Named Entity Recognition with Bilingual Constraints

Wanxiang Che[†] **Mengqiu Wang**[‡] **Christopher D. Manning**[‡] **Ting Liu**[†]

[†]{car, tliu}@ir.hit.edu.cn
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China, 150001

[‡]{mengqiu, manning}@stanford.edu
Computer Science Department
Stanford University
Stanford, CA, 94305

Abstract

Different languages contain complementary cues about entities, which can be used to improve Named Entity Recognition (NER) systems. We propose a method that formulates the problem of exploring such signals on unannotated parallel text as a simple Integer Linear Program, which encourages entity labels to agree via bilingual constraints. Parallel text NER experiments on the large OntoNotes 4.0 Chinese-English corpus show that the proposed method can improve strong baselines for both Chinese and English. In particular, Chinese performance improves by over 5% absolute F_1 score. We can then annotate a large amount of parallel text (80k sentences) using our method, and add it as up-training data to the original NER training corpus. The monolingual Chinese model retrained on this new combined dataset outperforms the strong baseline by over 3% F_1 score.

1 Introduction

Named Entity Recognition (NER) is an important task for many applications, such as information extraction and machine translation. State-of-the-art supervised NER methods require large amounts of annotated data, which are difficult and expensive to produce manually, especially for resource-poor languages.

A promising approach for improving NER performance without annotating more data is to exploit unannotated bilingual text (bitext), which are relatively easy to obtain for many language pairs, borrowing from the resources made available by statistical machine translation research.[1] Different languages contain complementary cues about entities. For example, in Figure 1, the word "本 (Ben)" is common in Chinese but rarely appears as a translated foreign name. However, its aligned word on the English side ("Ben") provides a strong clue that this is a person name. Judicious use of this type of bilingual cues can help to recognize errors a monolingual tagger would make, allowing us to produce more accurately tagged bitext. Each side of the tagged bitext can then be used to expand the original monolingual training dataset, which may lead to higher accuracy in the monolingual taggers.

Previous work such as Li et al. (2012) and Kim et al. (2012) demonstrated that bilingual corpus annotated with NER labels can be used to improve monolingual tagger performance. But a major drawback of their approaches are the need for manual annotation efforts to create such corpora. To avoid this requirement, Burkett et al. (2010) suggested a "multi-view" learning scheme based on re-ranking. Noisy output of a "strong" tagger is used as training data to learn parameters of a log-linear re-ranking model with additional bilingual features, simulated by a "weak" tagger. The learned parameters are then reused with the "strong" tagger to re-rank its own outputs for unseen inputs. Designing good "weak" taggers so that they complement the "view" of bilingual features in the log-linear re-ranker is crucial to the success of this algorithm. Unfortunately there is no principled way of designing such "weak" taggers.

In this paper, we would like to explore a conceptually much simpler idea that can also take ad-

[1] opus.lingfil.uu.se

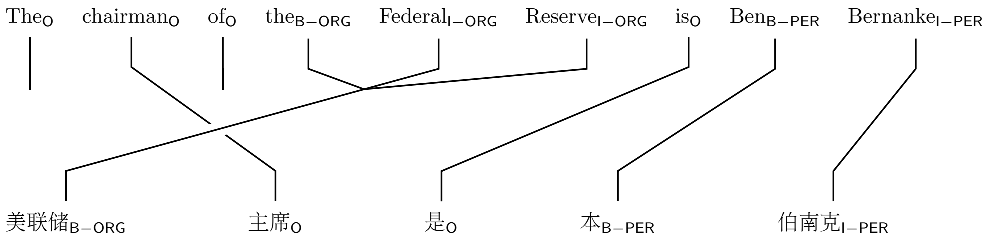


Figure 1: Example of NER labels between two word-aligned bilingual parallel sentences.

vantage of the large amount of unannotated bitext, without complicated machinery. More specifically, we introduce a joint inference method that formulates the bilingual NER tagging problem as an Integer Linear Program (ILP) and solves it during decoding. We propose a set of intuitive and effective bilingual linear constraints that encourage entities to agree across the two languages.

Experimental results on the OntoNotes 4.0 Chinese-English parallel corpus show that the proposed method can improve the strong Chinese NER baseline by over 5% F_1 score and also give small improvements over the English baseline. Moreover, by adding the automatically annotated data to the original NER training corpus and retraining the monolingual model using an up-training regimen (Petrov et al., 2010), we can improve monolingual Chinese NER performance by over 3% F_1 score.

2 Constraint-based Monolingual NER

NER is a sequence labeling task where we assign a named entity label to each word in an input sentence. One commonly used labeling scheme is the BIO scheme. The label B-X (Begin) represents the first word of a named entity of type X, for example, PER (Person) or LOC (Location). The label I-X (Inside) indicates that a word is part of an entity but not first word. The label O (Outside) is used for all non-entity words.[2] See Figure 1 for an example tagged sentence.

