# Named Entity Recognition Based on Bilingual Co-training

Yegang Li[1,2,3], Heyan Huang[1,2,*], Xingjian Zhao[1,2], and Shumin Shi[1,2]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
{lyg8256,hhy63,wisedo,bjssm}@bit.edu.cn
[2] Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing (Beijing Institute of Technology), Beijing, China
[3] Department of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, China
lyg8256@bit.edu.cn

**Abstract.** Named entity recognition (NER) is a very important task in natural language processing (NLP). In this paper we present a semi-supervised approach to extract bilingual named entity, starting from a bilingual corpus where the named entities are extracted independently for each language. Then a bilingual co-training algorithm is used to improve the named entity annotation quality, and iterative process is applied to extract named entity pairs with higher bilingual conformity ratio. This leads to a significant improvement of the monolingual named entity annotation quality for both languages. Experimental result shows that the annotation quality of Chinese NE is improved from 87.17 to 88.28, and improved 80.37 to 81.76 of English NE in F-measure.

**Keywords:** named entity recognition, bilingual co-training, natural language processing.

## 1 Introduction

NER is a frequently needed technology in natural language processing applications. State-of-the-art supervised statistical models for NER typically require large amount of labeled data and linguistic expertise to be sufficiently accurate, which makes it difficult to build high-quality models for so many languages.

Recently, there have been some works which offered hope for creating NER analyzers in many languages, using parallel English-foreign language data, a high-quality NER tagger for English, and projected annotations for the foreign language[1]. Parallel data has also been used to improve existing monolingual taggers or other analyzers in both languages[2].

Both Chinese and English have their own special features that can be employed for entity extraction. Chinese does not have white space for tokenization or capitalization, features which, for English, can help identify name boundaries and distinguish names

* Corresponding author.

P. Liu and Q. Su (Eds.): CLSW 2013, LNAI 8229, pp. 480–489, 2013.

from nominal. Using bilingual co-training allows us to capture such indicative information to improve Chinese name tagging. For example,

1. Results from Chinese name tagger： <PER>金庸新[*jinyongxin*]( *Jin Yongxin*)</PER >小说*[xiaoshuo]*(*novel*).
2. Results from English name tagger: the new novels of <PER>Jin Yong</PER>
3. Name tagging after using bilingual co-training: < PER >金庸[*jinyong*]( *Jin Yong*)</PER >新小说[*xinxiaoshuo*]( *new novels*)

“金庸新” and “金庸” in Chinese can be a PER name, while its English translation “Jin Yong” indicates that “金庸” is more likely to be a PER name than “金庸新”.

On the other hand, Chinese has some useful language-specific properties for entity extraction. For example, standard Chinese family names are generally single characters drawn from a fixed set of 437 family names, and almost all first names include one or two characters. The suffix words (if there are any) of ORG and GPE names belong to relatively distinguishable fixed lists. For example,

1. Results from English name tagger：The captain of a ferry boat who works on 〈PER〉 Lake Constance 〈/PER〉 ...
2. Results from Chinese name tagger: 在 〈 LOC 〉 康斯坦茨湖 [kangsitancihu]( Lake Constance)〈/LOC〉 工作的一艘渡船的船长...
3. Name tagging after using bilingual co-training: The captain of a ferry boat who works on 〈LOC〉 Lake Constance 〈/LOC〉 …

“Lake” in English can be the suffix word of either a PER or LOC name, while its Chinese translation “康斯坦茨湖” indicates that “Lake Constance” is more likely to be a LOC name.

To solve the problems, a semi-supervised learning framework with the help of parallel corpus is proposed in this paper. We propose a semi-supervised approach using a bilingual co-training model to carry out English and Chinese NER. To ease error propagations from the projected annotations, we correct the projected annotations with a maximum entropy model. The maximum entropy model addresses the Named Entity alignment of a bilingual corpus, which builds an alignment between each source NE and its translation NE in the target language. A Named Entity alignment, however, is not easy to obtain. It requires both Named Entity Recognition and alignment to be handled correctly. NEs may not be well recognized, or only parts of them may be recognized during NER. When aligning bilingual NEs in different languages, we need to handle many-to-many alignments. Although this makes the task more difficult, it greatly reduces the chance of errors introduced by previous steps and therefore produces much better performance on our task.

