumulating, which stems from the recognition stage, the baseline system also adopts the same expansion strategy described in Section 3.3 to enlarge the possi-

[3] The LDC2005T34 data-set consists of proofread bilingual entries: 73,352 person names, 76,460 location names and 68,960 organization names.

[4] http://mallet.cs.umass.edu/index.php/Main_Page

ble NE candidate set. However, only a slight improvement (68.4% type-sensitive F-score) is obtained, as shown in Table 2. Therefore, it is conjectured that the baseline alignment model is unable to achieve good performance if those features/factors proposed in this paper are not adopted.

4.2 The Recognition and Alignment Joint Model

To show the individual effect of each factor in the joint model, a series of experiments, from Exp0 to Exp11, are conducted. Exp0 is the *basic system*, which ignores monolingual candidate certainty scores, and also disregards mapping type and NE type consistency constraint by ignoring $P(Mtype_n \mid ew_{[n]}, RType)$ and $P(\delta \mid RType)$, and also replacing $P(cpn_{<n>} \mid Mtype_n, ew_{[n]}, RType)$ with $P(cpn_{<n>} \mid ew_{[n]})$ in Eq (3).

To show the effect of enforcing NE type consistency constraint on internal component mapping, Exp1 (named *Exp0+RType*) replaces $P(cpn_{<n>} \mid ew_{[n]})$ in Exp0 with $P(cpn_{<n>} \mid ew_{[n]}, RType)$; On the other hand, Exp2 (named *Exp0+MappingType*) shows the effect of introducing the component mapping type to Eq (3) by replacing $P(cpn_{<n>} \mid ew_{[n]})$ in Exp0 by $P(cpn_{<n>} \mid Mtype_n, ew_{[n]}) \times P(Mtype_n \mid ew_{[n]})$; Then Exp3 (named *Exp2+MappingTypeRatio*) further adds $P(\delta \mid RType)$ to Exp2, to manifest the contribution from the mapping type ratio. In addition, Exp4 (named *Exp0+RTypeReassignment*) adds the NE type reassignment score, Eq (4), to Exp0 to show the effect of enforcing NE-type consistency. Furthermore, Exp5 (named *All-BiFactors*) shows the full power of the set of proposed bilingual factors by turning on all the options mentioned above. As the bilingual alignment factors would favor the candidates with shorter lengths, $P([cpn_{<n>}, ew_{[n]}, Mtype_n]_{n=1}^{N}, \delta \mid RType, RENE)$, Eq (3), is further *normalized* into the following form:

$$\left[\prod_{n=1}^{N} P(cpn_{<n>} \mid Mtype_n, ew_{[n]}, RType) \times P(Mtype_n \mid ew_{[n]}, RType)\right]^{\frac{1}{N}} \times P(\delta \mid RType),$$

and is shown by Exp6 (named *All-N-BiFactors*).

To show the influence of additional information carried by those initially recognized NEs, Exp7 (named *Exp6+LeftD/RightD*) adds left and right distance information into Exp6, as that specified in Eq (5). To study the monolingual bigram capability, Exp8 (named *Exp6+Bigram*) adds the NEtype dependant bigram model of each language to Exp6. We use SRI Language Modeling Toolkit[5] (SRILM) (Stolcke, 2002) to train various character/word based bi-gram models with different NE types. Similar to what we have done on the bilingual alignment factor above, Exp9 (named *Exp6+N-Bigram*) adds the *normalized* NEtype dependant bigram to Exp6 for removing the bias induced by having different NE lengths. The normalized Chinese NEtype dependant bigram score is defined as $[\prod_{m=1}^{M} P(cc_m \mid cc_{m-1}, RType)]^{\frac{1}{M}}$. A Similar transformation is also applied to the English side.

Lastly, Exp10 (named *Fully-JointModel*) shows the full power of the proposed Recognition and Alignment Joint Model by adopting all the normalized factors mentioned above. The result of a MERT weighted version is further shown by Exp11 (named *Weighted-JointModel*).