Conditional Random Fields (CRF) (Lafferty et al., 2001) is a state-of-the-art sequence labeling model widely used in NER. A first-order linear-chain CRF defines the following conditional probability:

$$P_{CRF}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_i M_i(y_i, y_{i-1}|\mathbf{x}) \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output sequences, respectively, $Z(\mathbf{x})$ is the partition function, and M_i is the clique potential for edge clique i. Decoding in CRF involves finding the most likely output sequence that maximizes this objective, and is commonly done by the Viterbi algorithm.

Roth and Yih (2005) proposed an ILP inference algorithm, which can capture more task-specific and global constraints than the vanilla Viterbi algorithm. Our work is inspired by Roth and Yih (2005). But instead of directly solving the shortest-path problem in the ILP formulation, we re-define the conditional probability as:

$$P_{MAR}(\mathbf{y}|\mathbf{x}) = \prod_i P(y_i|\mathbf{x}) \quad (2)$$

where $P(y_i|\mathbf{x})$ is the marginal probability given by an underlying CRF model computed using *forward-backward* inference. Since the early HMM literature, it has been well known that using the marginal distributions at each position works well, as opposed to Viterbi MAP sequence labeling (Mérialdo, 1994). Our experimental results also supports this claim, as we will show in Section 6. Our objective is to find an optimal NER label sequence:

$$\begin{aligned}\hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} P_{MAR}(\mathbf{y}|\mathbf{x}) \quad (3)\\ &= \arg\max_{\mathbf{y}} \sum_i \log P(y_i|\mathbf{x})\end{aligned}$$

Then an ILP can be used to solve the inference problem as classification problem with constraints.

[2] While the performance of NER is measured at the entity level (not tag level).

The objective function is:

$$\max \sum_{i=1}^{|\mathbf{x}|} \sum_{y \in Y} z_i^y \log P_i^y \tag{4}$$

where Y is the set of all possible named entity labels. $P_i^y = P(y_i = y|\mathbf{x})$ is the CRF marginal probability that the i^{th} word is tagged with a label y, and z_i^y is an indicator that equals 1 *iff* the i^{th} word is labeled y; otherwise, z_i^y is 0.

If no constraints are identified, then Eq. (4) achieves maximum when all z_i^y are assigned to 1, which violates the condition that each word should only be assigned a single entity label. We can express this with constraints:

$$\forall i : \sum_{y \in Y} z_i^y = 1 \tag{5}$$

After adding the constraints, the probability of the sequence is maximized when each word is assigned the label with highest probability. However, some invalid results may still exist. For example a label O may be wrongly followed by a label I-X, although a named entity cannot start with the label I-X. Therefore, we can add the following constraints:

$$\forall i, \forall \textsf{X} : z_{i-1}^{\textsf{B-X}} + z_{i-1}^{\textsf{I-X}} - z_i^{\textsf{I-X}} \geq 0 \tag{6}$$

which specifies that when the i^{th} word is tagged with I-X ($z_i^{\textsf{I-X}} = 1$), then the previous word can only be tagged with B-X or I-X ($z_{i-1}^{\textsf{B-X}} + z_{i-1}^{\textsf{I-X}} \geq 1$).

3 NER with Bilingual Constraints

This section demonstrates how to jointly perform NER for two languages with bilingual constraints. We assume sentences have been aligned into pairs, and the word alignment between each pair of sentences is also given.

3.1 Hard Bilingual Constraints

We first introduce the simplest *hard* constraints, i.e., each word alignment pair should have the same named entity label. For example, in Figure 1, the Chinese word "美联储" was aligned with the English words "the", "Federal" and "Reserve". Therefore, they have the same named entity labels ORG.[3] Similarly, "本" and "Ben" as well as "伯南克" and "Bernanke" were all tagged with the label PER.

The objective function for bilingual NER can be expressed as follows:

$$\max \sum_{i=1}^{|\mathbf{x}_c|} \sum_{y \in Y} z_i^y \log P_i^y + \sum_{j=1}^{|\mathbf{x}_e|} \sum_{y \in Y} z_j^y \log P_j^y \tag{7}$$

where P_i^y and P_j^y are the probabilities of the i^{th} Chinese word and j^{th} English word to be tagged with a named entity label y, respectively. $\mathbf{x}_c$ and $\mathbf{x}_e$ are respectively the Chinese and English sentences.