The rest of this paper is organized as follows: In section 2, we discuss related work on NE recognition and alignment. In section 3, we discuss bilingual co-training algorithm. Section 4 gives the overall framework of NE alignment with our maximum entropy model. Feature functions are also explained in this section. We show

experimental results and compare them with baseline systems in Section 5. Section 6 concludes the paper and discusses future work.

## 2 Related Work

NER, including proper names, temporal and numerical expressions, has been widely addressed by symbolic, statistical as well as hybrid approaches. Its major part in information extraction (IE) and other NLP applications has been stated and encouraged by several editions of evaluation campaigns such as MUC[3], the CoNLL-2003 NER shared task[4] or ACE[5], where NER systems show near-human performances for the English language. Our model builds upon prior work on co-training and cross-lingual projection for named entities. Other interesting work on aligning named entities in two languages is reported in [1-2],[6-9].

Our bilingual co-training approach is related to bilingual labeling models presented in previous work. Still, there are some disadvantages for previous work based on parallel corpora. First, current NE alignment methods are not accurate enough, and many noises could be introduced during the word alignment stage. Second, *manual annotation* is usually obtained from a few limited domains, leading to a bad affect on statistical supervised learning methods. Following of previous work we focus on the task of bilingual NE recognition. In contrast, our bilingual co-training model does not require large amount of labeled data, since we conduct our experiments with the co-training algorithm.

## 3 Bilingual Co-training

Starting with a set of labeled data, co-training algorithms attempt to increase the amount of annotated data using large amounts of unlabeled data. The process may continue for several iterations. In natural language processing, co-training was generally found to bring improvement over the cases when no additional unlabeled data are used. One important aspect of co-training consists in the relation between the views used in learning. Blum and Mitchell[10] states conditional independence of the views as a required criterion for co-training. Abney[11] shows that the independence assumption can be relaxed, and co-training is still effective under a weaker independence assumption.

In this work, we apply co-training by regarding the parallel Chinese-English sentences as weaker independent views for NE identity. Instances are selected by bilingual conformity ratio on a set of unlabeled instances, and the instances most confidently labeled are added to the labeled data. This procedure preserves the distribution of labels in the labeled data as instances are labeled and added. The bilingual conformity ratio can be defined as follows:

$$conformity\_\text{ratio} = \frac{1}{n}\sum_{U}\frac{1}{K}\sum_{k=1}^{K} conformity(ws_i, wt_j)_k$$

$$conformity(ws_i, wt_j)_k = \begin{cases} 1 & T(ws_i) = T(wt_j) \\ 0 & T(ws_i) \neq T(wt_j) \end{cases}. \quad (1)$$

Where $(ws_i, wt_j)_k$ represents k($1 \le k \le K$) pair word in the parallel sentence, and $T(ws_i), T(wt_j)$ represent the annotations of NE. The bilingual co-training algorithm is discussed as follows.

---

1. Given:
   (a) A set Ls of source labeled examples
   (b) A set Lt of target labeled examples
   (c) A set Us of source unlabeled examples
   (d) A set Ut of target unlabeled examples

---

4. Classifiers
   (e) Use Ls to train the classifiers Classifier(s)
   (f) Use Lt to train the classifiers Classifier(t)
5. Loop for m iterations
   (g) Create a pairs pool $\overline{Us}$ and $\overline{Ut}$, with examples from Us and Ut, create $\overline{Us}$ and $\overline{Ut}$ by labeled the examples in $\overline{Us}$ and $\overline{Ut}$ with Classifier(s) and Classifier(t),calculate $conformity\_ratio(\overline{Us}, \overline{Ut})$, $\max \leftarrow conformity\_ratio(\overline{Us}, \overline{Ut})$, create $\tilde{Lt}^* \leftarrow null$, $\tilde{Ls}^* \leftarrow null$
   (h) Loop for 10 iterations
      (i) Create $(\hat{Ls}, \hat{Lt})$ with k pairs sentences extracted from $(\overline{Us}, \overline{Ut})$, projected annotations from $\hat{Ls}$ to $\hat{Lt}$ ,create $\tilde{Lt}$ by corrected the projected annotations with $\hat{Lt}$
      (ii) $classifier(t) \leftarrow classfier(Lt \cup \tilde{Lt})$, train Classifier(t), label $\overline{Ut}$ with Classifier(t), update $conformity\_ratio(\overline{Us}, \overline{Ut})$, if $conformity\_ratio(\overline{Us}, \overline{Ut}) > \max$, then $\max \leftarrow conformity\_ratio(\overline{Us}, \overline{Ut})$, $\tilde{Lt}^* \leftarrow \tilde{Lt}$
      (iii) $Lt \leftarrow Lt \cup \tilde{Lt}^*$, train classifier(t) with Lt: $classifier(t) \leftarrow classfier(Lt)$