Model	P (%)	R (%)	F (%)
Baseline	77.1 (67.1)	79.7 (69.8)	78.4 (68.4)
Exp0 (*Basic System*)	67.9 (62.4)	70.3 (64.8)	69.1 (63.6)
Exp1 (*Exp0 + Rtype*)	69.6 (65.7)	71.9 (68.0)	70.8 (66.8)
Exp2 (*Exp0 + MappingType*)	70.5 (65.3)	73.0 (67.5)	71.7 (66.4)
Exp3 (*Exp2 + MappingTypeRatio*)	72.0 (68.3)	74.5 (70.8)	73.2 (69.5)
Exp4 (*Exp0 + RTypeReassignment*)	70.2 (66.7)	72.7 (69.2)	71.4 (67.9)
Exp5 (*All-BiFactors*)	76.2 (**72.3**)	78.5 (**74.6**)	77.3 (**73.4**)
Exp6 (*All-N-BiFactors*)	**77.7** (**73.5**)	79.9 (**75.7**)	78.8 (**74.6**)
Exp7 (*Exp6 + LeftD/RightD*)	**83.5** (**77.7**)	**85.8** (**80.1**)	**84.6** (**78.9**)
Exp8 (*Exp6 + Bigram*)	**80.4** (**75.5**)	**82.7** (**77.9**)	**81.5** (**76.7**)
Exp9 (*Exp6 + N-Bigram*)	**82.7** (**77.1**)	**85.1** (**79.6**)	**83.9** (**78.3**)
Exp10 (*Fully-JointModel*)	**83.7** (**78.1**)	**86.2** (**80.7**)	**84.9** (**79.4**)
Exp11 (*Weighted-Joint Model*)	**85.9** (**80.5**)	**88.4** (**83.0**)	**87.1** (**81.7**)

Table 2. NEA Type-Insensitive (Type-Sensitive) Performance

Since most papers in the literature are evaluated only based on the boundaries of NEs, two kinds of performance are thus given here. The first one (named type-insensitive) only checks the scope of each NE without taking its associated NE type into consideration, and is reported

[5] http://www.speech.sri.com/projects/srilm/

as the main data at Table 2. The second one (named type-sensitive) would also evaluate the associated NE type of each NE, and is given within parentheses in Table 2. A large degradation is observed when NE type is also taken into account. The highlighted entries are those that are statistically better[6] than that of the baseline system.

4.3 ME Approach with Primitive Features

Although the proposed model has been derived above in a principled way, since all these proposed features can also be directly integrated with the well-known maximum entropy (ME) (Berger et al., 1996) framework without making any assumptions, one might wonder if it is still worth to deriving a model after all the related features have been proposed. To show that not only the features but also the adopted model contribute to the performance improvement, an ME approach is tested as follows for comparison. It directly adopts all those primitive features mentioned above as its inputs (including internal component mapping, initial and final NE type, NE bigram-based string, and left/right distance), without involving any related probability factors derived within the proposed model.

This ME method is implemented with a public package YASMET[7], and is tested under various training-set sizes (400, 4,000, 40,000, and 90,412 sentence-pairs). All those training-sets are extracted from the Training-Set-I mentioned above (a total of 298,302 NE pairs included are manually labeled). Since the ME approach is unable to utilize the bilingual NE dictionary (Training-Set-II), for fair comparison, this dictionary was also not used to train our models here. Table 3 shows the performance (F-score) using the same testing-set. The data within parentheses are relative improvements.