Similar to monolingual constrained NER (Section 2), monolingual constraints are added for each language as shown in Eqs. (8) and (9):

$$\forall i : \sum_{y \in Y} z_i^y = 1; \forall j : \sum_{y \in Y} z_j^y = 1 \tag{8}$$

$$\begin{aligned} \forall i, \forall \textsf{X} : z_i^{\textsf{B-X}} + z_i^{\textsf{I-X}} - z_{i+1}^{\textsf{B-X}} \geq 0 \\ \forall j, \forall \textsf{X} : z_j^{\textsf{B-X}} + z_j^{\textsf{I-X}} - z_{j+1}^{\textsf{B-X}} \geq 0 \end{aligned} \tag{9}$$

Bilingual constraints are added in Eq. (10):

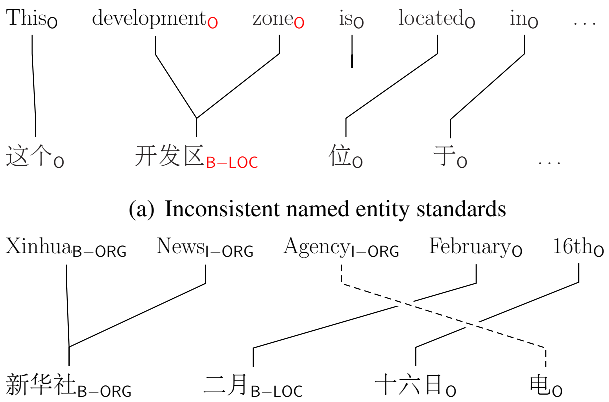
$$\forall (i,j) \in A, \forall \textsf{X} : z_i^{\textsf{B-X}} + z_i^{\textsf{I-X}} = z_j^{\textsf{B-X}} + z_j^{\textsf{I-X}} \tag{10}$$

where $A = \{(i,j)\}$ is the word alignment pair set, i.e., the i^{th} Chinese word and the j^{th} English word were aligned together. Chinese word i is tagged with a named entity type X ($z_i^{\textsf{B-X}} + z_i^{\textsf{I-X}} = 1$), *iff* English word j is tagged with X ($z_j^{\textsf{B-X}} + z_j^{\textsf{I-X}} = 1$). Therefore, these *hard* bilingual constraints guarantee that when two words are aligned, they are tagged with the same named entity label.

However, in practice, aligned word pairs do not always have the same named entity label because of the difference in annotation standards across different languages. For example, in Figure 2(a), the Chinese word "开发区" is a location. However, it is aligned to the words, "development" and "zone", which are not named entity labels in English. Word alignment error is another problem that can cause violation of hard constraints. In Figure 2(b), the English word "Agency" is wrongly aligned with the Chinese word "电 (report)". Thus, these two words cannot be tagged with the same label.

To address these two problems, we present a probabilistic model for bilingual NER which can lead to

[3] The prefix B- and I- are ignored.



(a) Inconsistent named entity standards

(b) Word alignment error

Figure 2: Errors of hard bilingual constraints method.

an optimization problem with two *soft* bilingual constraints:

1) allow word-aligned pairs to have different named entity labels; 2) consider word alignment probabilities to reduce the influence of wrong word alignments.

3.2 Soft Constraints with Label Uncertainty

The new probabilistic model for bilingual NER is:

$$\begin{aligned} P(\mathbf{y}_c, \mathbf{y}_e|\mathbf{x}_c, \mathbf{x}_e, A) &= \frac{P(\mathbf{y}_c, \mathbf{y}_e, \mathbf{x}_c, \mathbf{x}_e, A)}{P(\mathbf{x}_c, \mathbf{x}_e, A)} \\ = \frac{P(\mathbf{y}_c, \mathbf{x}_c, \mathbf{x}_e, A)}{P(\mathbf{x}_c, \mathbf{x}_e, A)} &\cdot \frac{P(\mathbf{y}_e, \mathbf{x}_c, \mathbf{x}_e, A)}{P(\mathbf{x}_c, \mathbf{x}_e, A)} \\ &\cdot \frac{P(\mathbf{y}_c, \mathbf{y}_e, \mathbf{x}_c, \mathbf{x}_e, A) P(\mathbf{x}_c, \mathbf{x}_e, A)}{P(\mathbf{y}_c, \mathbf{x}_c, \mathbf{x}_e, A) P(\mathbf{y}_e, \mathbf{x}_c, \mathbf{x}_e, A)} \qquad (11) \\ \approx P(\mathbf{y}_c|\mathbf{x}_c) P(\mathbf{y}_e|\mathbf{x}_e) &\frac{P(\mathbf{y}_c, \mathbf{y}_e|A)}{P(\mathbf{y}_c|A) P(\mathbf{y}_e|A)} \qquad (12) \end{aligned}$$

where $\mathbf{y}_c$ and $\mathbf{y}_e$ respectively denotes Chinese and English named entity output sequences. A is the set of word alignment pairs.

