Joint Word Alignment and Bilingual Named Entity Recognition Using Dual Decomposition

Mengqiu Wang
Stanford University
Stanford, CA 94305
mengqiu@cs.stanford.edu

Wanxiang Che
Harbin Institute of Technology
Harbin, China, 150001
car@ir.hit.edu.cn

Christopher D. Manning
Stanford University
Stanford, CA 94305
manning@cs.stanford.edu

Abstract

Translated bi-texts contain complementary language cues, and previous work on Named Entity Recognition (NER) has demonstrated improvements in performance over monolingual taggers by promoting agreement of tagging decisions between the two languages. However, most previous approaches to bilingual tagging assume word alignments are given as fixed input, which can cause cascading errors. We observe that NER label information can be used to correct alignment mistakes, and present a graphical model that performs bilingual NER tagging jointly with word alignment, by combining two monolingual tagging models with two unidirectional alignment models. We introduce additional cross-lingual edge factors that encourage agreements between tagging and alignment decisions. We design a dual decomposition inference algorithm to perform joint decoding over the combined alignment and NER output space. Experiments on the OntoNotes dataset demonstrate that our method yields significant improvements in both NER and word alignment over state-of-the-art monolingual baselines.

1 Introduction

We study the problem of Named Entity Recognition (NER) in a bilingual context, where the goal is to annotate parallel bi-texts with named entity tags. This is a particularly important problem for machine translation (MT) since entities such as person names, locations, organizations, etc. carry much of the information expressed in the source sentence. Recognizing them provides useful information for phrase detection and word sense disambiguation (e.g., “melody” as in a female name has a different translation from the word “melody” in a musical sense), and can be directly leveraged to improve translation quality (Babych and Hartley, 2003). We can also automatically construct a named entity translation lexicon by annotating and extracting entities from bi-texts, and use it to improve MT performance (Huang and Vogel, 2002; Al-Onaizan and Knight, 2002). Previous work such as Burkett et al. (2010b), Li et al. (2012) and Kim et al. (2012) have also demonstrated that bi-texts annotated with NER tags can provide useful additional training sources for improving the performance of standalone monolingual taggers.

Because human translation in general preserves semantic equivalence, bi-texts represent two perspectives on the same semantic content (Burkett et al., 2010b). As a result, we can find complementary cues in the two languages that help to disambiguate named entity mentions (Brown et al., 1991). For example, the English word “Jordan” can be either a last name or a country. Without sufficient context it can be difficult to distinguish the two; however, in Chinese, these two senses are disambiguated: “乔丹” as a last name, and “约旦” as a country name.

In this work, we first develop a bilingual NER model (denoted as BI-NER) by embedding two monolingual CRF-based NER models into a larger undirected graphical model, and introduce additional edge factors based on word alignment (WA). Because the new bilingual model contains many cyclic cliques, exact inference is intractable. We employ a dual decomposition (DD) inference algorithm (Bertsekas, 1999; Rush et al., 2010) for performing approximate inference. Unlike most

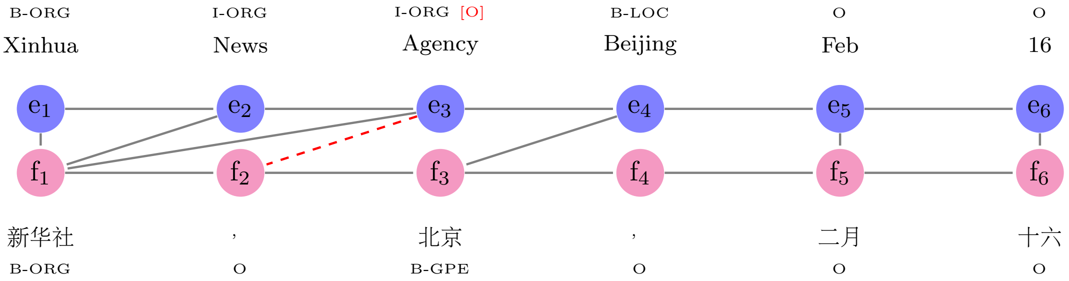

Figure 1: Example of NER labels between two word-aligned bilingual parallel sentences. The [O] tag is an example of a wrong tag assignment. The dashed alignment link between e_3 and f_2 is an example of alignment error.

previous applications of the DD method in NLP, where the model typically factors over two components and agreement is to be sought between the two (Rush et al., 2010; Koo et al., 2010; DeNero and Macherey, 2011; Chieu and Teow, 2012), our method decomposes the larger graphical model into many overlapping components where each alignment edge forms a separate factor. We design clique potentials over the alignment-based edges to encourage entity tag agreements. Our method does not require any manual annotation of word alignments or named entities over the bilingual training data.

The aforementioned BI-NER model assumes fixed alignment input given by an underlying word aligner. But the entity span and type predictions given by the NER models contain complementary information for correcting alignment errors. To capture this source of information, we present a novel extension that combines the BI-NER model with two uni-directional HMM-based alignment models, and perform joint decoding of NER and word alignments. The new model (denoted as BI-NER-WA) factors over five components: one NER model and one word alignment model for each language, plus a joint NER-alignment model which not only enforces NER label agreements but also facilitates message passing among the other four components. An extended DD decoding algorithm is again employed to perform approximate inference.

We give a formal definition of the Bi-NER model in Section 2, and then move to present the Bi-NER-WA model in Section 3.

