Cross-lingual Projected Expectation Regularization for Weakly Supervised Learning

Mengqiu Wang and **Christopher D. Manning**
Computer Science Department
Stanford University
Stanford, CA 94305 USA
{mengqiu,manning}@cs.stanford.edu

Abstract

We consider a multilingual weakly supervised learning scenario where knowledge from annotated corpora in a resource-rich language is transferred via bitext to guide the learning in other languages. Past approaches project labels across bitext and use them as features or gold labels for training. We propose a new method that projects model expectations rather than labels, which facilities transfer of model uncertainty across language boundaries. We encode expectations as constraints and train a discriminative CRF model using Generalized Expectation Criteria (Mann and McCallum, 2010). Evaluated on standard Chinese-English and German-English NER datasets, our method demonstrates F_1 scores of 64% and 60% when no labeled data is used. Attaining the same accuracy with supervised CRFs requires 12k and 1.5k labeled sentences. Furthermore, when combined with labeled examples, our method yields significant improvements over state-of-the-art supervised methods, achieving best reported numbers to date on Chinese OntoNotes and German CoNLL-03 datasets.

1 Introduction

Supervised statistical learning methods have enjoyed great popularity in Natural Language Processing (NLP) over the past decade. The success of supervised methods depends heavily upon the availability of large amounts of annotated training data. Manual curation of annotated corpora is a costly and time consuming process. To date, most annotated resources resides within the English language, which hinders the adoption of supervised learning methods in many multilingual environments.

To minimize the need for annotation, significant progress has been made in developing unsupervised and semi-supervised approaches to NLP (Collins and Singer 1999; Klein 2005; Liang 2005; Smith 2006; Goldberg 2010; *inter alia*) . More recent paradigms for semi-supervised learning allow modelers to directly encode knowledge about the task and the domain as constraints to guide learning (Chang et al., 2007; Mann and McCallum, 2010; Ganchev et al., 2010). However, in a multilingual setting, coming up with effective constraints require extensive knowledge of the foreign[1] language.

Bilingual parallel text (bitext) lends itself as a medium to transfer knowledge from a resource-rich language to a foreign languages. Yarowsky and Ngai (2001) project labels produced by an English tagger to the foreign side of bitext, then use the projected labels to learn a HMM model. More recent work applied the projection-based approach to more language-pairs, and further improved performance through the use of type-level constraints from tag dictionary and feature-rich generative or discriminative models (Das and Petrov, 2011; Täckström et al., 2013).

In our work, we propose a new projection-based method that differs in two important ways. First, we never explicitly project the labels. Instead, we project expectations over the labels. This projection

---

[1]For experimental purposes, we designate English as the resource-rich language, and other languages of interest as "foreign". In our experiments, we simulate the resource-poor scenario using Chinese and German, even though in reality these two languages are quite rich in resources.

acts as a soft constraint over the labels, which allows us to transfer more information and uncertainty across language boundaries. Secondly, we encode the expectations as constraints and train a model by minimizing divergence between model expectations and projected expectations in a Generalized Expectation (GE) Criteria (Mann and McCallum, 2010) framework.

We evaluate our approach on Named Entity Recognition (NER) tasks for English-Chinese and English-German language pairs on standard public datasets. We report results in two settings: a weakly supervised setting where no labeled data or a small amount of labeled data is available, and a semi-supervised settings where labeled data is available, but we can gain predictive power by learning from unlabeled bitext.

2 Related Work

Most semi-supervised learning approaches embody the principle of learning from constraints. There are two broad categories of constraints: multi-view constraints, and external knowledge constraints.

Examples of methods that explore multi-view constraints include self-training (Yarowsky, 1995; McClosky et al., 2006),[2] co-training (Blum and Mitchell, 1998; Sindhwani et al., 2005), multi-view learning (Ando and Zhang, 2005; Carlson et al., 2010), and discriminative and generative model combination (Suzuki and Isozaki, 2008; Druck and McCallum, 2010).

