Effective Bilingual Constraints for Semi-supervised Learning of Named Entity Recognizers

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

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

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

Abstract

Most semi-supervised methods in Natural Language Processing capitalize on unannotated resources in a single language; however, information can be gained from using parallel resources in more than one language, since translations of the same utterance in different languages can help to disambiguate each other. We demonstrate a method that makes effective use of vast amounts of *bilingual* text (a.k.a. bi-text) to improve monolingual systems. We propose a factored probabilistic sequence model that encourages both cross-language and intra-document consistency. A simple Gibbs sampling algorithm is introduced for performing approximate inference. Experiments on English-Chinese Named Entity Recognition (NER) using the OntoNotes dataset demonstrate that our method is significantly more accurate than state-of-the-art monolingual CRF models in a bilingual test setting. Our model also improves on previous work by Burkett et al. (2010), achieving a relative error reduction of 10.8% and 4.5% in Chinese and English, respectively. Furthermore, by annotating a moderate amount of unlabeled bi-text with our bilingual model, and using the tagged data for uptraining, we achieve a 9.2% error reduction in Chinese over the state-of-the-art Stanford monolingual NER system.

Introduction

Supervised learning algorithms have been met with great success in many areas of Natural Language Processing (NLP). It is well-known that the performance of supervised learners increases when more labeled training examples become available. In most application scenarios, however, manually labeled data are extremely limited in quantity and costly to produce. On the other hand, we live in an age of abundance of unannotated data — as regards NLP, there has been an explosion in the amount of freely available web and news texts. One would expect to greatly increase the coverage of a system if such large amounts of additional data can be incorporated in a judicious manner.

A number of semi-supervised techniques have been introduced to tackle this problem, such as bootstrapping (Yarowsky 1995; Collins and Singer 1999; Riloff and Jones 1999), multi-view learning (Blum and Mitchell 1998; Ganchev et al. 2008) and structural learning (Ando and Zhang 2005). Most previous semi-supervised work is situated in a monolingual setting where all unannotated data are available only in a single language.

However, in recent years, a vast amount of translated parallel texts have been generated in our increasingly connected multilingual world. While such bi-texts have primarily been leveraged to train statistical machine translation (SMT) systems, contemporary research has increasingly considered the possibilities of utilizing parallel corpora to improve systems outside of SMT. For example, Yarowsky and Ngai (2001) projects the part-of-speech labels assigned by a supervised model in one language (e.g. English) onto word-aligned parallel text in another language (e.g. Chinese) where less manually annotated data is available. Similar ideas were also employed by Das and Petrov (2011) and Fu, Qin, and Liu (2011).

A severe limitation of methods employing bilingual projection is that they can only be applied to test scenarios where parallel sentence pairs are available. It is more desirable to improve monolingual system performance, which is more broadly applicable. Previous work such as Li et al. (2012) and Kim, Toutanova, and Yu (2012) successfully demonstrated that manually-labeled bilingual corpora can be used to improve monolingual system performance. This approach, however, encounters the difficulty that manually annotated bilingual corpora are even harder to come by than monolingual ones.

In this work, we consider a semi-supervised learning scheme using unannotated bi-text. For a given language pair (e.g., English-Chinese), we expect one language (e.g. English) to have more annotated training resources than the other (e.g. Chinese), and thus there exists a strong monolingual model (for English) and a weaker model (for Chinese). Since bi-text contains translations across the two languages, an aligned sentence pair would exhibit some semantic and syntactic similarities. Thus we can constrain the two models to agree with each other by making joint predictions that are skewed towards the more informed model. In general, errors made in the lower-resource model will be corrected by the higher-resource model, but we also anticipate that these joint predictions will have higher quality for both languages than the output of a monolingual model alone. We can then apply this bilingual annotation method to a large amount of