If we assume that named entity label assignments in Chinese is only dependent on the observed Chinese sentence, then we can drop the A and $\mathbf{x}_e$ term in the first part of Eq. (11), and arrive at the first part of Eq. (12); similarly we can use the same assumption to derive the second factor in Eq. (12) for English; alternatively, if we assume the named entity label assignments are only dependent on the cross-lingual word associations via word alignment, then we can drop $\mathbf{x}_c$ and $\mathbf{x}_e$ terms in the third factor of Eq. (11) and arrive at the third factor of Eq. (12). These factors represent the two major sources of information in the model: monolingual surface observation, and cross-lingual word associations.

The first two factors of Eq. (12) can be further decomposed into the product of probabilities of all words in each language sentence like Eq. (3).

Assuming that the labels are independent between different word alignment pairs, then the last factor of Eq. (12) can be decomposed into:

$$\begin{aligned} \frac{P(\mathbf{y}_c, \mathbf{y}_e|A)}{P(\mathbf{y}_c|A)P(\mathbf{y}_e|A)} &= \prod_{a \in A} \frac{P(y_{c_a} y_{e_a})}{P(y_{c_a})P(y_{e_a})} \\ &= \prod_{a \in A} \lambda_a^{y_c y_e} \qquad (13) \end{aligned}$$

where y_{c_a} and y_{e_a} respectively denotes Chinese and English named entity labels in a word alignment pair a. $\lambda^{y_c y_e} = \frac{P(y_c y_e)}{P(y_c)P(y_e)}$ is the pointwise mutual information (PMI) score between a Chinese named entity label y_c and an English named entity label y_e. If $y_c = y_e$, then the score will be high; otherwise the score will be low. A number of methods for calculating the scores are provided at the end of this section.

We use ILP to maximize Eq. (12). The new objective function is expressed as follow:

$$\begin{aligned} \max \sum_{i=1}^{|\mathbf{x}_c|} \sum_{y \in Y} z_i^y \log P_i^y + \sum_{j=1}^{|\mathbf{x}_e|} \sum_{y \in Y} z_j^y \log P_j^y \\ + \sum_{a \in A} \sum_{y_c \in Y} \sum_{y_e \in Y} z_a^{y_c y_e} \log \lambda_a^{y_c y_e} \qquad (14) \end{aligned}$$

where $z_a^{y_c y_e}$ is an indicator that equals 1 *iff* the Chinese named entity label is y_c and the English named entity label is y_e, given a word alignment pair a; otherwise, $z_a^{y_c y_e}$ is 0.

Monolingual constraints such as Eqs. (8) and (9) need to be added. In addition, one and only one possible named entity label pair exists for a word alignment pair. This condition can be expressed as the following constraints:

$$\forall a \in A : \sum_{y_c \in Y} \sum_{y_e \in Y} z_a^{y_c y_e} = 1 \qquad (15)$$

When the label pair of a word alignment pair is determined, the corresponding monolingual named entity labels can also be identified. This rule can be expressed by the following constraints:

$$\forall a = (i, j) \in A : z_a^{y_c y_e} \leq z_i^{y_c}, z_a^{y_c y_e} \leq z_j^{y_e} \quad (16)$$

Thus, if $z_a^{y_c y_e} = 1$, then $z_i^{y_c}$ and $z_j^{y_e}$ must be both equal to 1. Here, the i^{th} Chinese word and the j^{th} English word are aligned together.

In contrast to hard bilingual constraints, inconsistent named entity labels for an aligned word pair are allowed in soft bilingual constraints, but are given lower $\lambda^{y_c y_e}$ scores.

To calculate the $\lambda^{y_c y_e}$ score, an annotated bilingual NER corpus is consulted. We count from all word alignment pairs the number of times y_c and y_e occur together ($C(y_c y_e)$) and separately ($C(y_c)$ and $C(y_e)$). Afterwards, $\lambda^{y_c y_e}$ is calculated with maximum likelihood estimation as follows:

$$\lambda^{y_c y_e} = \frac{P(y_c y_e)}{P(y_c)P(y_e)} = \frac{N \times C(y_c y_e)}{C(y_c)C(y_e)} \quad (17)$$

where N is the total number of word alignment pairs.

However, in this paper, we assume that no named entity annotated bilingual corpus is available. Thus, the above method is only used as `Oracle`. A realistic method for calculating the $\lambda^{y_c y_e}$ score requires the use of two initial monolingual NER models, such as baseline CRF, to predict named entity labels for each language on an unannotated bitext. We count from this automatically tagged corpus the statistics mentioned above. This method is henceforth referred to as `Auto`.