---

(i) Loop for 10 iterations

(i) Create $(\hat{L}s, \hat{L}t)$ with k pairs of sentences extracted from $(\overline{U}s, \overline{U}t)$, projected annotations from $\hat{L}t$ to $\hat{L}s$, create $\tilde{L}s$ by corrected the projected annotations with $\tilde{L}s$.

(ii) $classifier(s) \leftarrow classfier(Ls \bigcup \tilde{L}s)$, train Classifier(s), labeled $\overline{U}s$ with Classifier(s), update $conformity\_ratio(\overline{U}s, \overline{U}t)$, if $conformity\_ratio(\overline{U}s, \overline{U}t) > \max$, then $\max \leftarrow conformity\_ratio(\overline{U}s, \overline{U}t)$, $\tilde{L}s^* \leftarrow \tilde{L}s$.

(iii) $Ls \leftarrow Ls \bigcup \tilde{L}s^*$, $classifier(s) \leftarrow classfier(Ls)$.

## 4 Corrective NE Projection Annotation

To ease error propagations from the projected annotations, we corrected the projected annotations with a Corrective model.

### 4.1 Projection NE Candidate

For each word in the source language NE, we find all the possible projection word in target language through the word alignment [12]. Next, we have all the projection words as the "seed" data. With an open-ended window for each seed, all the possible sequences located within the window are considered as possible candidates for NE projection. Their lengths range from 1 to the empirically determined length of the window. During the best candidate projection NE selection, the NE alignment model discussed as follows is applied to search the best projection NE.

### 4.2 NE Alignment Model

There are several valuable features that can be used for NE alignment. Considering the advantages of the maximum entropy model [13] to integrate different kinds of features, we use this framework to handle our problem. Suppose the English NE $Ne_c^d$, and the Chinese NE $Nc_a^b$. Suppose also that we have M feature functions $f_m(a_k, Nc_a^b, N\tilde{e}_c^d), m = 1,2,\cdots,M$. For each feature function, we have a model parameter $\lambda_m, m = 1,2,\cdots,M$. The alignment probability can be defined as follows [14]:

$$P(a_k \mid Nc_a^b, N\tilde{e}_c^d) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m f_m\left(a_k, Nc_a^b, N\tilde{e}_c^d\right)\right)}{\sum_{A} \exp\left(\sum_{m=1}^{M} \lambda_m f_m\left(a_k, Nc_a^b, N\tilde{e}_c^d\right)\right)} \tag{2}$$

In our approach, we adopt 3 features: translation feature, the source NE and target NE's co-occurrence feature, and length of NE pair feature. Next, we discuss these three features in detail.

**Translation Feature**

The translation feature here is used to represent how close an NE pair is based on translation probabilities. Given a parallel corpus aligned at the sentence level, we can achieve the translation probability between English chunk and Chinese chunk with IBM Model [15].

$$P(F \mid E) = \frac{1}{(n+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{n} t(f_j \mid e_i) \tag{3}$$

Suppose the candidate English NE, $E = e_1, e_2, \cdots, e_m$ consists of m English words and the candidate Chinese NE, $C = c_1, c_2, \cdots, c_n$ is composed of n Chinese characters. The translation probability of the NE pair is computed as follows,

$$P(Nc \mid Ne) = \frac{1}{(n+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{n} t(c_j \mid e_i) \qquad 1 \le j \le n, 1 \le i \le m \tag{4}$$

We defined translation feature as follows,

$$f_m\left(a_k, Nc_a^b, N\tilde{e}_c^d\right) = \log(P(Nc_a^b \mid N\tilde{e}_c^d)) + \log(P(N\tilde{e}_c^d \mid Nc_a^b)) \tag{5}$$

The scores between the candidate Chinese NEs and the English NEs are calculated via this formula as the value of translation feature.