[*] Corresponding author

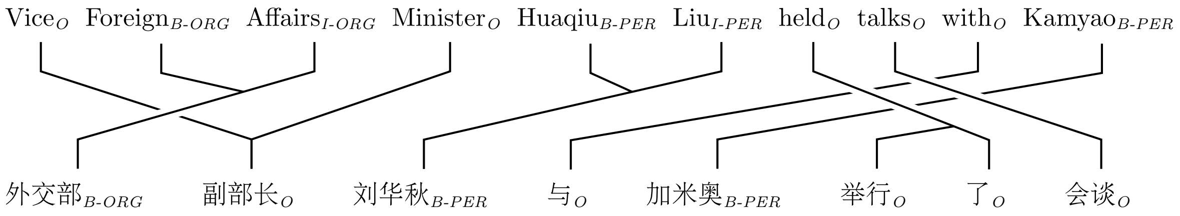

Figure 1: Example of NER labels between two word-aligned bilingual parallel sentences.

unannotated bi-text, and use the resulting annotated data as additional training data to train a new monolingual model with better coverage.[1]

Burkett et al. (2010) proposed a similar framework with a "multi-view" learning scheme where k-best outputs of two monolingual taggers are reranked using a complex self-trained reranking model. In our work, we propose a simple decoding method based on Gibbs sampling that eliminates the need for training complex reranking models. In particular, we construct a new factored probabilistic model by chaining together two Conditional Random Field monolingual models with a bilingual constraint model, which encourages soft label agreements. We then apply Gibbs sampling to find the best labels under the new factored model. We can further improve the quality of bilingual prediction by incorporating an additional model, expanding upon Finkel, Grenager, and Manning (2005), that enforces global label consistency for each language.

Experiments on Named Entity Recognition (NER) show that our bilingual method yields significant improvements over the state-of-the-art Stanford NER system. When evaluated over the standard OntoNotes English-Chinese dataset in a bilingual setting, our models achieve a F_1 error reduction of 18.6% in Chinese and 9.9% in English. Our method also improves over Burkett et al. (2010) with a relative error reduction of 10.8% and 4.5% in Chinese and English, respectively. Furthermore, we automatically label a moderate-sized set of 80k sentence pairs using our bilingual model, and train new monolingual models using an uptraining scheme. The resulting monolingual models demonstrate an error reduction of 9.2% over the Stanford NER systems for Chinese.[2]

Monolingual NER with CRF

Named Entity Recognition is an important task in NLP. It serves as a first step in turning unstructured text into structured data, and has broad applications in news aggregation, question answering, and bioNLP. Given an input sentence, an NER tagger identifies words that are part of a named entity, and assigns the entity type and relative position information. For example, in the commonly used `BIO` tagging scheme, a tag such as `B-PERSON` indicates the word is the beginning of a *person name* entity; and a `I-LOCATION` tag marks the word to be inside a *location* entity. All words marked with tag `O` are not part of any entity. Figure 1 illustrates a tagged sentence pair in English and Chinese.

Current state-of-the-art supervised NER systems employ an undirected graphical model called Conditional Random Field (CRF) (Lafferty, McCallum, and Pereira 2001). Given an input sentence $\mathbf{x}$, a linear-chain structured CRF defines the following conditional probability for tag sequence $\mathbf{y}$:

$$P_{mono}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\prod_{i}\exp(\sum_{j}\lambda_j f_j(y_i, y_{i-1}|\mathbf{x})) \quad (1)$$

where f_j is the jth feature function, λ_j is the feature weight, and $Z(\mathbf{x})$ is the partition function.

Bilingual NER Constraints

A pair of aligned sentences in two languages contain complementary cues to aid the analysis of each other. For example, in Figure 1, it is not immediately obvious whether the phrase "Foreign Affairs" on the English side refers to an organization (Ministry of Foreign Affairs), or general foreign affairs. But the aligned word on the Chinese side is a lot less ambiguous, and can be easily identified as an *organization* entity.

Another example is that in the Chinese training data we have never seen the translation of the name "Kamyao". As a result, the tagger cannot make use of lexical features, and so has to rely on less informative contextual features to predict if it is a *geo-political entity (GPE)* or a *person*. But we have seen the aligned word on the English side being tagged as *person*, and thus can infer that the Chinese aligned entity should also be a *person*.