A simpler approach is to manually set the value of $\lambda^{y_c y_e}$: if $y_c = y_e$ then we assign a larger value to $\lambda^{y_c y_e}$; else we assign an ad-hoc smaller value. In fact, if we set $\lambda^{y_c y_e} = 1$ *iff* $y_c = y_e$; otherwise, $\lambda^{y_c y_e} = 0$, then the soft constraints backs off to hard constraints. We refer to this set of soft constraints as `Soft-label`.

3.3 Constraints with Alignment Uncertainty

So far, we assumed that a word alignment set A is known. In practice, only the word alignment probability P_a for each word pair a is provided. We can set a threshold θ for P_a to tune the set A: $a \in A$ *iff* $P_a \geq \theta$. This condition can be regarded as a kind of *hard word alignment*. However, the following problem exists: the smaller the θ, the noisier the word alignments are; the larger the θ, the more possible word alignments are lost. To ameliorate this problem, we introduce another set of soft bilingual constraints.

We can re-express Eq. (13) as follows:

$$\prod_{a \in A} \lambda_a^{y_c y_e} = \prod_{a \in \mathscr{A}} (\lambda_a^{y_c y_e})^{I_a} \quad (18)$$

where $\mathscr{A}$ is the set of all word pairs between two languages. $I_a = 1$ *iff* $P_a \geq \theta$; otherwise, $I_a = 0$.

We can then replace the hard indicator I_a with the word alignment probability P_a, Eq. (14) is then transformed into the following equation:

$$\max \sum_i^{|W_c|} \sum_{y \in Y} z_i^y \log P_i^y + \sum_j^{|W_e|} \sum_{y \in Y} z_j^y \log P_j^y + \sum_{a \in \mathscr{A}} \sum_{y_c \in Y} \sum_{y_e \in Y} z_a^{y_c y_e} P_a \log \lambda_a^{y_c y_e} \quad (19)$$

We name the set of constraints above `Soft-align`, which has the same constraints as `Soft-label`, i.e., Eqs. (8), (9), (15) and (16).

4 Experimental Setup

We conduct experiments on the latest OntoNotes 4.0 corpus (LDC2011T03). OntoNotes is a large, manually annotated corpus that contains various text genres and annotations, such as part-of-speech tags, named entity labels, syntactic parse trees, predicate-argument structures and co-references (Hovy et al., 2006). Aside from English, this corpus also contains several Chinese and Arabic corpora. Some of these corpora contain bilingual parallel documents. We used the Chinese-English parallel corpus with named entity labels as our development and test data. This corpus includes about 400 document pairs (chtb_0001-0325, ectb_1001-1078). We used odd-numbered documents as development data and even-numbered documents as test data. We used all other portions of the named entity annotated corpus as training data for the monolingual systems. There were a total of ∼660 Chinese documents (∼16k sentences) and ∼1,400 English documents (∼39k sentences). OntoNotes annotates 18 named entity types, which include person, location, date and money. In this paper, we selected the four most common

Chinese NER Templates
00: 1 (class bias param)
01: $w_{i+k}, -1 \le k \le 1$
02: $w_{i+k-1} \circ w_{i+k}, 0 \le k \le 1$
03: $\mathsf{shape}(w_{i+k}), -4 \le k \le 4$
04: $\mathsf{prefix}(w_i, k), 1 \le k \le 4$
05: $\mathsf{prefix}(w_{i-1}, k), 1 \le k \le 4$
06: $\mathsf{suffix}(w_i, k), 1 \le k \le 4$
07: $\mathsf{suffix}(w_{i-1}, k), 1 \le k \le 4$
08: $\mathsf{radical}(w_i, k), 1 \le k \le \mathsf{len}(w_i)$
Unigram Features
$y_i \circ$ 00 – 08
Bigram Features
$y_{i-1} \circ y_i \circ$ 00 – 08

Table 1: Basic features of Chinese NER.

named entity types, i.e., PER (Person), LOC (Location), ORG (Organization) and GPE (Geo-Political Entities), and discarded the others.

Since the bilingual corpus is only aligned at the document level, we performed sentence alignment using the Champollion Tool Kit (CTK).[4] After removing sentences with no aligned sentence, a total of 8,249 sentence pairs were retained.

We used the BerkeleyAligner,[5] to produce word alignments over the sentence-aligned datasets. BerkeleyAligner also gives posterior probabilities P_a for each aligned word pair.

We used the CRF-based Stanford NER tagger (using Viterbi decoding) as our baseline monolingual NER tool.[6] English features were taken from Finkel et al. (2005). Table 1 lists the basic features of Chinese NER, where $\circ$ means string concatenation and y_i is the named entity label of the i^{th} word w_i. Moreover, $\mathsf{shape}(w_i)$ is the shape of w_i, such as date and number. $\mathsf{prefix/suffix}(w_i, k)$ denotes the k-characters prefix/suffix of w_i. $\mathsf{radical}(w_i, k)$ denotes the radical of the k^{th} Chinese character of w_i.[7] $\mathsf{len}(w_i)$ is the number of Chinese characters in w_i.