2 Bilingual NER by Agreement

The inputs to our models are parallel sentence pairs (see Figure 1 for an example in English and Chinese). We denote the sentences as $\mathbf{e}$ (for English) and $\mathbf{f}$ (for Chinese). We assume access to two monolingual linear-chain CRF-based NER models that are already trained. The English-side CRF model assigns the following probability for a tag sequence $\mathbf{y}^e$:

$$P_{CRF_e}(\mathbf{y}^e|\mathbf{e}) = \frac{\prod\limits_{v_i \in \mathcal{V}^e} \psi(v_i) \prod\limits_{(v_i,v_j) \in \mathcal{D}^e} \omega(v_i, v_j)}{\mathcal{Z}^e(\mathbf{e})}$$

where $\mathcal{V}^e$ is the set of vertices in the CRF and $\mathcal{D}^e$ is the set of edges. $\psi(v_i)$ and $\omega(v_i, v_j)$ are the node and edge clique potentials, and $\mathcal{Z}^e(\mathbf{e})$ is the partition function for input sequence $\mathbf{e}$ under the English CRF model. We let $k(\mathbf{y}^e)$ be the un-normalized log-probability of tag sequence y^e, defined as:

$$k(\mathbf{y}^e) = \log \left(\prod_{v_i \in \mathcal{V}^e} \psi(v_i) \prod_{(v_i,v_j) \in \mathcal{D}^e} \omega(v_i, v_j) \right)$$

Similarly, we define model P_{CRF_f} and un-normalized log-probability $l(\mathbf{y}^f)$ for Chinese.

We also assume that a set of word alignments ($\mathcal{A} = \{(i,j) : e_i \leftrightarrow f_j\}$) is given by a word aligner and remain fixed in our model.

For clarity, we assume $\mathbf{y}^e$ and $\mathbf{y}^f$ are binary variables in the description of our algorithms. The extension to the multi-class case is straight-forward and does not affect the core algorithms.

2.1 Hard Agreement

We define a BI-NER model which imposes hard agreement of entity labels over aligned word pairs. At inference time, we solve the following optimization problem:

$$
\begin{aligned}
&\max_{\mathbf{y}^e,\mathbf{y}^f} \log\left(P_{CRF_e}\left(\mathbf{y}^e\right)\right) + \log\left(P_{CRF_f}\left(\mathbf{y}^f\right)\right) \\
=&\max_{\mathbf{y}^e,\mathbf{y}^f} k(\mathbf{y}^e) + l(\mathbf{y}^f) - \log \mathcal{Z}_e(\mathbf{e}) - \log \mathcal{Z}_f(\mathbf{f}) \\
\simeq&\max_{\mathbf{y}^e,\mathbf{y}^f} k(\mathbf{y}^e) + l(\mathbf{y}^f) \\
&\ni y_i^e = y_j^f \;\; \forall (i,j) \in \mathcal{A}
\end{aligned}
$$

We dropped the $\mathcal{Z}_e(\mathbf{e})$ and $\mathcal{Z}_f(\mathbf{f})$ terms because they remain constant at inference time.

The Lagrangian relaxation of this term is:

$$
\begin{aligned}
&L\left(\mathbf{y}^e, \mathbf{y}^f, \mathbf{U}\right) = \\
&\quad k\left(\mathbf{y}^e\right) + l\left(\mathbf{y}^f\right) + \sum_{(i,j)\in\mathcal{A}} u(i,j)\left(y_i^e - y_j^f\right)
\end{aligned}
$$

where $u(i,j)$ are the Lagrangian multipliers.

Instead of solving the Lagrangian directly, we can form the dual of this problem and solve it using dual decomposition (Rush et al., 2010):

$$
\begin{aligned}
\min_{\mathbf{U}} \Bigg(&\max_{y^e} \left[k\left(\mathbf{y}^e\right) + \sum_{(i,j)\in\mathcal{A}} u(i,j) y_i^e \right] \\
&+ \max_{y^f} \left[l\left(\mathbf{y}^f\right) - \sum_{(i,j)\in\mathcal{A}} u(i,j) y_j^f \right] \Bigg)
\end{aligned}
$$

Similar to previous work, we solve this DD problem by iteratively updating the sub-gradient as depicted in Algorithm 1. T is the maximum number of iterations before early stopping, and α_t is the learning rate at time t. We adopt a learning rate update rule from Koo et al. (2010) where α_t is defined as $\frac{1}{N}$, where N is the number of times we observed a consecutive dual value increase from iteration 1 to t.

A thorough introduction to the theoretical foundations of dual decomposition algorithms is beyond the scope of this paper; we encourage unfamiliar readers to read Rush and Collins (2012) for a full tutorial.

2.2 Soft Agreement

The previously discussed hard agreement model rests on the core assumption that aligned words must have identical entity tags. In reality, however, this assumption does not always hold. Firstly, assuming words are correctly aligned, their entity tags may not agree due to inconsistency in annotation standards. In Figure 1, for example, the word "Beijing" can be either a Geo-Political Entity (GPE) or a location. The Chinese annotation standard may enforce that "Beijing" should always be tagged as GPE when it is mentioned in isolation, while the English standard may require the annotator to judge based on word usage context. The assumption in the hard agreement model can also be violated if there are word alignment errors.