It is straight-forward to see that accurate word alignment is essential in such an analysis. Fortunately, there are automatic word alignment systems used in MT research that produce robust and accurate alignment results, and our method will use the output of one (Liang, Taskar, and Klein 2006).

Hard Agreement Constraints

Drawing on the above observations, we first propose a simple bilingual constraint model that enforces *hard* agreements.

We define the following probability for an output sequence pair $\mathbf{y}_c$ and $\mathbf{y}_e$ for Chinese and English input sentences $\mathbf{x}_c$ and $\mathbf{x}_e$, respectively:

$$P_{bi}(\mathbf{y}_c, \mathbf{y}_e) = \prod_{A=\{a^c,a^e\}} \mathbb{I}(y_{a^c}, y_{a^e}) \quad (2)$$

[1] This training regimen is also referred to as "uptraining" (Petrov et al. 2010).

[2] All of our code is made available at `nlp.stanford.edu/software/CRF-NER.shtml`.

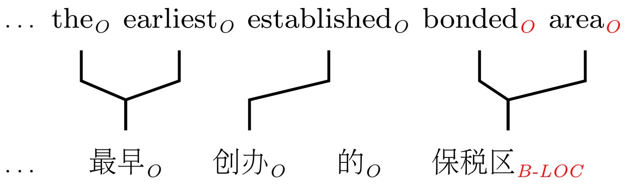
… the$_O$ earliest$_O$ established$_O$ bonded$_O$ area$_O$
… 最早$_O$ 创办$_O$ 的$_O$ 保税区$_{B\text{-}LOC}$

Figure 2: Example of annotation standard inconsistency

where A is the set of all aligned word pairs, and $\mathbb{I}(y_{a^c}, y_{a^e})$ is an indicator function that equals 1 if $y_{a^c} = y_{a^e}$, and 0 otherwise.

Soft Agreement Constraints

If we apply hard agreement constraints, any output sequence pairs that disagree on any tag pair will be assigned zero probability. Such a hard constraint is not always satisfied in practice, since annotation standards in different languages can differ. An example is given in Figure 2, where the phrase mention of "bonded area" is considered a *location* in the Chinese gold-standard, but not in the English gold-standard.

We can soften these constraints by replacing the 1 and 0 values in indicator function $\mathbb{I}(y_{a^c}, y_{a^e})$ with a probability measure. We first tag a set of unannotated bilingual sentence pairs using two baseline monolingual CRF taggers. Then we collect counts of aligned entity tag pairs from the auto-generated tagged data. The value $\mathbb{I}(y_{a^c}, y_{a^e})$ is chosen to be the pairwise mutual information score of the entity pair (y_{a^c}, y_{a^e}). This version of constraints is denoted as `auto`.

Alignment Uncertainty

When we consider the previous two sets of bilingual constraints, we assume the word alignments are given by some off-the-shelf alignment model which outputs a set of "hard" alignments. In practice, most statistical word alignment models assign a probability to each alignment pair, and "hard" alignments are produced by cutting off alignment pairs that fall below a threshold value.

To take into account alignment uncertainties, we modify function $\mathbb{I}(y_{a^c}, y_{a^e})$ by exponentiating its value to the power of the alignment probability to give a new function: $\mathbb{U}(y_{a^c}, y_{a^e}) = \mathbb{I}(y_{a^c}, y_{a^e})^{P(y_{a^c}, y_{a^e})}$. The intuition behind this modification is that pairs with a higher alignment probability will reflect more probability fluctuation when different label assignments are considered.

For example, consider an extreme case where a particular pair of aligned words has alignment probability 0. Then the value of the $\mathbb{U}$ function will always be 1 regardless of what tags are assigned to the two words, thus reducing the impact of different choices of tags for this pair in the overall tag sequence assignment.

Gibbs Sampling with Factored Models

In a monolingual setting, exact inference in a standard linear-chain CRF can be done by applying the Viterbi algorithm to find the most likely output sequence. But when we consider the joint probability of an output sequence pair in a bilingual setting, especially when we apply the aforementioned bilingual constraints, cyclic cliques are introduced into the Markov random field which make exact inference algorithms intractable.