To make the baseline CRF taggers stronger, we added word clustering features to improve generalization over unseen data for both Chinese and English. Word clustering features have been successfully used in several English tasks, including NER (Miller et al., 2004) and dependency parsing (Koo et al., 2008). To our knowledge, this work is the first use of word clustering features for Chinese NER. A C++ implementation of the Brown word clustering algorithms (Brown et al., 1992) was used to obtain the word clusters (Liang, 2005).[8] Raw text was obtained from the fifth edition of Chinese Gigaword (LDC2011T13). One million paragraphs from Xinhua news section were randomly selected, and the Stanford Word Segmenter with LDC standard was applied to segment Chinese text into words.[9] About 46 million words were obtained which were clustered into 1,000 word classes.

5 Threshold Tuning

During development, we tuned the word alignment probability thresholds to find the best value. Figure 3 shows the performance curves.

When the word alignment probability threshold θ is set to 0.9, the hard bilingual constraints perform quite well for both Chinese and English. But as the thresholds value gets smaller, and more noisy word alignments are introduced, we see the hard bilingual constraints method starts to perform badly.

In `Soft-label` setting, where inconsistent label assignments within aligned word pairs are allowed but penalized, different languages have different optimal threshold values. For example, Chinese has an optimal threshold of 0.7, whereas English has 0.2. Thus, the optimal thresholds for different languages need to be selected with care when `Soft-label` is applied in practice.

`Soft-align` eliminates the need for careful tuning of word alignment thresholds, and therefore can be more easily used in practice. Experimental results of `Soft-align` confirms our hypothesis – the performance of both Chinese and English NER systems improves with decreasing threshold. However, we can still improve efficiency by setting a low threshold to prune away very unlikely word alignments. We set the threshold to 0.1 for `Soft-align` to increase speed, and we observed very minimal performance lost when doing so.

We also found that automatically estimated bilin-

[4] champollion.sourceforge.net

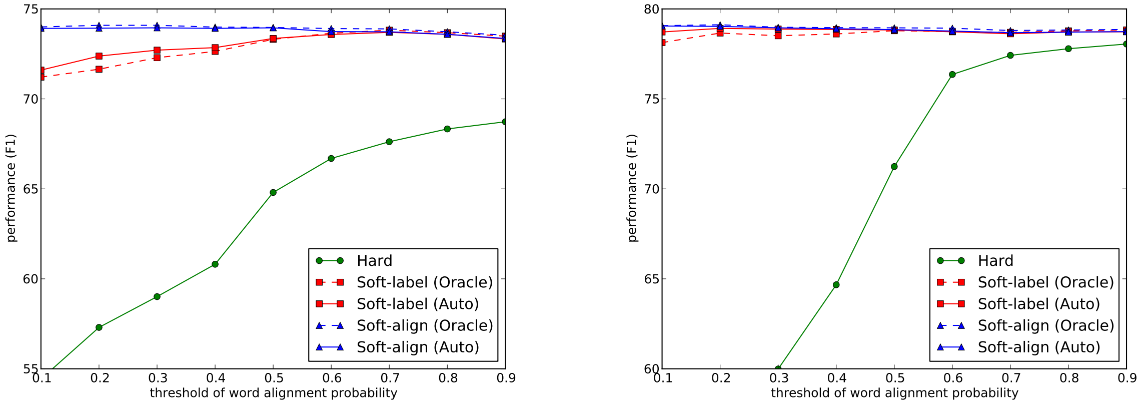
[5] code.google.com/p/berkeleyaligner

[6] nlp.stanford.edu/software/CRF-NER.shtml, which includes our English and Chinese NER implementations.

[7] The radical of a Chinese character can be found at: www.unicode.org/charts/unihan.html

[8] github.com/percyliang/brown-cluster

[9] nlp.stanford.edu/software/segmenter.shtml



(a) Chinese

(b) English

Figure 3: Performance curves of different bilingual constraints methods on development set.

gual label PMI scores (`Auto`) gave comparable results to `Oracle`. Therefore this technique is effective for computing the PMI scores, avoiding the need of manually annotating named entity bilingual corpus.

6 Bilingual NER Results

The main results on Chinese and English test sets with the optimal word alignment threshold for each method are shown in Table 2.

The CRF-based Chinese NER with and without word clustering features are compared here. The word clustering features significantly ($p < 0.01$) improved the performance of Chinese NER, [10] giving us a strong Chinese NER baseline.[11] The effectiveness of word clustering for English NER has been proved in previous work.

The performance of ILP with only monolingual constraints is quite comparable with the CRF results, especially on English. The greater ILP performance on English is probably due to more accurate marginal probabilities estimated by the English CRF model.