An early example of using knowledge as constraints in weakly-supervised learning is the work by Collins and Singer (1999). They showed that the addition of a small set of "seed" rules greatly improve a co-training style unsupervised tagger. Chang et al. (2007) proposed a constraint-driven learning (CODL) framework where constraints are used to guide the selection of best self-labeled examples to be included as additional training data in an iterative EM-style procedure. The kind of constraints used in applications such as NER are the ones like "the words CA, Australia, NY are LOCATION" (Chang et al., 2007). Notice the similarity of this partic-

ular constraint to the kinds of features one would expect to see in a discriminative MaxEnt model. The difference is that instead of learning the validity (or weight) of this feature from labeled examples — since we do not have them — we can constrain the model using our knowledge of the domain. Druck et al. (2009) also demonstrated that in an active learning setting where annotation budget is limited, it is more efficient to label features than examples. Other sources of knowledge include lexicons and gazetteers (Druck et al., 2007; Chang et al., 2007).

While it is straight-forward to see how resources such as a list of city names can give a lot of mileage in recognizing locations, we are also exposed to the danger of over-committing to hard constraints. For example, it becomes problematic with city names that are ambiguous, such as Augusta, Georgia.[3] To soften these constraints, Mann and McCallum (2010) proposed the Generalized Expectation (GE) Criteria framework, which encodes constraints as a regularization term over some score function that measures the divergence between the model's expectation and the target expectation. The connection between GE and CODL is analogous to the relationship between hard (Viterbi) EM and soft EM, as illustrated by Samdani et al. (2012).

Another closely related work is the Posterior Regularization (PR) framework by Ganchev et al. (2010). In fact, as Bellare et al. (2009) have shown, in a discriminative model these two methods optimize exactly the same objective.[4] The two differ in optimization details: PR uses a EM algorithm to approximate the gradients which avoids the expensive computation of a covariance matrix between features and constraints, whereas GE directly calculates the gradient. However, later results (Druck, 2011) have shown that using the Expectation Semiring techniques of Li and Eisner (2009), one can compute the exact gradients of GE in a Conditional Random Fields (CRF) (Lafferty et al., 2001) at costs

---

[2]A multi-view interpretation of self-training is that the self-tagged additional data offers new views to learners trained on existing labeled data.

[3]This is a city in the state of Georgia in USA, famous for its golf courses. It is ambiguous since both Augusta and Georgia can also be used as person names.

[4]The different terminology employed by GE and PR may be confusing to discerning readers, but the "expectation" in the context of GE means the same thing as "marginal posterior" as in PR.

no greater than computing the gradients of ordinary CRF. And empirically, GE tends to perform more accurately than PR (Bellare et al., 2009; Druck, 2011).

Obtaining appropriate knowledge resources for constructing constraints remain as a bottleneck in applying GE and PR to new languages. However, a number of past work recognizes parallel bitext as a rich source of linguistic constraints, naturally captured in the translations. As a result, bitext has been effectively utilized for unsupervised multilingual grammar induction (Alshawi et al., 2000; Snyder et al., 2009), parsing (Burkett and Klein, 2008), and sequence labeling (Naseem et al., 2009).

A number of recent work also explored bilingual constraints in the context of simultaneous bilingual tagging, and showed that enforcing agreements between language pairs give superior results than monolingual tagging (Burkett et al., 2010; Che et al., 2013; Wang et al., 2013a). Burkett et al. (2010) also demonstrated a *uptraining* (Petrov et al., 2010) setting where tag-induced bitext can be used as additional monolingual training data to improve monolingual taggers. A major drawback of this approach is that it requires a readily-trained tagging models in each languages, which makes a weakly supervised setting infeasible. Another intricacy of this approach is that it only works when the two models have comparable strength, since mutual agreements are enforced between them.