Markov Chain Monte Carlo (MCMC) methods offer a simple and elegant solution for approximate inference by constructing a Markov chain whose stationary distribution is the target distribution.

In this work, we adopt a specific MCMC sampling method called Gibbs sampling (Geman and Geman 1984). We define a Markov chain over output sequences by observing a simple transition rule: from a current sequence assignment at time $t-1$, we can transition into the next sequence at time t by changing the label at any position i. And the distribution over these transitions is defined as:

$$P(\mathbf{y}^t|\mathbf{y}^{t-1}) = P(y_i^t|\mathbf{y}_{-i}^{t-1}, \mathbf{x}) \tag{3}$$

where $\mathbf{y}_{-i}^{t-1}$ is the set of all labels except y_i at time $t-1$.

To apply the bilingual constraints during decoding, we formulate a new *factored* model by combining the two monolingual CRF models (one for each language) with the bilingual constraint model via a simple product.[3] The resulting model is of the following form:

$$P(\mathbf{y}_c, \mathbf{y}_e|\mathbf{x}_c, \mathbf{x}_e) = P_{mono}(\mathbf{y}_c|\mathbf{x}_c)P_{mono}(\mathbf{y}_e|\mathbf{x}_e)P_{bi}(\mathbf{y}_c, \mathbf{y}_e) \tag{4}$$

Obtaining the state transition model $P(y_i^t|\mathbf{y}_{-i}^{t-1}, \mathbf{x})$ for the monolingual CRF models is straight-forward. In the case of a first order linear-chain CRF, the Markov blanket is the neighboring two cliques. Given the Markov blanket of state i, the label at position i is independent of all other states. Thus we can compute the transition model simply by normalizing the product of the neighboring clique potentials. Finkel, Grenager, and Manning (2005) gave a more detailed account of how to compute this quantity.

The transition probability of label y_{c_i} in the bilingual constraint model is defined as $\prod_{(y_{c_i}, y_{e_k}) \in A} \mathbb{U}(y_{c_i}, y_{e_k})$, where y_{e_k} is a word aligned to y_{c_i}.

At decoding time, we walk the Markov chain by taking samples at each step. We start from some random assignment of the label sequence, and at each step we randomly sample a new value for y_i at a randomly chosen position i. After a fixed number of steps, we output a complete sequence as our final solution. In practice, MCMC sampling could be quite slow and inefficient, especially when the input sentence is long. To speed up the sampling process, we initialize the state sequence from the best sequences found by Viterbi decoding using only the monolingual models.

A bigger problem with vanilla Gibbs sampling is that the random samples we draw do not necessarily give us the most likely state sequence, as given by Viterbi in the exact inference case. One way to tackle this problem is to borrow the *simulated annealing* technique from optimization research

[3] This model double-counts the state sequence conditioned on a given observation, and therefore is likely deficient. However, we do not find this to be a problem in practice.

(Kirkpatrick, Gelatt, and Vecchi 1983). We redefine the transition probability in Eqn. 3 as:

$$P(\mathbf{y}^t|\mathbf{y}^{t-1}) = \frac{P(y_i^t|\mathbf{y}_{-i}^{t-1},\mathbf{x})^{1/c_t}}{\sum_j P(y_j^t|\mathbf{y}_{-j}^{t-1},\mathbf{x})^{1/c_t}} \qquad (5)$$

where $\mathbf{c} = \{c_0 \dots c_T\}$ is the schedule of annealing "temperature," with $0 \leq c_i \leq 1$. The distribution becomes sharper as the value of c_i move towards 0. In our experiments we adopted a linear cooling schedule, where $c_0 = 1$, and $c_{t+1} = c_t - 1/T$. This technique has been shown to be effective by Finkel, Grenager, and Manning (2005).