Algorithm 1 DD inference algorithm for hard agreement model.

```
$\forall (i,j) \in \mathcal{A}:\ u(i,j) = 0$
for $t \leftarrow 1$ to $T$ do
    $\mathbf{y}^{\mathbf{e}*} \leftarrow \operatorname{argmax} k\left(\mathbf{y}^e\right) + \sum_{(i,j)\in A} u(i,j) y_i^e$
    $\mathbf{y}^{\mathbf{f}*} \leftarrow \operatorname{argmax} l\left(\mathbf{y}^f\right) - \sum_{(i,j)\in A} u(i,j) y_j^f$
    if $\forall (i,j) \in \mathcal{A}$: $y_i^{e*} = y_j^{f*}$ then
        return $\left(\mathbf{y}^{e*}, \mathbf{y}^{f*}\right)$
    end if
    for all $(i,j) \in \mathcal{A}$ do
        $u(i,j) \leftarrow u(i,j) + \alpha_t \left(y_j^{f*} - y_i^{e*}\right)$
    end for
end for
return $\left(\mathbf{y}^{\mathbf{e}*}_{(\mathbf{T})}, \mathbf{y}^{\mathbf{f}*}_{(\mathbf{T})}\right)$
```

In order to model this uncertainty, we extend the two previously independent CRF models into a larger undirected graphical model, by introducing a cross-lingual edge factor $\phi(i,j)$ for every pair of word positions $(i,j) \in \mathcal{A}$. We associate a clique potential function $h_{(i,j)}(y_i^e, y_j^f)$ for $\phi(i,j)$:

$$
h_{(i,j)}\left(y_i^e, y_j^f\right) = \operatorname{pmi}\left(y_i^e, y_j^f\right)^{\hat{P}(e_i, f_j)}
$$

where $\operatorname{pmi}(y_i^e, y_j^f)$ is the point-wise mutual information (PMI) of the tag pair, and we raise it to the power of a posterior alignment probability $\hat{P}(e_i, f_j)$. For a pair of NEs that are aligned with low probability, we cannot be too sure about the association of the two NEs, therefore the model should not impose too much influence from the bilingual agreement model; instead, we will let the monolingual NE models make their decisions, and trust that those are the best estimates we can come up with when we do not have much confidence in their bilingual association. The use of the posterior alignment probability facilitates this purpose.

Initially, each of the cross-lingual edge factors will attempt to assign a pair of tags that has the highest PMI score, but if the monolingual taggers do not agree, a penalty will start accumulating over this pair, until some other pair that agrees better with the monolingual models takes the top spot.

Simultaneously, the monolingual models will also be encouraged to agree with the cross-lingual edge factors. This way, the various components effectively trade penalties indirectly through the cross-lingual edges, until a tag sequence that maximizes the joint probability is achieved.

Since we assume no bilingually annotated NER corpus is available, in order to get an estimate of the PMI scores, we first tag a collection of unannotated bilingual sentence pairs using the monolingual CRF taggers, and collect counts of aligned entity pairs from this auto-generated tagged data.

Each of the $\phi(i,j)$ edge factors (e.g., the edge between node f_3 and e_4 in Figure 1) overlaps with each of the two CRF models over one vertex (e.g., f_3 on Chinese side and e_4 on English side), and we seek agreement with the Chinese CRF model over tag assignment of f_j, and similarly for e_i on English side. In other words, no direct agreement between the two CRF models is enforced, but they both need to agree with the bilingual edge factors.

The updated optimization problem becomes:

$$\max_{\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}}\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}}\mathbf{y}^{\mathbf{e}^{(\mathbf{h})}}\mathbf{y}^{\mathbf{f}^{(\mathbf{h})}}} k\left(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}}\right) + l\left(\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}}\right) + \sum_{(i,j)\in\mathcal{A}} h_{(i,j)}\left(y_i^{e^{(h)}}, y_j^{f^{(h)}}\right)$$

$$\ni \forall (i,j)\in\mathcal{A}\colon \left(y_i^{e^{(k)}} = y_i^{e^{(h)}}\right) \wedge \left(y_j^{f^{(l)}} = y_j^{f^{(h)}}\right)$$

where the notation $y_i^{e^{(k)}}$ denotes tag assignment to word e_i by the English CRF and $y_i^{e^{(h)}}$ denotes assignment to word e_i by the bilingual factor; $y_j^{f^{(l)}}$ denotes the tag assignment to word f_j by the Chinese CRF and $y_j^{f^{(h)}}$ denotes assignment to word f_j by the bilingual factor.

The updated DD algorithm is illustrated in Algorithm 2 (case 2). We introduce two separate sets of dual constraints $\mathbf{w}^e$ and $\mathbf{w}^f$, which range over the set of vertices on their respective half of the graph. Decoding the edge factor model $h_{(i,j)}(y_i^e, y_j^f)$ simply involves finding the pair of tag assignments that gives the highest PMI score, subject to the dual constraints.

The way DD algorithms work in decomposing undirected graphical models is analogous to other message passing algorithms such as loopy belief propagation, but DD gives a stronger optimality guarantee upon convergence (Rush et al., 2010).

3 Joint Alignment and NER Decoding

In this section we develop an extended model in which NER information can in turn be used to improve alignment accuracy. Although we have seen more than a handful of recent papers that apply the dual decomposition method for joint inference problems, all of the past work deals with cases where the various model components have the same inference output space (e.g., dependency parsing (Koo et al., 2010), POS tagging (Rush et al., 2012), etc.). In our case the output space is the much more complex joint alignment and NER tagging space. We propose a novel dual decomposition variant for performing inference over this joint space.

Most commonly used alignment models, such as the IBM models and HMM-based aligner are unsupervised learners, and can only capture simple distortion features and lexical translational features due to the high complexity of the structure prediction space. On the other hand, the CRF-based NER models are trained on manually annotated data, and admit richer sequence and lexical features. The entity label predictions made by the NER model can potentially be leveraged to correct alignment mistakes. For example, in Figure 1, if the tagger knows that the word "Agency" is tagged I-ORG, and if it also knows that the first comma in the Chinese sentence is not part of any entity, then we can infer it is very unlikely that there exists an alignment link between "Agency" and the comma.