Projection-based methods can be very effective in weakly-supervised scenarios, as demonstrated by Yarowsky and Ngai (2001), and Xi and Hwa (2005). One problem with projected labels is that they are often too noisy to be directly used as training signals. To mitigate this problem, Das and Petrov (2011) designed a label propagation method to automatically induce a tag lexicon for the foreign language to smooth the projected labels. Fossum and Abney (2005) filter out projection noise by combining projections from from multiple source languages. However, this approach is not always viable since it relies on having parallel bitext from multiple source languages. Li et al. (2012) proposed the use of crowd-sourced Wiktionary as additional resources for inducing tag lexicons. More recently, Täckström et al. (2013) combined token-level and type-level constraints to constrain legitimate label sequences and and recalibrate the probability distribution in a CRF. The tag dictionary used for POS tagging are analogous to the gazetteers and name lexicons used for NER by Chang et al. (2007).

Our work is also closely related to Ganchev et al. (2009). They used a two-step projection method similar to Das and Petrov (2011) for dependency parsing. Instead of using the projected linguistic structures as ground truth (Yarowsky and Ngai, 2001), or as features in a generative model (Das and Petrov, 2011), they used them as constraints in a PR framework. Our work differs by projecting expectations rather than Viterbi one-best labels. We also choose the GE framework over PR. Experiments in Bellare et al. (2009) and Druck (2011) suggest that in a discriminative model (like ours), GE is more accurate than PR. More recently, Ganchev and Das (2013) further extended this line of work to directly train discriminative sequence models using cross lingual projection with PR. The types of constraints applied in this new work are similar to the ones in the monolingual PR setting proposed by Ganchev et al. (2010), where the total counts of labels of a particular kind are expected to match some fraction of the projected total counts. Our work differ in that we enforce expectation constraints at token level, which gives tighter guidance to learning the model.

3 Approach

Given bitext between English and a foreign language, our goal is to learn a CRF model in the foreign language from little or no labeled data. Our method performs **C**ross-**Li**ngual **P**rojected **E**xpectation **R**egularization (**CLiPER**).

For every aligned sentence pair in the bitext, we first compute the posterior marginal at each word position on the English side using a pre-trained English CRF tagger; then for each aligned English word, we project its posterior marginal as expectations to the aligned word position on the foreign side. Figure 1 shows a snippet of a sentence from real corpus. Notice that if we were to directly project the Viterbi best assignment from English to Chinese, all three Chinese words that are named entities would have gotten the wrong tags. But projecting the English CRF model expectations preserves some uncertainties, informing the Chinese model that there is a 40%

O:1.0000 PER:0.6369 PER:0.6377 O:1.0000 O:1.0000 PER:0.5925 PER:0.5925
LOC:0.0000 LOC:0.3250 LOC:0.3256 LOC:0.0000 LOC:0.0000 ORG:0.4060 ORG:0.4061
ORG:0.0000 ORG:0.0308 ORG:0.0307 ORG:0.0000 ORG:0.0000 O:0.0012 O:0.0011
GPE:0.0000 GPE:0.0042 GPE:0.0042 GPE:0.0000 GPE:0.0000 LOC:0.0003 LOC:0.0003
PER:0.0000 O:0.0032 O:0.0037 PER:0.0000 PER:0.0000 GPE:0.0000 GPE:0.0000

a   reception   in   Luobu   Linka   ······   met   with   representatives   of   Zhongguo   Ribao