Global Consistency Constraints

A distinctive feature of the proposed factored model and Gibbs sampling inference is the ability to incorporate non-local constraints that are not easily captured in a traditional Markov network model. The bilingual constraint model described earlier is certainly a benefactor of this unique characteristic.

Still, there are further linguistic constraints that we can apply to improve the NER system. For example, many previous papers have made the observation that occurrences of the same word sequence within a given document are unlikely to take on different entity types (Bunescu and Mooney 2004; Sutton and McCallum 2004; Finkel, Grenager, and Manning 2005; *inter alia*) . Similar to Finkel, Grenager, and Manning (2005), we devise a global consistency model as follows:

$$P_{glo}(\mathbf{y}|\mathbf{x}) = \prod_{\gamma\in\Gamma} \phi_\gamma^{\#(\gamma,\mathbf{y},\mathbf{x})} \qquad (6)$$

Γ is the set of all possible entity type violations, ϕ_γ is the penalty parameter for violation type γ, and $\#(\gamma,\mathbf{y},\mathbf{x})$ is the count of violations γ in sequence $\mathbf{y}$. For example, if the word sequence "China Daily" has occurred both as *GPE* and *organization* exactly once, then the penalty ϕ_γ for *GPE-to-organization* violation will apply once. The parameter values of ϕ_γ are estimated empirically by counting the occurrences of entity pairs of the same word sequence in the training data.

We can now factor in one global consistency model for each language by taking the product of Eqn. 4 with Eqn. 6. The same Gibbs sampling procedure applies unchanged to this new factored model. At test time, instead of tagging one sentence at a time, we group together sentences that belong to the same document, and tag one document at a time.

Enhancing Recall

A flaw of the Finkel, Grenager, and Manning (2005) model described above is that consistency is enforced by applying penalties to entity type violations. But if a word is not tagged with an entity type, it will not receive any penalty since no entity type violations would occur. Therefore, this model has the tendency of favoring null annotations, which can result in losses in model recall.

Chinese NER Templates
00: 1 (class bias param)
01: $w_{i+k}, -1 \leq k \leq 1$
02: $w_{i+k-1} \circ w_{i+k}, 0 \leq k \leq 1$
03: shape$(w_{i+k}), -4 \leq k \leq 4$
04: prefix$(w_i, k), 1 \leq k \leq 4$
05: prefix$(w_{i-1}, k), 1 \leq k \leq 4$
06: suffix$(w_i, k), 1 \leq k \leq 4$
07: suffix$(w_{i-1}, k), 1 \leq k \leq 4$
08: radical$(w_i, k), 1 \leq k \leq$ len(w_i)
09: distsim$(w_{i+k}), -1 \leq k \leq 1$
Unigram Features
$y_i \circ$ 00 – 09
Bigram Features
$y_{i-1} \circ y_i \circ$ 00 – 09

Table 1: Basic features of Chinese NER. $\circ$ means string concatenation and y_i is the named entity label of the i^{th} word w_i. shape(w_i) is the shape of w_i, such as date and number. prefix/suffix(w_i, k) denotes the k-characters prefix/suffix of w_i. radical(w_i, k) denotes the radical of the k^{th} Chinese character of w_i.[4] len(w_i) is the number of Chinese characters in w_i. distsim(w_i, k) denotes the distributional similarity features based on large word clusters.

We fix this deficiency in Finkel, Grenager, and Manning (2005) by introducing a new "reward" parameter δ, which has value > 0. δ is activated each time we see a matching pair of entities for the same word occurrence. The new P_{glo} is modified as:

$$P_{glo}(\mathbf{y}|\mathbf{x}) = \delta^{\#(\delta,\mathbf{y},\mathbf{x})} \prod_{\gamma\in\Gamma} \phi_\gamma^{\#(\gamma,\mathbf{y},\mathbf{x})} \qquad (7)$$

where $\#(\delta,\mathbf{y},\mathbf{x})$ is the activation count of δ in sequence $\mathbf{y}$.

This model is in fact a naive Bayes model, where the parameters δ and ϕ are empirically estimated (a value of 2 is used for δ in our experiments, based on tuning on a development set). A similar global consistency model was shown to be effective in Rush et al. (2012), where parameters were also tuned on a development set.