Model	400	4,000	40,000	90,412
ME framework	36.5 (0%)	50.4 (0%)	62.6 (0%)	67.9 (0%)
Un-weighted-JointModel	+4.6 (+12.6%)	+4.5 (+8.9%)	+4.3 (+6.9%)	+4.1 (+6.0%)
Weighted-JointModel	+5.0 (+13.7%)	+4.7 (+9.3%)	+4.6 (+7.3%)	+4.5 (+6.6%)

Table 3. Comparison between ME Framework and Derived Model on the Testing-Set

[6] Statistical significance test is measured on 95% confidence level on 1,000 re-sampling batches (Zhang et al., 2004)
[7] http://www.fjoch.com/YASMET.html

The improvement indicated in Table 3 clearly illustrates the benefit of deriving the model shown in Eq (2). Since a reasonably derived model not only shares the same training-set with the primitive ME version above, but also enjoys the additional knowledge introduced by the human (i.e., the assumptions/constraints implied by the model), it is not surprising to find out that a good model does help, and that it also becomes more noticeable as the training-set gets smaller.

5 Error Analysis and Discussion

Although the proposed model has substantially improved the performance of both NE alignment and recognition, some errors still remain. Having examined those type-insensitive errors, we found that they can be classified into four categories: (A) Original NEs or their components are already not one-to-one mapped (23%). (B) NE components are one-to-one linked, but the associated NE anchors generated from the initial recognition stage are either missing or spurious (24%). Although increasing the number of output candidates generated from the initial recognition stage might cover the missing problem, possible side effects might also be expected (as the complexity of the alignment task would also be increased). (C) Mapping types are not assumed by the model (27%). For example, one NE is abbreviated while its counterpart is not; or some loanwords or out-of-vocabulary terms are translated neither semantically nor phonetically. (D) Wrong NE scopes are selected (26%). Errors of this type are uneasy to resolve, and their possible solutions are beyond the scope of this paper.

Examples of above category (C) are interesting and are further illustrated as follows. As an instance of abbreviation errors, a Chinese NE "葛兰素制药厂 (GlaxoSmithKline Factory)" is tagged as "葛兰素/PRR 制药厂/n", while its counterpart in the English side is simply abbreviated as "GSK" (or replaced by a pronoun "it" sometimes). Linking "葛兰素" to "GSK" (or to the pronoun "it") is thus out of reach of our model. It seems an abbreviation table (or even anaphora analysis) is required to recover these kind of errors.

As an example of errors resulting from loanwords; Japanese kanji "明仁" (the name of a Japanese emperor) is linked to the English word "Akihito". Here the Japanese kanji "明仁" is directly adopted as the corresponding Chinese characters (as those characters were originally borrowed from Chinese), which would be pro-

nounced as “Mingren” in Chinese and thus deviates greatly from the English pronunciation of “Akihito”. Therefore, it is translated neither semantically nor phonetically. Further extending the model to cover this new conversion type seems necessary; however, such a kind of extension is very likely to be language pair dependent.

6 Capability of the Proposed Model

In addition to improving NE alignment, the proposed joint model can also boost the performance of NE recognition in both languages. The corresponding differences in performance (of the weighted version) when compared with the initial NER (ΔP, ΔR and ΔF) are shown in Table 4. Again, those marked entries indicate that they are statistically better than that of the original NER.

NEtype	ΔP (%): C/E	ΔR (%): C/E	ΔF (%): C/E
PER	**+5.4** / **+6.4**	+2.2 / **+2.6**	**+3.9** / **+4.6**
LOC	**+4.0** / **+3.4**	-0.2 / **+2.7**	+1.8 / **+3.0**
ORG	**+7.0** / **+3.9**	**+5.6** / **+9.1**	**+6.2** / **+6.4**
ALL	**+5.3** /**+5.2**	**+2.4** / **+4.0**	**+3.9** / **+4.6**

Table 4. Improvement in Chinese/English NER

The result shows that the proposed joint model has a clear win over the initial NER for either Chinese or English NER. In particular, ORG seems to have yielded the greatest gain amongst NE types, which matches our previous observations that the boundaries of Chinese ORG are difficult to identify with the information only coming from the Chinese sentence, while the type of English ORG is uneasy to classify with the information only coming from the English sentence.