To capture this intuition, we extend the Bi-NER model to jointly perform word alignment and NER decoding, and call the resulting model Bi-NER-WA. As a first step, instead of taking the output from an aligner as fixed input, we incorporate two uni-directional aligners into our model. We name the Chinese-to-English aligner model as $m(\mathbf{B}^e)$ and the reverse directional model $n(\mathbf{B}^f)$. $\mathbf{B}^e$ is a matrix that holds the output of the Chinese-to-English aligner. Each $b^e(i,j)$ binary variable in $\mathbf{B}^e$ indicates whether f_j is aligned to e_i; similarly we define output matrix $\mathbf{B}^f$ and $b^f(i,j)$ for Chinese. In our experiments, we used two HMM-based alignment models. But in principle we can adopt any alignment model as long as we can perform efficient inference over it.

We introduce a cross-lingual edge factor $\zeta(i,j)$ in the undirected graphical model for every pair of word indices (i,j), which predicts a binary vari-

Algorithm 2 DD inference algorithm for joint alignment and NER model. A line marked with (2) means it applies to the BI-NER model; a line marked with (3) means it applies to the BI-NER-WA model.

```
$S \leftarrow \mathcal{A}$                                              (2)
$S \leftarrow \{(i, j)\colon \forall i \in |\mathbf{e}|, \forall j \in |\mathbf{f}|\}$                        (3)
$\forall i \in |\mathbf{e}| : w_i^e = 0; \forall j \in |\mathbf{f}| : w_j^f = 0$                     (2,3)
$\forall (i, j) \in \mathcal{S} :\ d^e(i, j) = 0, d^f(i, j) = 0$                     (3)
for $t \leftarrow 1$ to $T$ do
  $\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}*} \leftarrow \text{argmax}\ k\left(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}}\right) + \sum_{i\in|\mathbf{e}|} w_i^e y_i^{e^{(k)}}$          (2,3)
  $\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}*} \leftarrow \text{argmax}\ l\left(\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}}\right) + \sum_{i\in|\mathbf{f}|} w_j^f y_j^{f^{(l)}}$          (2,3)
  $\mathbf{B}^{e*} \leftarrow \text{argmax}\ m\left(\mathbf{B}^e\right) + \sum_{(i,j)} d^e(i,j) b^e(i,j)$          (3)
  $\mathbf{B}^{f*} \leftarrow \text{argmax}\ n\left(\mathbf{B}^f\right) + \sum_{(i,j)} d^f(i,j) b^f(i,j)$          (3)
  for all $(i, j) \in S$ do
    $(y_i^{e^{(h)*}} y_j^{f^{(h)*}}) \leftarrow -w_i^e y_i^{e^{(h)}} - w_j^f y_j^{f^{(h)}}$
    $+ \text{argmax}\ h_{(i,j)}(y_i^{e^{(q)}} y_j^{f^{(q)}})$                    (2)
    $(y_i^{e^{(q)*}} y_j^{f^{(q)*}} a(i,j)^*) \leftarrow -w_i^e y_i^{e^{(q)}} - w_j^f y_j^{f^{(q)}}$
    $+ \text{argmax}\ q_{(i,j)}(y_i^{e^{(q)}} y_j^{f^{(q)}} a(i,j))$
    $- d^e(i,j)a(i,j) - d^f(i,j)a(i,j)$                    (3)
  end for
  Conv = $(\mathbf{y}^{e^{(k)}} = \mathbf{y}^{e^{(q)}} \wedge \mathbf{y}^{f^{(l)}} = \mathbf{y}^{f^{(q)}})$                (2)
  Conv = $(\mathbf{B}^e = \mathbf{A} = \mathbf{B}^f \wedge \mathbf{y}^{e^{(k)}} = \mathbf{y}^{e^{(q)}} \wedge \mathbf{y}^{f^{(l)}} = \mathbf{y}^{f^{(q)}})$  (3)
  if Conv = true , then
    return $\left(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}*}, \mathbf{y}^{\mathbf{f}^{(\mathbf{l})}*}\right)$                    (2)
    return $\left(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}*}, \mathbf{y}^{\mathbf{f}^{(\mathbf{l})}*}, \mathbf{A}\right)$                 (3)
  else
    for all $i \in |\mathbf{e}|$ do
      $w_i^e \leftarrow w_i^e + \alpha_t \left(y_i^{e^{(q|h)*}} - y_i^{e^{(k)*}}\right)$          (2,3)
    end for
    for all $j \in |\mathbf{f}|$ do
      $w_j^f \leftarrow w_j^f + \alpha_t \left(y_j^{f^{(q|h)*}} - y_j^{f^{(l)*}}\right)$          (2,3)
    end for
    for all $(i, j) \in S$ do
      $d^e(i,j) \leftarrow d^e(i,j) + \alpha_t \left(a^{e*}(i,j) - b^{e*}(i,j)\right)$   (3)
      $d^f(i,j) \leftarrow d^f(i,j) + \alpha_t \left(a^{f*}(i,j) - b^{f*}(i,j)\right)$   (3)
    end for
  end if
end for
return $\left(\mathbf{y}_{(\mathbf{T})}^{\mathbf{e}^{(\mathbf{k})}*}, \mathbf{y}_{(\mathbf{T})}^{\mathbf{f}^{(\mathbf{l})}*}\right)$                      (2)
return $\left(\mathbf{y}_{(\mathbf{T})}^{\mathbf{e}^{(\mathbf{k})}*}, \mathbf{y}_{(\mathbf{T})}^{\mathbf{f}^{(\mathbf{l})}*}, \mathbf{A}_{(T)}\right)$                  (3)
```

able $a(i, j)$ for an alignment link between e_i and f_j. The edge factor also predicts the entity tags for e_i and f_j.