Experimental Setup

To compare the proposed bilingual constraint decoding algorithm against traditional monolingual methods, we evaluate on a large, manually annotated parallel corpus that contains named entity annotation in both Chinese and English. The corpus we use is the latest version (v4.0) of the OntoNotes corpus (Hovy et al. 2006), which includes 401 pairs of Chinese and English documents (chtb_0001-0325, ectb_1001-1078). We use odd-numbered documents as the development set and even-numbered documents as the blind test set.

These document pairs are aligned at document level, but not at sentence or word level. To obtain sentence alignment, we use the Champollion Tool Kit (CTK).[5] After discarding sentences with no aligned counterpart, a total of 8,249 sentence pairs were retained. We induce word alignment using the BerkeleyAligner toolkit (Liang, Taskar, and Klein

[4] www.unicode.org/charts/unihan.html

[5] champollion.sourceforge.net

2006).[6] The aligner outputs the posterior probability for each aligned word pair. To increase efficiency, we prune away all alignments that have probability less than 0.1.

We adopt the state-of-the-art monolingual Stanford NER tagger as a strong baseline for both English and Chinese. For English, we use the default tagger setting from Finkel, Grenager, and Manning (2005). For Chinese, we use an improved set of features over the default tagger, which are listed in Table 1. Both models make use of distributional similarity features taken from word clusters trained on large amounts of non-overlapping data. 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 documents (∼16k sentences) and 1,400 documents (∼39k sentences) for Chinese and English, respectively.

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 entities types are marked with tag O.

In all of the Gibbs sampling experiments, a fixed number of 2000 sampling steps are taken, and a linear cooling schedule is used in the deterministic annealing procedure.

In order to compare our method with past work, we obtained code from Burkett et al. (2010) and reproduced their experiment setting for the OntoNotes data. An extra set of 5,000 unannotated parallel sentence pairs are used for training the reranker, and the reranker model selection was performed on the development dataset.

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), where we repeatedly draw random samples with replacement from the output of the two systems, and compare the test statistics (e.g. absolute difference in F_1 score) of the new samples with the observed test statistics. We used 1000 sampling iterations in our experiments.

Bilingual NER Results

The main results on Chinese and English test sets are shown in Table 2. The first row (CRF) shows the baseline monolingual model performance. As we can see, the performance on Chinese is much lower than on English. This is partially attributed to the fact that the Chinese NER tagger was trained on less than half as much data, but it is also because NER in Chinese is a harder problem (e.g., there are no capitalization features in Chinese, which is a very strong indicator of named entities in English).

By enforcing hard agreement constraints, we can see from row hard that there is an increase of about 1.4% in absolute F_1 score on the Chinese side, but at the expense of a 0.9% drop on the English side. The tradeoff mainly occurs in recall.

	Chinese			English		
	P	R	F_1	P	R	F_1
CRF	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**
hard	76.19	64.47	**69.84**	82.13	72.85	77.21
manual	80.02	65.85	**72.24**	82.87	74.56	**78.50**
auto	78.53	66.90	**72.25**	82.11	75.40	**78.62**
auto+aP	79.17	68.46	**73.43**	82.05	75.56	**78.67**

Table 2: Results on bilingual parallel test set. F_1 scores that are statistically significantly better than the CRF baseline is highlighted in bold.

When we loosen the bilingual constraint to allow soft-agreement by simply assigning a hand-picked value (0.02) to aligned entities of different types (row manual), we observe a significant increase in accuracy in both Chinese and English. This suggests that the soft alignment successfully accounted for the cases where annotation standards differ in the two languages. In particular, the Chinese results are 3.8% better than the monolingual baseline, a 12% relative error reduction.

When we replace the arbitrary hand-picked soft-agreement probabilities with empirical counts from the auto-tagged dataset (row auto), we see a small increase in recall on both sides, but a drop in precision for Chinese. However, accounting for alignment uncertainty (row auto+aP) increases both precision and recall for Chinese, resulting in another 1.2% increase in absolute F_1 score over the auto model.