在   罗布林卡   举行   的   招待会   ······   会见   了   中国   日报   代表

PER:0.6373 (罗布林卡) O:1.0000 (招待会) O:1.0000 (会见) PER:0.5925 (中国) PER:0.5925 (日报) O:1.0000 (代表)
LOC:0.3253 LOC:0.0000 LOC:0.0000 ORG:0.4060 ORG:0.4061 LOC:0.0000
ORG:0.0307 ORG:0.0000 ORG:0.0000 O:0.0012 O:0.0011 ORG:0.0000
GPE:0.0042 GPE:0.0000 GPE:0.0000 LOC:0.0003 LOC:0.0003 GPE:0.0000
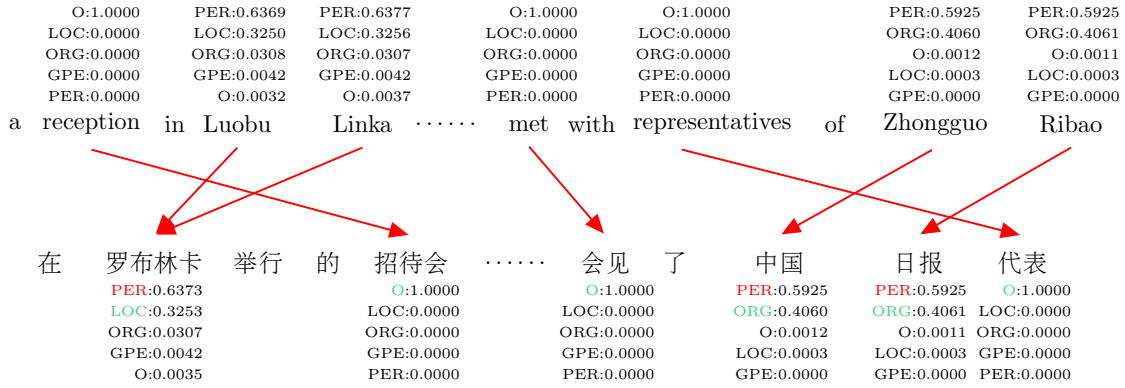O:0.0035 PER:0.0000 PER:0.0000 GPE:0.0000 GPE:0.0000 PER:0.0000

Figure 1: Diagram illustrating the projection of model expectation from English to Chinese. The posterior probabilities assigned by the English CRF model is shown above each English word; automatically induced word alignments are shown in red; the correct projected labels for Chinese words are shown in green, and incorrect labels are shown in red.

chance that "中国日报" (*China Daily*) is an organization in this context.

We would like to learn a CRF model in the foreign language that has similar expectations as the projected expectations from English. To this end, we adopt the Generalized Expectation (GE) Criteria framework introduced by Mann and McCallum (2010). In the remainder of this section, we follow the notation used in (Druck, 2011) to explain our approach.

3.1 CLiPER

The general idea of GE is that we can express our preferences over models through constraint functions. A desired model should satisfy the imposed constraints by matching the expectations on these constraint functions with some target expectations (attained by external knowledge like lexicons or in our case transferred knowledge from English). We define a constraint function ϕ_{i,l_j} for each word position i and output label assignment l_j. $\phi_{i,l_j} = 0$ is a constraint in that position i cannot take label l_j.

The set $\{l_1, \cdots, l_m\}$ denotes all possible label assignment for each y_i, and m is number of label values. A_i is the set of English words aligned to Chinese word i. ϕ_{i,l_j} are defined for all position i such that $A_i \neq \emptyset$. In other words, the constraint function applies only to Chinese word positions that have at least one aligned English word. Each $\phi_{i,l_j}(\mathbf{y})$ can be treated as a Bernoulli random variable, and we concatenate the set of all ϕ_{i,l_j} into a *random vector*

$\phi(\mathbf{y})$, where $\phi_k = \phi_{i,l_j}$ if $k = i * m + j$. We drop the $(\mathbf{y})$ in ϕ for simplicity.

The target expectation over ϕ_{i,l_j}, denoted as $\tilde{\phi}_{i,l_j}$, is the expectation of assigning label l_j to English word A_i under the English conditional probability model. When multiple English words are aligned to the same foreign word, we average the expectations.

The expectation over ϕ under a conditional probability model $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$ is denoted as $E_{P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}[\phi]$, and simplified as $E_{\boldsymbol{\theta}}[\phi]$ whenever it is unambiguous.

The conditional probability model $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$ in our case is defined as a standard linear-chain CRF:[5]

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x};\boldsymbol{\theta})}\exp\left(\sum_i^n \boldsymbol{\theta}\mathbf{f}(\mathbf{x}, y_i, y_{i-1})\right)$$

where $\mathbf{f}$ is a set of feature functions; $\boldsymbol{\theta}$ are the matching parameters to learn; $n = |\mathbf{x}|$.