The new edge potential q is defined as:

$$q_{(i,j)}\left(y_i^e, y_j^f, a(i,j)\right) =$$

$$\log(P(a(i,j) = 1)) + S(y_i^e, y_j^f | a(i,j))^{P(a(i,j)=1)}$$

$$S(y_i^e, y_j^f | a(i,j)) = \begin{cases} \text{pmi}(y_i^e, y_j^f), \text{if } a(i,j) = 1 \\ 0, \text{else} \end{cases}$$

$P(a(i,j) = 1)$ is the alignment probability assigned by the bilingual edge factor between node e_i and f_j. We initialize this value to $\hat{P}(e_i, f_j) = \frac{1}{2}(P_m(e_i, f_j) + P_n(e_i, f_j))$, where $P_m(e_i, f_j)$ and $P_n(e_i, f_j)$ are the posterior probabilities assigned by the HMM-aligners.

The joint optimization problem is defined as:

$$\max_{\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}} \mathbf{y}^{\mathbf{f}^{(\mathbf{l})}} \mathbf{y}^{\mathbf{e}^{(\mathbf{h})}} \mathbf{y}^{\mathbf{f}^{(\mathbf{h})}} \mathbf{B}^e \mathbf{B}^f \mathbf{A}} k(\mathbf{y}^{\mathbf{e}^{(\mathbf{k})}}) + l(\mathbf{y}^{\mathbf{f}^{(\mathbf{l})}}) +$$

$$m(\mathbf{B}^e) + n(\mathbf{B}^f) + \sum_{(i\in|\mathbf{e}|, j\in|\mathbf{f}|)} q_{(i,j)}(y_i^{e^h}, y_j^{f^{(h)}}, a(i,j))$$

$$\ni \forall (i,j)\colon \left(b^e(i,j) = a(i,j)\right) \wedge \left(b^f(i,j) = a(i,j)\right)$$

$$\wedge \text{ if } a(i,j) = 1 \text{ then } \left(y_i^{e^{(k)}} = y_i^{e^{(h)}}\right) \wedge \left(y_j^{f^{(l)}} = y_j^{f^{(h)}}\right)$$

We include two dual constraints $d^e(i, j)$ and $d^f(i, j)$ over alignments for every bilingual edge factor $\zeta(i, j)$, which are applied to the English and Chinese sides of the alignment space, respectively.

The DD algorithm used for this model is given in Algorithm 2 (case 3). One special note is that after each iteration when we consider updates to the dual constraint for entity tags, we only check tag agreements for cross-lingual edge factors that have an alignment assignment value of 1. In other words, cross-lingual edges that are not aligned do not affect bilingual NER tagging.

Similar to $\phi(i, j)$, $\zeta(i, j)$ factors do not provide that much additional information other than some selectional preferences via PMI score. But the real power of these cross-language edge cliques is that they act as a liaison between the NER and alignment models on each language side, and encourage these models to indirectly agree with each other by having them all agree with the edge cliques.

It is also worth noting that since we decode the alignment models with Viterbi inference, additional constraints such as the neighborhood constraint proposed by DeNero and Macherey (2011) can be easily integrated into our model. The neighborhood constraint enforces that if f_j is aligned to e_i, then f_j can only be aligned to e_{i+1} or e_{i-1} (with a small penalty), but not any other word position. We report results of adding neighborhood constraints to our model in Section 6.

4 Experimental Setup

We evaluate on the large OntoNotes (v4.0) corpus (Hovy et al., 2006) which contains manually annotated NER tags for both Chinese and English. Document pairs are sentence aligned using the Champollion Tool Kit (Ma, 2006). After discarding sentences with no aligned counterpart, a total of 402 documents and 8,249 parallel sentence pairs were used for evaluation. We will refer to this evaluation set as *full-set*. We use odd-numbered documents as the dev set and even-numbered documents as the blind test set. We did not perform parameter tuning on the dev set to optimize performance, instead we fix the initial learning rate to 0.5 and maximum iterations to 1,000 in all DD experiments. We only use the dev set for model development.

The Stanford CRF-based NER tagger was used as the monolingual component in our models (Finkel et al., 2005). It also serves as a state-of-the-art monolingual baseline for both English and Chinese. For English, we use the default tagger setting from Finkel et al. (2005). For Chinese, we use an improved set of features over the default tagger, which includes distributional similarity features trained on large amounts of non-overlapping data.[1]

We train the two CRF models on all portions of the OntoNotes corpus that are annotated with named entity tags, except the parallel-aligned portion which we reserve for development and test purposes. In total, there are about 660 training documents ($\sim$16k sentences) for Chinese and 1,400 documents ($\sim$39k sentences) for English.

Out of the 18 named entity types that are annotated in OntoNotes, which include person, location, date, money, and so on, we select the four most commonly seen named entity types for evaluation. They are *person*, *location*, *organization* and *GPE*. All entities of these four types are converted to the standard BIO format, and background tokens and all other entity types are marked with tag `O`. When we consider label agreements over aligned word pairs in all bilingual agreement models, we ignore the distinction between `B-` and `I-` tags.