The ILP model with hard bilingual constraints gives a slight performance improvement on Chinese, but affects performance negatively on English. Once we introduced labeling uncertainties into the `Soft-label` bilingual constraints, we see a very significant ($p < 0.01$) performance boost on Chinese. This method also improves the recall on English, with a smaller decrease in precision. Overall, it improves English F_1 score by about 0.4%, which is unfortunately not statistically significant.

Compared with `Soft-label`, the final `Soft-align` method can further improve performance on both Chinese and English. This is likely to be because: 1) `Soft-align` includes more word alignment pairs, thereby improving recall; and 2) uses probabilities to cut wrong word alignments, thereby improving precision. In particular, compared with the strong CRF baseline, the gain on Chinese side is almost 5.5% in absolute F_1 score.

Decoding efficiency of different methods are shown in the last column of Table 2.[12] Compared with Viterbi decoding in CRF, monolingual ILP decoding is about 2.3 times slower. Bilingual ILP decoding, with either hard or soft constraints, is significantly slower than the monolingual methods. The reason is that the number of monolingual ILP constraints doubles, and there are additionally many more bilingual constraints. The difference in speed between the `Soft-label` and `Soft-align` methods is attributed to the difference in number of word alignment pairs.

[10] We use paired bootstrap resampling significance test (Efron and Tibshirani, 1993).

[11] To the best of our knowledge, there was no performance report of state-of-the-art NER results on the latest OntoNotes dataset.

[12] CPU: Intel Xeon E5-2660 2.20GHz. And the speed calculation of ILP inference methods exclude the time needed to obtain marginal probabilities from the CRF models.

	Chinese			English			Speed
	P	R	F_1	P	R	F_1	sent / s
CRF (No Cluster)	74.74	56.17	64.13	–	–	–	–
CRF (Word Cluster)	76.90	63.32	69.45	**82.95**	76.67	79.68	317.3
Monolingual ILP	76.20	63.06	69.01	82.88	76.68	79.66	138.0
Hard	74.38	65.78	69.82	82.66	75.36	78.84	21.1
`Soft-label` (`Auto`)	77.37	71.14	74.13	81.36	**78.74**	80.03	5.9
`Soft-align` (`Auto`)	**77.71**	**72.51**	**75.02**	81.94	78.35	**80.10**	1.5

Table 2: Results on bilingual parallel test set.

Since each sentence pair can be decoded independently, parallelization the decoding process can result in significant speedup.

7 Semi-supervised NER Results

The above results show the usefulness of our method in a bilingual setting, where we are presented with sentence aligned data, and are tagging both languages at the same time. To have a greater impact on general monolingual NER systems, we employ a semi-supervised learning setting. First, we tag a large amount of unannotated bitext with our bilingual constraint-based NER tagger. Then we mix the automatically tagged results with the original monolingual Chinese training data to train a new model.

Our bitext is derived from the Chinese-English part of the Foreign Broadcast Information Service corpus (FBIS, LDC2003E14). The best performing bilingual model `Soft-align` with threshold $\theta = 0.1$ was used under the same experimental setting as described in Section 4

Method	#sent	P	R	F_1
CRF	$\sim$16k	76.90	63.32	69.45
Semi	10k	77.60	66.51	71.62
	20k	77.28	67.26	71.92
	40k	77.40	67.81	72.29
	80k	77.44	68.64	72.77

Table 3: Semi-supervised results on Chinese test set.

Table 3 shows that the performance of the semi-supervised method improves with more additional data. We simply appended these data to the original training data. We also have done the experiments to down weight the additional training data by duplicating the original training data. There was some slight improvements, but not very significant. Finally, when we add 80k sentences, the F_1 score is improved by 3.32%, which is significantly ($p < 0.01$) better than the baseline, and most of the contribution comes from recall improvement.

Before the end of experimental section, let us summarize the usage of different kinds of data resources used in our experiments, as shown in Table 4, where ✓ and × denote whether the corresponding resources are required. In the bilingual case, only the monolingual named entity annotated data (NE-mono) is necessary to train a monolingual NER tagger during training. Unannotated bitext (Bitext) is required by the word aligner and our bilingual NER tagger. Named entity annotated bitext (NE-bitext) is used to evaluate our bilingual model during the test. In the semi-supervised case, besides the original NE-mono data, the Bitext is used as input to our bilingual NER tagger to product additional training data. To evaluate the final NER model, only NE-mono is needed.

		NE-mono	Bitext	NE-bitext
Bilingual	train	✓	×	×
	test	×	✓	✓
Semi	train	✓	✓	×
	test	✓	×	×

Table 4: Summarization of the data resource usage

8 Related Work

Previous work explored the use of bilingual corpora to improve existing monolingual analyzers. Huang et al. (2009) proposed methods to improve parsing performance using bilingual parallel corpus. Li et al. (2012) jointly labeled bilingual named entities with a cyclic CRF model, where approximate inference was done using loopy belief propagation.