The objective function to maximize in a standard CRF is the log probability over a collection of labeled documents:

$$L_{CRF}(\boldsymbol{\theta}) = \sum_{a=1}^{a'} \log P(\mathbf{y}_a^*|\mathbf{x}_a;\boldsymbol{\theta}) \tag{1}$$

a' is the number of labeled sentences. $\mathbf{y}^*$ is an observed label sequence.

The objective function to maximize in GE is defined as the sum over all unlabeled examples on the

---

[5] We simplify notation by dropping the L_2 regularizer in the CRF definition, but apply it in our experiments.

foreign side of bitext, denoted as $\mathbf{x}_b$, over some cost function S between the model expectation over ϕ ($E_{\boldsymbol{\theta}}[\phi]$) and the target expectation ($\tilde{\phi}$).

We choose S to be the negative L_2^2 squared error sum[6] defined as:

$$L_{GE}(\boldsymbol{\theta}) = \sum_{b=1}^{n'} S\left(E_{P(\mathbf{y}_b|\mathbf{x}_b;\boldsymbol{\theta})}[\phi(\mathbf{y}_b)], \tilde{\phi}_b\right)$$

$$= \sum_{b=1}^{n'} -\|\tilde{\phi}_b - E_{\boldsymbol{\theta}}[\phi(\mathbf{y}_b)]\|_2^2 \qquad (2)$$

n' is the total number of unlabeled bitext sentence pairs.

When both labeled and bitext training data are available, the joint objective is the sum of Eqn. 1 and 2. Each is computed over the labeled training data and foreign half in the bitext, respectively.

We can optimize this joint objective by computing the gradients and use a gradient-based optimization method such as L-BFGS. Gradients of L_{CRF} decomposes down to the gradients over each labeled training example $(\mathbf{x}, \mathbf{y}^*)$. Computing the gradient of L_{GE} decomposes down to the gradients of $S(E_{P(\mathbf{y}|\mathbf{x}_b;\boldsymbol{\theta}}[\phi])$ for each unlabeled foreign sentence $\mathbf{x}$ and the constraints over this example ϕ. The gradients can be calculated as:

$$\frac{\partial}{\partial\boldsymbol{\theta}}S(E_{\boldsymbol{\theta}}[\phi]) = -\frac{\partial}{\partial\boldsymbol{\theta}}\left(\tilde{\phi} - E_{\boldsymbol{\theta}}[\phi]\right)^T\left(\tilde{\phi} - E_{\boldsymbol{\theta}}[\phi]\right)$$

$$= 2\left(\tilde{\phi} - E_{\boldsymbol{\theta}}[\phi]\right)^T\left(\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}[\phi]\right)$$

We redefine the penalty vector $\mathbf{u} = 2\left(\tilde{\phi} - E_{\boldsymbol{\theta}}[\phi]\right)$ to be u. $\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}[\phi]$ is a matrix where each column contains the gradients for a particular model feature θ with respect to all constraint functions ϕ. It can be

---

[6]In general, other loss functions such as KL-divergence can also be used for S. We found L_2^2 to work well in practice.

computed as:

$$\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}[\phi] = \sum_{\mathbf{y}}\phi(\mathbf{y})\frac{\partial}{\partial\boldsymbol{\theta}}P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$$

$$= \sum_{\mathbf{y}}\phi(\mathbf{y})\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{Z(\mathbf{x};\boldsymbol{\theta})}\exp(\boldsymbol{\theta}^T\mathbf{f}(\mathbf{x},\mathbf{y}))\right)$$

$$= \sum_{\mathbf{y}}\phi(\mathbf{y})\left(\frac{1}{Z(\mathbf{x};\boldsymbol{\theta})}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\exp(\boldsymbol{\theta}^T\mathbf{f}(\mathbf{x},\mathbf{y}))\right)\right.$$
$$\left.+ \exp(\boldsymbol{\theta}^T\mathbf{f}(\mathbf{x},\mathbf{y}))\left(\frac{\partial}{\partial\boldsymbol{\theta}}\frac{1}{Z(\mathbf{x};\boldsymbol{\theta})}\right)\right)$$