Comparing against Burkett et al. (2010) (second row from the top), we can see that both our method and Burkett et al. (2010) significantly outperform the monolingual CRF baseline. This suggests that methods that explore bilingual language cues do have great utility in the NER task. Our best model (auto+aP) gives a significant gain over Burkett et al. (2010) on Chinese (by 2.2%), but trails behind on English by 0.7%. However, we will show in the next section some further improvements to our method by modeling global label consistency, which allows us to outperform Burkett et al. (2010) on both languages.

Results on Global Consistency

Table 3 shows results on the test set after factoring in a global consistency model. Adding global consistency to the monolingual baseline (mono) increases performance on English (consistent with results from previous work (Finkel, Grenager, and Manning 2005)), but hurts Chinese results, especially in recall.

A possible explanation is that CRF models for English are more certain about which words are entities (by having strong indicative features such as word capitalization), and thus a penalty does not persuade the model to label a word as a non-entity. However, in the Chinese case, the CRF model is weaker, and thus less certain about words being an entity or not. It is also much more likely that the same word

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

	Chinese			English		
	P	R	F$_1$	P	R	F$_1$
mono	76.89	61.64	68.42	81.98	74.59	78.11
+*global*	77.30	58.96	66.90	83.89	74.88	79.13
+*global'*	75.23	68.12	71.50	82.31	77.63	79.90
auto	78.53	66.90	72.25	82.11	75.40	78.62
+*global*	79.02	64.57	71.07	**84.02**	75.73	79.66
+*global'*	76.17	71.04	73.52	82.87	**78.84**	**80.81**
auto+aP	79.17	68.46	73.43	82.05	75.56	78.67
+*global*	**79.31**	65.93	72.01	84.01	75.81	79.70
+*global'*	76.43	**72.32**	**74.32**	82.30	78.35	80.28

Table 3: Results of enforcing global consistency. *global* is the global consistency without "reward" parameter, and *global'* is the one with "reward" parameter. "mono" is the monolingual CRF baseline. Best number in each column is highlighted in bold.

(string) will be both an entity and a common word in Chinese than English. In some cases, the model will be better off marking a word as a non-entity, than risking taking a penalty for labeling it inconsistently. By applying the "reward" function, we see a drastic increase in recall on both Chinese and English, with a relatively small sacrifice in precision on Chinese. The overall F$_1$ score increases by about 3.1% and 0.8% in Chinese and English, respectively.

Similar results can be found when we apply global consistency to the bilingual model (`auto`). Again we see a recall-precision tradeoff between models with or without a "reward" function. But overall, we observe a significant increase in performance when global consistency with a reward function is factored in.

Modeling alignment uncertainty continues to improve the Chinese results when the global consistency model is added, but shows a small performance decrease on the English side. But the gain on the Chinese side is more significant than the loss on English side.

The best overall F$_1$ scores are achieved when bilingual constraints, global consistency with reward, and alignment uncertainty are conjoined. The combined model outperforms the CRF monolingual baseline, with an error reduction of 18.6% for Chinese and 9.9% for English. This model also significantly improves over the method of Burkett et al. (2010) with an error reduction of 10.8% for Chinese and 4.5% for English.

Beyond the difference in model performance, our method is much easier to understand and implement than Burkett et al. (2010). Their method involves simulating a multi-view learning environment using "weakened" monolingual models to train a reranking model, and transplanting the parameters of the "weakened" models to "strong" models at test time in a practical but ad-hoc manner.

Semi-supervised NER Results

In the previous section we demonstrated the utility of our proposed method in a bilingual setting, where parallel sentence pairs are tagged together and directly evaluated. In reality, this is not the common use case. Most down-stream NLP applications operate in a monolingual environment. Therefore, in order to benefit general monolingual NLP systems, we propose a semi-supervised learning setting where we use the bilingual tagger to annotate a large amount of unannotated bilingual text, then we take the tagged sentences on the Chinese side to retrain a monolingual Chinese tagger.