We report standard NER measures (entity precision (P), recall (R) and F_1 score) on the test set. Statistical significance tests are done using the paired bootstrap resampling method (Efron and Tibshirani, 1993).

For alignment experiments, we train two unidirectional HMM models as our baseline and monolingual alignment models. The parameters of the HMM were initialized by IBM Model 1 using the agreement-based EM training algorithms from Liang et al. (2006). Each model is trained for 2 iterations over a parallel corpus of 12 million English words and Chinese words, almost twice as much data as used in previous work that yields state-of-the-art unsupervised alignment results (DeNero and Klein, 2008; Haghighi et al., 2009; DeNero and Macherey, 2011).

Word alignment evaluation is done over the sections of OntoNotes that have matching gold-standard word alignment annotations from GALE Y1Q4 dataset.[2] This subset contains 288 documents and 3,391 sentence pairs. We will refer to this subset as *wa-subset*. This evaluation set is over 20 times larger than the 150 sentences set used in most past evaluations (DeNero and Klein, 2008; Haghighi et al., 2009; DeNero and Macherey, 2011).

Alignments input to the BI-NER model are produced by thresholding the averaged posterior probability at 0.5. In joint NER and alignment experiments, instead of posterior thresholding, we take the direct intersection of the Viterbi-best alignment of the two directional models. We report the standard P, R, F_1 and Alignment Error Rate (AER) measures for alignment experiments.

An important past work to make comparisons with is Burkett et al. (2010b). Their method is similar to ours in that they also model bilingual agreement in conjunction with two CRF-based monolingual models. But instead of using just the PMI scores of bilingual NE pairs, as in our work, they employed a feature-rich log-linear model to capture bilingual correlations. Parameters in their log-linear model require training with bilingually annotated data, which is not readily available. To counter this problem, they proposed an "up-training" method which simulates a supervised learning environment by pairing a weak classifier with strong classifiers, and train the bilingual model to rank the output of the strong classifier highly among the N-best outputs of the weak classifier. In order to compare directly with their method, we obtained the code behind Burkett et al. (2010b) and reproduced their experimental setting for the OntoNotes data. An extra set of 5,000 unannotated parallel sentence pairs are used for

[1] The exact feature set and the CRF implementation can be found here: `http://nlp.stanford.edu/software/CRF-NER.shtml`

[2] LDC Catalog No. LDC2006E86.

	Chinese			English		
	P	R	F_1	P	R	F_1
Mono	76.89	61.64	68.42	81.98	74.59	78.11
Burkett	77.52	65.84	71.20	82.28	76.64	79.36
Bi-soft	**79.14**	**71.55**	**75.15**	**82.58**	**77.96**	**80.20**

Table 1: NER results on bilingual parallel test set. Best numbers on each measure that are statistically significantly better than the monolingual baseline and Burkett et al. (2010b) are highlighted in bold.

training the reranker, and the reranker model selection was performed on the development dataset.

5 Bilingual NER Results

The main results on bilingual NER over the test portion of *full-set* are shown in Table 1. We initially experimented with the hard agreement model, but it performs quite poorly for reasons we discussed in Section 2.2. The BI-NER model with soft agreement constraints, however, significantly outperforms all baselines. In particular, it achieves an absolute F_1 improvement of 6.7% in Chinese and 2.1% in English over the CRF monolingual baselines.

A well-known issue with the DD method is that when the model does not necessarily converge, then the procedure could be very sensitive to hyper-parameters such as initial step size and early termination criteria. If a model only gives good performance with well-tuned hyper-parameters, then we must have manually annotated data for tuning, which would significantly reduce the applicability and portability of this method to other language pairs and tasks. To evaluate the parameter sensitivity of our model, we run the model from 50 to 3000 iterations before early stopping, and with 6 different initial step sizes from 0.01 to 1. The results are shown in Figure 2. The soft agreement model does not seem to be sensitive to initial step size and almost always converges to a superior solution than the baseline.

6 Joint NER and Alignment Results

We present results for the BI-NER-WA model in Table 2. By jointly decoding NER with word alignment, our model not only maintains significant improvements in NER performance, but also yields significant improvements to alignment performance. Overall, joint decoding with NER alone yields a 10.8% error reduction in AER over the baseline HMM-aligners, and also gives improvement over BI-NER in NER. Adding additional neighborhood constraints gives a further 6% error reduction in AER, at the cost of a small loss in Chinese NER. In terms of word alignment results, we see great increases in F_1 and recall, but precision goes down significantly. This is because the joint decoding algorithm promotes an effect of "soft-union", by encouraging the two unidirectional aligners to agree more often. Adding the neighborhood constraints further enhances this union effect.

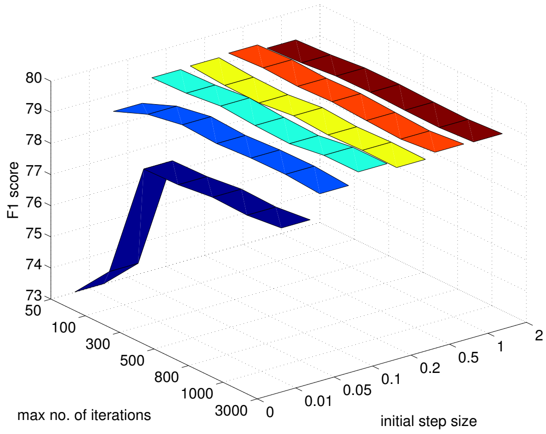

Figure 2: Performance variance of the soft agreement models on the Chinese dev dataset, as a function of step size (*x-axis*) and maximum number of iterations before early stopping (*y-axis*).