**Co-occurrence Feature**

If a source NE and a target NE co-occur very often, there exists a big possibility that they align to each other. This probability is a good indication for determining bilingual NE alignment. The co-occurrence feature can be defined as follows,

$$f_m\left(a_k, Nc_a^b, N\tilde{e}_c^d\right) = \frac{count(Nc_a^b, N\tilde{e}_c^d)}{count(Nc_a^b, *)} + \frac{count(Nc_a^b, N\tilde{e}_c^d)}{count(*, N\tilde{e}_c^d)} \tag{6}$$

Where $count(Nc_a^b, N\tilde{e}_c^d)$ is the number of times $Nc_a^b$ and $N\tilde{e}_c^d$ appear together. ($Nc_a^b$,*) is the number of times that $Nc_a^b$ appears.(*, $N\tilde{e}_c^d$) is the number of times that $N\tilde{e}_c^d$ appears.

**Length Feature**

When translating NE across languages, we notice that the difference of their length is also a good indication for determining their relation. The length feature [16] can be defined as follows,

$$f_m\left(a_k, Nc_a^b, N\tilde{e}_c^d\right) \approx f_m\left(a_k, \left|Nc_a^b\right|, \left|N\tilde{e}_c^d\right|\right) = \frac{\left|Nc_a^b\right| - \delta\left|N\tilde{e}_c^d\right|}{\sqrt{\left(\left|Nc_a^b\right|+1\right)^{\sigma^2}}}$$

$$\delta = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{count(Ne_i)}{count(Nc_i)}\right), \sigma^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{count(Ne_j)}{count(Nc_j)} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{count(Ne_i)}{count(Nc_i)}\right)^2\right) \quad (7)$$

Where count($Ne_i$) is the character number of $Ne_i$ and count($Nc_j$) is the character number of $Nc_j$.

## 5 Experimental Results

### 5.1 Experimental Setup

We perform experiments to investigate the performance of the above framework. We take the NiuTrans[17] bilingual training data consists of 100K Chinese-English sentence pairs as our unlabeled corpus. The training and test data sets consist of the Penn Xinhua News corpus sections 1–206 and section 206-305, respectively. They are annotated with 3 types of NE such as person name (PER), location name (LOC) and organization name (ORG).

To achieve the most probable aligned Chinese NE, we use the published package YASMET[1] to conduct parameter training and re-ranking of all the NE candidates. YASMET requires supervised learning for the training of the maximum entropy model. We acquire a large annotated training set from NiuTrans corpus.

This paper maintains a pool of 1000 unlabeled instances by random selection. The classifier proposes labels for the instances in this pool. We choose 100 instances for each classifier with high confidence while preserving the class distribution observed in the initial labeled data, and add them to the labeled data.

[1] http://www.isi.edu/~och/YASMET.html

The automatically generated annotation was then evaluated by calculating precision and recall with respect to this gold standard. Precision is defined as

$$P = \frac{\#of \quad correct \quad annotated \quad NEs}{\#of \quad all \quad annotated \quad NEs}$$

Recall is defined as : $R = \frac{\#of \quad correct \quad annotated \quad NEs}{\#of \quad all \quad correct \quad NEs}$

The F-score, a combined measure of NE annotation's precision and recall, is defined as: $F = \frac{2PR}{P+R}$

### 5.2 Baseline System

We formulate the named entity recognition task as the classification of each word with context to one of the classes that represent region information and named entity's semantic class. MALLET[2] includes tools for sequence tagging for applications such as named-entity extraction from text. Algorithms include Hidden Markov Models, Maximum Entropy Markov Models, and Conditional Random Fields (CRF).

We employ a CRF method in Chinese language as Chinese NER baseline system, and employ MALLET in English language as English NER baseline system. The Chinese and English baseline NE tagging performance on different entity types is shown in Table 1 as follows.

**Table 1.** Baseline F-Measure (%) of NE Tagging

| *NE Type* | *Chinese NE(F-value)* | *English NE(F-value)* |
|---|---|---|
| PER | 89.59 | 81.22 |
| LOC | 88.48 | 80.43 |
| ORG | 84.54 | 79.69 |
| ALL | 87.17 | 80.37 |

### 5.3 Bilingual Co-training Results

Based on the testing strategies discussed in Section 3, we perform all the experiments and get the performance for bilingual co-training. NE tagging performance on different entity types are shown in Table 2 as follows.