Method	# train sent	P	R	F$_1$
CRF baseline	$\sim$16k	76.89	61.64	68.42
Self-training	+80k	75.15	59.06	66.14
Semi with Burkett	+80k	76.30	63.46	**69.29**
Semi with auto+aP	+80k	77.40	66.10	**71.31**
	+40k	76.97	65.60	**70.83**
	+10k	77.48	64.95	**70.66**

Table 4: Semi-supervised results on Chinese test set. F$_1$ scores that are statistically significantly better than the CRF baseline is highlighted in bold.

To evaluate the effectiveness of this approach, we used the Chinese-English part of the Foreign Broadcast Information Service corpus (FBIS, LDC2003E14), and tagged it with the `auto+aP` model. Unlike the OntoNotes dataset, this corpus does not contain document boundaries. In order to apply the document-level label consistency model, we divide the test set into blocks of ten sentences, and use the blocks as pseudo-documents.

Results from self-training, as well as results from uptraining using model outputs from Burkett et al. (2010) are shown in Table 4. We can see that by using 80,000 additional sentences, our method gives a significant boost ($\sim$2.9%, an error reduction of $\sim$9.2%) over the CRF baseline. Our method also improves over Burkett et al. (2010) by a significant margin.

The gains are more pronounced in recall than precision, which suggests that the semi-supervised approach using bilingual data is very effective in increasing the coverage of the monolingual tagger. On the other hand, monolingual self-training hurts performance in both precision and recall.

We also report results on the effect of using increasing amounts of unannotated bilingual data. When only 10k sentences are added to the Chinese side, we already see a 5.2% error reduction over the CRF baseline.

Conclusions

We introduced a factored model with a Gibbs sampling inference algorithm, that can be used to produce more accurate tagging results for a parallel corpus. Our model makes use of cross-language bilingual constraints and intra-document consistency constraints. We further demonstrated that unlabeled parallel corpora tagged with our bilingual model can then be used to improve monolingual tagging results, using an uptraining scheme. The model presented here is not restricted to the NER task only, but can be adopted to improve other natural language applications as well, such as syntactic parsing and semantic analysis.

Acknowledgments

The authors would like to thank Rob Voigt and the four 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

Ando, R. K., and Zhang, T. 2005. A high-performance semi-supervised learning method for text chunking. In *Proceedings of ACL*.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*.

Bunescu, R., and Mooney, R. J. 2004. Collective information extraction with relational markov networks. In *Proceedings of ACL*.

Burkett, D.; Petrov, S.; Blitzer, J.; and Klein, D. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of CoNLL*.

Collins, M., and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of EMNLP*.

Das, D., and Petrov, S. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*.

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

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Fu, R.; Qin, B.; and Liu, T. 2011. Generating chinese named entity data from a parallel corpus. In *Proceedings of IJCNLP*.

Ganchev, K.; Graca, J.; Blitzer, J.; and Taskar, B. 2008. Multi-view learning over structured and non-identical outputs. In *Proceedings of UAI*.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transitions on Pattern Analysis and Machine Intelligence* 6:721–741.

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2006. Ontonotes: the 90% solution. In *Proceedings of NAACL-HLT*.

Kim, S.; Toutanova, K.; and Yu, H. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of ACL*.

Kirkpatrick, S.; Gelatt, C. D.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Li, Q.; Li, H.; Ji, H.; Wang, W.; Zheng, J.; and Huang, F. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of CIKM*.

Liang, P.; Taskar, B.; and Klein, D. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*.

Petrov, S.; Chang, P.-C.; Ringgaard, M.; and Alshawi, H. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of EMNLP*.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI*.

Rush, A. M.; Reichert, R.; Collins, M.; and Globerson, A. 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. In *Proceedings of EMNLP*.

Sutton, C., and McCallum, A. 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*.

Yarowsky, D., and Ngai, G. 2001. Inducing multilingual POS taggers and NP brackets via robust projection across aligned corpora. In *Proceedings of NAACL*.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL*.