7 Error Analysis and Discussion

We can examine the example in Figure 3 to gain an understanding of the model's performance. In this example, a snippet of a longer sentence pair is shown with NER and word alignment results. The monolingual Chinese tagger provides a strong cue that word f_6 is a person name because the unique 4-character word pattern is commonly associated with foreign names in Chinese, and also the word is immediately preceded by the word "president". The English monolingual tagger, however, confuses the aligned word e_0 with a GPE.

Our bilingual NER model is able to correct this error as expected. Similarly, the bilingual model corrects the error over e_{11}. However, the model also propagates labeling errors from the English side over the entity "Tibet Autonomous Region" to the Chinese side. Nevertheless, the resulting Chinese tags are arguably more useful than the original tags assigned by the baseline model.

In terms of word alignment, the HMM models failed badly on this example because of the long

	NER-Chinese			NER-English			word alignment			
	P	R	F_1	P	R	F_1	P	R	F_1	AER
HMM-WA	-	-	-	-	-	-	**90.43**	40.95	56.38	43.62
Mono-CRF	82.50	66.58	73.69	84.24	78.70	81.38	-	-	-	-
Bi-NER	**84.87**	75.30	79.80	**84.47**	81.45	82.93	-	-	-	-
Bi-NER-WA	84.42	**76.34**	**80.18**	84.25	**82.20**	**83.21**	77.45	50.43	61.09	38.91
Bi-NER-WA+NC	84.25	75.09	79.41	84.28	82.17	**83.21**	76.67	**54.44**	**63.67**	**36.33**

Table 2: Joint alignment and NER test results. +NC means incorporating additional neighbor constraints from DeNero and Macherey (2011) to the model. Best number in each column is highlighted in bold.

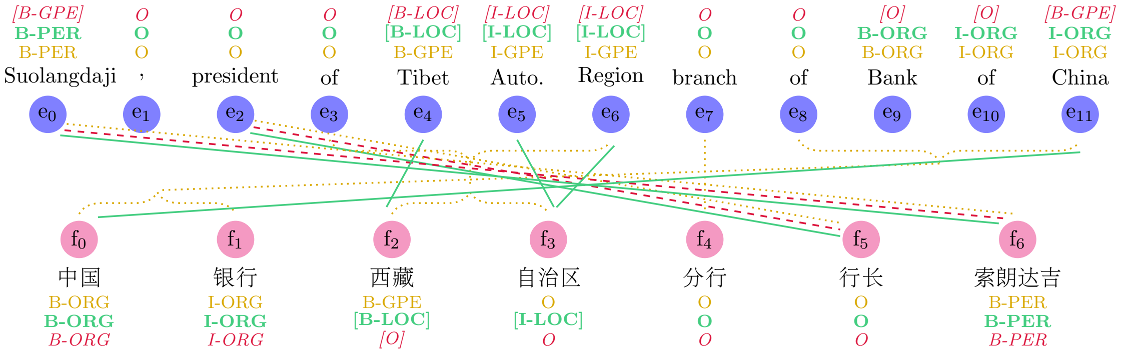
[B-GPE] O O O [B-LOC] [I-LOC] [I-LOC] O O [O] [O] [B-GPE]
B-PER O O O [B-LOC] [I-LOC] [I-LOC] O O B-ORG I-ORG I-ORG
B-PER O O O B-GPE I-GPE I-GPE O O B-ORG I-ORG I-ORG
Suolangdaji , president of Tibet Auto. Region branch of Bank of China
e0 e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 e11
f0 f1 f2 f3 f4 f5 f6
中国 银行 西藏 自治区 分行 行长 索朗达吉
B-ORG I-ORG B-GPE O O O B-PER
B-ORG I-ORG [B-LOC] [I-LOC] O O B-PER
B-ORG I-ORG [O] O O O B-PER

Figure 3: An example output of our BI-NER-WA model. Dotted alignment links are the oracle, dashed links are alignments from HMM baseline, and solid links are outputs of our model. Entity tags in the gold line (closest to nodes e_i and f_j) are the gold-standard tags; in the **green** line (second closest to nodes) are output from our model; and in the *crimson* line (furthest from nodes) are baseline output.

distance swapping phenomena. The two unidirectional HMMs also have strong disagreements over the alignments, and the resulting baseline aligner output only recovers two links. If we were to take this alignment as fixed input, most likely we would not be able to recover the error over e_{11}, but the joint decoding method successfully recovered 4 more links, and indirectly resulted in the NER tagging improvement discussed above.

8 Related Work

The idea of employing bilingual resources to improve over monolingual systems has been explored by much previous work. For example, Huang et al. (2009) improved parsing performance using a bilingual parallel corpus. In the NER domain, Li et al. (2012) presented a cyclic CRF model very similar to our BI-NER model, and performed approximate inference using loopy belief propagation. The feature-rich CRF formulation of bilingual edge potentials in their model is much more powerful than our simple PMI-based bilingual edge model. Adding a richer bilingual edge model might well further improve our results, and this is a possible direction for further experimentation. However, a big drawback of this approach is that training such a feature-rich model requires manually annotated bilingual NER data, which can be prohibitively expensive to generate. How and where to obtain training signals without manual supervision is an interesting and open question. One of the most interesting papers in this regard is Burkett et al. (2010b), which explored an "up-training" mechanism by using the outputs from a strong monolingual model as ground-truth, and simulated a learning environment where a bilingual model is trained to help a "weakened" monolingual model to recover the results of the strong model. It is worth mentioning that since our method does not require additional training and can take pretty much any existing model as "black-box" during decoding, the richer and more accurate bilingual model learned from Burkett et al. (2010b) can be directly plugged into our model.