$$= \sum_{\mathbf{y}}\phi(\mathbf{y})\left(P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\mathbf{f}(\mathbf{x},\mathbf{y})^T\right.$$
$$\left. - P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\sum_{\mathbf{y}'}P(\mathbf{y}'|\mathbf{x};\boldsymbol{\theta})\mathbf{f}(\mathbf{x},\mathbf{y}')^T\right)$$

$$= \sum_{\mathbf{y}}P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\sum_{\mathbf{y}}\phi(\mathbf{y})\mathbf{f}(\mathbf{x},\mathbf{y})^T$$
$$- \left(\sum_{\mathbf{y}}P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\phi(\mathbf{y})\right)\left(\sum_{\mathbf{y}}P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\mathbf{f}(\mathbf{x},\mathbf{y})^T\right)$$

$$= \mathrm{COV}_{P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}\left(\phi(\mathbf{y}), \mathbf{f}(\mathbf{x},\mathbf{y})\right) \qquad (3)$$

$$= E_{\boldsymbol{\theta}}[\phi\mathbf{f}^T] - E_{\boldsymbol{\theta}}[\phi]E_{\boldsymbol{\theta}}[\mathbf{f}^T] \qquad (4)$$

Eqn. 3 gives the intuition of how optimization works in GE. In each iteration of L-BFGS, the model parameters are updated according to their covariance with the constraint features, scaled by the difference between current expectation and target expectation. The term $E_{\boldsymbol{\theta}}[\phi\mathbf{f}^T]$ in Eqn. 4 can be computed using a dynamic programming (DP) algorithm, but solving it directly requires us to store a matrix of the same dimension as $\mathbf{f}^T$ in each step of the DP. We can reduce the complexity by using the same trick as in (Li and Eisner, 2009) for computing Expectation Semiring. The resulting algorithm has complexity $O(nm^2)$, which is the same as the standard forward-backward inference algorithm for CRF. (Druck, 2011, 93) gives full details of this derivation.

3.2 Hard vs. soft Projection

Projecting expectations instead of one-best label assignments from English to foreign language can be thought of as a soft version of the method described in (Das and Petrov, 2011) and (Ganchev et

al., 2009). Soft projection has its advantage: when the English model is not certain about its predictions, we do not have to commit to the current best prediction. The foreign model has more freedom to form its own belief since any marginal distribution it produces would deviates from a flat distribution by just about the same amount. In general, preserving uncertainties till later is a strategy that has benefited many NLP tasks (Finkel et al., 2006). Hard projection can also be treated as a special case in our framework. We can simply recalibrate posterior marginal of English by assigning probability mass 1 to the most likely outcome, and zero everything else out, effectively taking the argmax of the marginal at each word position. We refer to this version of expectation as the "hard" expectation. In the hard projection setting, GE training resembles a "project-then-train" style semi-supervised CRF training scheme (Yarowsky and Ngai, 2001; Täckström et al., 2013). In such a training scheme, we project the one-best predictions of English CRF to the foreign side through word alignments, then include the newly "tagged" foreign data as additional training data to a standard CRF in the foreign language. Rather than projecting labels on a per-word basis, Yarowsky and Ngai (2001) also explored an alternative method for noun-phrase (NP) bracketing task that amounts to projecting the spans of NPs based on the observation that individual NPs tend to retain their sequential spans across translations. We experimented with the same method for NER, but found that this method of projecting the NE spans does not help in reducing noise and actually lowers model performance.