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

**Table 2.** Bilingual co-training F-Measure (%) of NE Tagging

| *NE Type* | *Chinese NE(F-value)* | *English NE(F-value)* |
|---|---|---|
| PER | 90.86 | 82.31 |
| LOC | 89.53 | 82.01 |
| ORG | 85.71 | 80.42 |
| ALL | 88.28 | 81.76 |

We restrict the bilingual co-training model to use only features similar to the ones used by the baseline model. We obtain performance much better than that of the baseline model. In this set of experiments, we compare our bilingual co-training method with a named entity recognition system based on CRF method. As shown in table 2, NE improves F-measure over the supervised baseline. For Chinese, it achieves 88.28% in F-measure of all type of NE, which outperforms the supervised baseline by 1.11%. For English, it achieves 81.76% in F-measure of all type of NE, which outperforms the supervised baseline by 1.39%.

The projected data is not completely clean and brings some errors into the final results. But it avoids the acquisition of large annotated training set and the performance is still much better than traditional recognition models.

## 6 Conclusions

Traditional word alignment approaches cannot come up with satisfactory results for NE alignment. Consider that bilingual text can provide valuable additional information for named entity tagging, we propose a novel approach using a bilingual co-training model for NE recognition. The following part proved the assumption that more unlabeled data is used, the better performance is got. This is because additional cross-domain data makes the classification of NE tags more accurate in cross-domain.

While our approach has only been tested on Chinese and English so far, we can expect that it is applicable to other language pairs. The approach is independent of the baseline tagging/extraction system, and so can be used to improve systems with varied learning schemes.

Due to the inconsistency of NE translation, some projected NE is not correct. We may need some manually-generated rules to fix this. This problem will be investigated in the future.

**Acknowledgements.** This work was supported by project of the National Natural Science Foundation of China (No. 61132009, 61201352, 61202244), and the National Basic Research Program of China (973 Program) (2013CB329300, 2013CB329606), and MSRA UR Project (95116953).

## References


1. Das, D., Petrov, S.: Unsupervised part-of-speech tagging with bilingual graph-based projections. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 600–609 (June 2011)
2. Burkett, D., Petrov, S., Blitzer, J., Klein, D.: Learning better monolingual models with unannotated bilingual text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden, pp. 46–54 (July 2010)
3. Marsh, E., Perzanowski, D.: Muc-7 evaluation of ie technology. In: Overview of Proceedings of the Seventh Message Understanding Conference (MUC-7), vol. 20 (1998)
4. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language independent named entity recognition. In: Proceedings of CoNLL, Edmonton, Canada, pp. 142–147 (2003)
5. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction program-tasks, data, and evaluation. In: Proceedings of LREC, vol. 4, pp. 837–840 (2004)
6. Huang, F., Vogel, S., Waibel, A.: Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-Feature Cost Minimization. In: ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, Sapporo, Japan, pp. 9–16 (2003)
7. Moore, R.C.: Learning Translations of Named-Entity Phrases from Parallel Corpora. In: EACL 2003, Budapest, Hungary, pp. 259–266 (2003)
8. Donghui, F., Yajuan, L., Ming, Z.: A new approach for English-Chinese named entity alignment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP, pp. 372–379 (2004)
9. Sungchul, K., Kristina, T., Hwanjo, Y.: Multilingual Named Entity Recognition using Parallel Data and Metadata from Wikipedia. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, July 8-14, pp. 694–702 (2012)
10. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT 1998), Wisconsin, MI, pp. 92–100 (1998)
11. Abney, S.P.: Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 360–367 (2002)
12. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
13. Berger Adam, L., Della Pietra Stephen, A., Della Pietra Vincent, J.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–72 (1996)
14. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 295–302. Association for Computational Linguistics (July 2002)
15. Brown, P.F., Della Pietra Stephen, A., Della Pietra Vincent, J., et al.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2), 263–311 (1993)
16. Church, K.W.: Char align: A program for aligning bilingual texts at the character level. In: The 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp. 1–8 (1993)
17. Tong, X., Jingbo, Z., Hao, Z., et al.: NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics System Demonstrations, Jeju, Korea, pp. 19–24 (July 2012)