These methods require manually annotated bilingual corpora, which are expensive to construct, and hard to obtain. Kim et al. (2012) proposed a method of labeling bilingual corpora with named entity labels automatically based on Wikipedia. However, this method is restricted to topics covered by Wikipedia.

Similar to our work, Burkett et al. (2010) also assumed that annotated bilingual corpora are scarce. Beyond the difference discussed in Section 1, their re-ranking strategy may lose the correct named entity results if they are not included in the top-N outputs. Furthermore, we consider the word alignment probabilities in our method which can reduce the influence of word alignment errors. Finally, we test our method on a large standard publicly available corpus (8,249 sentences), while they used a much smaller (200 sentences) manually annotated bilingual NER corpus for results validation.

In addition to bilingual corpora, bilingual dictionaries are also useful resources. Huang and Vogel (2002) and Chen et al. (2010) proposed approaches for extracting bilingual named entity pairs from unannotated bitext, in which verification is based on bilingual named entity dictionaries. However, large-scale bilingual named entity dictionaries are difficult to obtain for most language pairs.

Yarowsky and Ngai (2001) proposed a projection method that transforms high-quality analysis results of one language, such as English, into other languages on the basis of word alignment. Das and Petrov (2011) applied the above idea to part-of-speech tagging with a more complex model. Fu et al. (2011) projected English named entities onto Chinese by carefully designed heuristic rules. Although this type of method does not require manually annotated bilingual corpora or dictionaries, errors in source language results, wrong word alignments and inconsistencies between the languages limit application of this method.

Constraint-based monolingual methods by using ILP have been successfully applied to many natural language processing tasks, such as Semantic Role Labeling (Punyakanok et al., 2004), Dependency Parsing (Martins et al., 2009) and Textual Entailment (Berant et al., 2011). Zhuang and Zong (2010) proposed a joint inference method for bilingual semantic role labeling with ILP. However, their approach requires training an alignment model with a manually annotated corpus.

9 Conclusions

We proposed a novel ILP based inference algorithm with bilingual constraints for NER. This method can jointly infer bilingual named entities without using any annotated bilingual corpus. We investigate various bilingual constraints: hard and soft constraints. Out empirical study on large-scale OntoNotes Chinese-English parallel NER data showed that `Soft-align` method, which allows inconsistent named entity labels between two aligned words and considers word alignment probabilities without requiring tuning any parameters, can significantly improve over the performance of a strong Chinese NER baseline. Our work is the first to evaluate performance on a large-scale standard dataset. Finally, we can also improve monolingual Chinese NER performance significantly, by combining the original monolingual training data with new NER data obtained from bitext tagged by our method. The final ILP-based bilingual NER tagger with soft constraints is publicly available at: `github.com/carfly/bi_ilp`

Future work could apply the bilingual constraint-based method to other tasks, such as part-of-speech tagging and relation extraction.

Acknowledgments

The authors would like to thank Rob Voigt and the three anonymous reviewers for their valuable comments and suggestions. We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 61133012, the National "863" Project via grant 2011AA01A207 and 2012AA011102, the Ministry of Education Research of Social Sciences Youth funded projects via grant 12YJCZH304, the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181 and the support of the DARPA Broad Operational Language Translation (BOLT) program through IBM.

Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 610–619, Portland, Oregon, USA, June. Association for Computational Linguistics.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 46–54, Uppsala, Sweden, July. Association for Computational Linguistics.

Yufeng Chen, 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, Sweden, July. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. 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*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

B. Efron and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ruiji Fu, Bing Qin, and Ting Liu. 2011. Generating chinese named entity data from a parallel corpus. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 264–272, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fei Huang and Stephan Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI 2002, Washington, DC, USA. IEEE Computer Society.

Liang Huang, Wenbin Jiang, and Qun Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1222–1231, Singapore, August. Association for Computational Linguistics.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 694–702, Jeju Island, Korea, July. Association for Computational Linguistics.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio, June. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, Honolulu, Hawaii, October.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, MIT.

Andre Martins, Noah Smith, and Eric Xing. 2009. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 342–350, Suntec, Singapore, August. Association for Computational Linguistics.

Bernard Mérialdo. 1994. Tagging english text with a probabilistic model. *Comput. Linguist.*, 20(2):155–171.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of Coling 2004*, pages 1346–1352, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Dan Roth and Wen-tau Yih. 2005. Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 736–743, New York, NY, USA. ACM.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tao Zhuang and Chengqing Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 304–314, Cambridge, MA, October. Association for Computational Linguistics.