A similar dual decomposition algorithm to ours was proposed by Riedel and McCallum (2011) for biomedical event detection. In their Model 3, the trigger and argument extraction models are reminiscent of the two monolingual CRFs in our model; additional binding agreements are enforced over every protein pair, similar to how we enforce agreement between every aligned word pair. Martins et al. (2011b) presented a new DD method that combines the power of DD with the augmented Lagrangian method. They showed that their method can achieve faster convergence than traditional sub-gradient methods in models with many overlapping components (Martins et al., 2011a). This method is directly applicable to our work.

Another promising direction for improving NER performance is in enforcing global label consistency across documents, which is an idea that has been greatly explored in the past (Sutton and McCallum, 2004; Bunescu and Mooney, 2004; Finkel et al., 2005). More recently, Rush et al. (2012) and Chieu and Teow (2012) have shown that combining local prediction models with global consistency models, and enforcing agreement via DD is very effective. It is straightforward to incorporate an additional global consistency model into our model for further improvements.

Our joint alignment and NER decoding approach is inspired by prior work on improving alignment quality through encouraging agreement between bi-directional models (Liang et al., 2006; DeNero and Macherey, 2011). Instead of enforcing agreement in the alignment space based on best sequences found by Viterbi, we could opt to encourage agreement between posterior probability distributions, which is related to the posterior regularization work by Graça et al. (2008). Cromières and Kurohashi (2009) proposed an approach that takes phrasal bracketing constraints from parsing outputs, and uses them to enforce phrasal alignments. This idea is similar to our joint alignment and NER approach, but in our case the phrasal constraints are indirectly imposed by entity spans. We also differ in the implementation details, where in their case belief propagation is used in both training and Viterbi inference.

Burkett et al. (2010a) presented a supervised learning method for performing joint parsing and word alignment using log-linear models over parse trees and an ITG model over alignment. The model demonstrates performance improvements in both parsing and alignment, but shares the common limitations of other supervised work in that it requires manually annotated bilingual joint parsing and word alignment data.

Chen et al. (2010) also tackled the problem of joint alignment and NER. Their method employs a set of heuristic rules to expand a candidate named entity set generated by monolingual taggers, and then rank those candidates using a bilingual named entity dictionary. Our approach differs in that we provide a probabilistic formulation of the problem and do not require pre-existing NE dictionaries.

9 Conclusion

We introduced a graphical model that combines two HMM word aligners and two CRF NER taggers into a joint model, and presented a dual decomposition inference method for performing efficient decoding over this model. Results from NER and word alignment experiments suggest that our method gives significant improvements in both NER and word alignment. Our techniques make minimal assumptions about the underlying monolingual components, and can be adapted for many other tasks such as parsing.

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, and the support of the U.S. Defense Advanced Research Projects Agency (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 DARPA, or the US government.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of ACL*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*.

Dimitri P. Bertsekas. 1999. *Nonlinear Programming*. Athena Scientific, New York.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of ACL*.

Razvan Bunescu and Raymond J. Mooney. 2004. Collective information extraction with relational Markov networks. In *Proceedings of ACL*.

David Burkett, John Blitzer, and Dan Klein. 2010a. Joint parsing and alignment with weakly synchronized grammars. In *Proceedings of NAACL-HLT*.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010b. Learning better monolingual models with unannotated bilingual text. In *Proceedings of CoNLL*.

Yufeng Chen, Chengqing Zong, and Keh-Yih Su. 2010. On jointly recognizing and aligning bilingual named entities. In *Proceedings of ACL*.

Hai Leong Chieu and Loo-Nin Teow. 2012. Combining local and non-local information with dual decomposition for named entity recognition from text. In *Proceedings of 15th International Conference on Information Fusion (FUSION)*.

Fabien Cromières and Sadao Kurohashi. 2009. An alignment algorithm using belief propagation and a structure-based distortion model. In *Proceedings of EACL/ IJCNLP*.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of ACL*.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of ACL*.

Brad Efron and Robert 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 ACL*.

Joao Graça, Kuzman Ganchev, and Ben Taskar. 2008. Expectation maximization and posterior constraints. In *Proceedings of NIPS*.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of ACL*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of NAACL-HLT*.

Fei Huang and Stephan Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 2002 International Conference on Multimodal Interfaces (ICMI)*.

Liang Huang, Wenbin Jiang, and Qun Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of EMNLP*.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from Wikipedia. In *Proceedings of ACL*.

Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, and David Sontag. 2010. Dual decomposition for parsing with non-projective head automata. In *Proceedings of EMNLP*.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of CIKM*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC*.

André F. T. Martins, Noah A. Smith, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2011a. Dual decomposition with many overlapping components. In *Proceedings of EMNLP*.

Andre F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2011b. Augmenting dual decomposition for map inference. In *Proceedings of the International Workshop on Optimization for Machine Learning (OPT 2010)*.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of EMNLP*.

Alexander M. Rush and Michael Collins. 2012. A tutorial on dual decomposition and Lagrangian relaxation for inference in natural language processing. *JAIR*, 45:305–362.

Alexander M. Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of EMNLP*.

Alexander M. Rush, Roi Reichert, Michael Collins, and Amir Globerson. 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. In *Proceedings of EMNLP*.

Charles Sutton and Andrew McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. In *Proceedings of ICML Workshop on Statistical Relational Learning and Its connections to Other Fields*.