Though not shown in the tables, it is also observed that the proposed approach achieves a 28.9% reduction on the spurious (false positive) and partial tags over the initial Chinese NER, as well as 16.1% relative error reduction compared with the initial English NER. In addition, total 27.2% wrong Chinese NEs and 40.7% wrong English NEs are corrected into right NE types. However, if the mapping type ratio is omitted, only 21.1% wrong Chinese NE types and 34.8% wrong English NE types can be corrected. This clearly indicates that the ratio is essential for identifying NE types.

With the benefits shown above, the alignment model could thus be used to train the monolingual NE recognition model via semi-supervised learning. This advantage is important for updating the NER model from time to time, as various domains frequently have different sets of NEs and new NEs also emerge with time.

Since the Chinese NE recognizer we use is not an open source toolkit, it cannot be used to carry out semi-supervised learning. Therefore, only the English NE recognizer and 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 different seed corpora (100, 400, 4,000 and 40,000 sentence-pairs). Table 5 shows the results of semi-supervised learning after convergence for adopting only the English NER model (*NER-Only*), the baseline alignment model (*NER+Baseline*), and our un-weighted joint model (*NER+JointModel*) respectively. The *Initial-NER* row indicates the initial performance of the NER model re-trained from different seed corpora. The data within parentheses are relative improvement over *Initial-NER*. Note that the testing set is still the same as before.

As Table 5 shows, with the NER model alone, the performance may even deteriorate after convergence. This is due to the fact that maximizing likelihood does not imply minimizing the error rate. However, with additional mapping constraints from the aligned sentence of another language, the alignment module could guide the searching process to converge to a more desirable point in the parameter space; and these additional constraints become more effective as the seed-corpus gets smaller.

Model	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+Baseline	+4.9 (+13.4%)	+3.4 (5.8%)	+1.7 (2.4%)	+0.7 (0.9%)
NER+Joint Model	+10.7 (+29.2%)	+8.7 (+14.8%)	+4.8 (+6.7%)	+2.3 (+2.9%)

Table 5. Testing-Set Performance for Semi-Supervised Learning of English NE Recognition

7 Conclusion

In summary, our experiments show that the new monolingual candidate certainty factors are more effective than the tagging cost (only bigram model) adopted in the baseline system. Moreover, both the mapping type ratio and the entity type consistency constraint are very helpful in identifying the associated NE boundaries and types. After having adopted the features and enforced

the constraint mentioned above, the proposed framework, which jointly recognizes and aligns bilingual named entities, achieves a remarkable 42.1% imperfection reduction on type-sensitive F-score (from 68.4% to 81.7%) in our Chinese-English NE alignment task.

Although the experiments are conducted on the Chinese-English language pair, it is expected that the proposed approach can also be applied to other language pairs, as no language dependent linguistic feature (or knowledge) is adopted in the model/algorithm used.

Acknowledgments

The research work has been partially supported by the National Natural Science Foundation of China under Grants No. 60975053, 90820303, and 60736014, the National Key Technology R&D Program under Grant No. 2006BAH03B02, and also the Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2006AA010108-4.

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 (ACL)*, pages 400-408.

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):39-72, March.

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.

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.

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*. Sappora, Japan.

Ji, Heng and Ralph Grishman. 2006. Analysis and Repair of Name Tagger Errors. In *Proceedings of COLING/ACL 06*, Sydney, Australia.

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.

Moore, R. C.. 2003. Learning Translations of Named-Entity Phrases from Parallel Corpora. In *Proceedings of 10th Conference of the European Chapter of ACL*, Budapest, Hungary.

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). July 8-10, 2003. Sapporo, Japan. Pages: 160-167.

Stolcke, A. 2002. SRILM -- An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver.

Wu, Youzheng, Jun Zhao and Bo Xu. 2005. Chinese Named Entity Recognition Model Based on Multiple Features. In *Proceedings of HLT/EMNLP* 2005, pages 427-434.

Zhang, Ying, Stephan Vogel, and Alex 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 2051--2054.