Besides the difference in projecting expectations rather than hard labels, our method and the "project-then-train" scheme also differ by optimizing different objectives: CRF optimizes maximum conditional likelihood of the observed label sequence, whereas GE minimizes squared error between model's expectation and "hard" expectation based on the observed label sequence. In the case where squared error loss is replaced with a KL-divergence loss, GE has the same effect as marginalizing out all positions with unknown projected labels, allowing more robust learning of uncertainties in the model. As we will show in the experimen-

	O	PER	LOC	ORG	GPE
O	291339	391	141	1281	221
PER	1263	6721	5	56	73
LOC	409	23	546	123	133
ORG	2423	143	52	8387	196
GPE	566	239	69	668	6604

	O	PER	LOC	ORG	MISC
O	81209	24	38	155	103
PER	77	5725	41	69	10
LOC	49	40	3743	127	60
ORG	178	102	142	4075	91
MISC	175	41	30	114	1826

Table 1: Raw counts in the error confusion matrix of English CRF models. Top table contains the counts on OntoNotes test data, and bottom table contains CoNLL-03 test data counts. Rows are the true labels and columns are the observed labels. For example, item at row 2, column 3 of the top table reads: we observed 5 times where the true label should be PERSON, but English CRF model output label LOCATION.

tal results in Section 4.2, soft projection in combination of the GE objective significantly outperforms the project-then-train style CRF training scheme.

3.3 Source-side noise

An additional source of noise comes from errors generated by the source-side English CRF models. We know that the English CRF models gives F_1 score of 81.68% on the OntoNotes dataset for English-Chinese experiment, and 90.45% on the CoNLL-03 dataset for English-German experiment. We present a simple way of modeling English-side noise by picturing the following process: the labels assigned by the English CRF model (denoted as $\mathbf{y}$) are some noised version of the true labels (denoted as $\mathbf{y}^*$). We can recover the probability of the true labels by marginalizing over the observed labels: $P(\mathbf{y}^*|\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{y}^*|\mathbf{y}) * P(\mathbf{y}|\mathbf{x})$. $P(\mathbf{y}|\mathbf{x})$ is the posterior probabilities given by the CRF model, and we can approximate $P(\mathbf{y}^*|\mathbf{y})$ by the column-normalized error confusion matrix shown in Table 1. This source-side noise model is likely to be overly simplistic. Generally speaking, we could build much more sophisticated noising model for the source-side, possibly conditioning on context, or capturing higher-order label sequences.

4 Experiments

We conduct experiments on Chinese and German NER. We evaluate CLiPER in two learning settings: weakly supervised and semi-supervised. In the weakly supervised setting, we simulate the condition of having no labeled training data, and evaluate the model learned from bitext alone. We then vary the amount of labeled data available to the model, and examine the model's learning curve. In the semi-supervised setting, we assume our model has access to the full labeled data; our goal is to improve performance of the supervised method by learning from additional bitext.

4.1 Dataset and setup

We used the latest version of Stanford NER Toolkit[7] as our base CRF model in all experiments. Features for English, Chinese and German CRFs are documented extensively in (Che et al., 2013) and (Faruqui and Padó, 2010) and omitted here for brevity. It it worth noting that the current Stanford NER models include recent improvements from semi-supervise learning approaches that induces distributional similarity features from large word clusters. These models represent the current state-of-the-art in supervised methods, and serve as a very strong baseline.

For Chinese NER experiments, we follow the same setup as Che et al. (2013) to evaluate on the latest OntoNotes (v4.0) corpus (Hovy et al., 2006).[8] A total of 8,249 sentences from the parallel Chinese and English Penn Treebank portion [9] are reserved for evaluation. Odd-numbered documents are used as development set, and even-numbered documents are held out as blind test set. The rest of OntoNotes annotated with NER tags are used to train the English and Chinese CRF base taggers. There are about 16k and 39k labeled sentences for Chinese and English training, respectively. The English CRF tagger trained on this training corpus gives F_1 score of 81.68% on the OntoNotes test set. Four entities types[10] are used for both Chinese and English with a IO tagging scheme.[11] The English-Chinese

bitext comes from the Foreign Broadcast Information Service corpus (FBIS).[12] We randomly sampled 80k parallel sentence pairs to use as bitext in our experiments. It is first sentence aligned using the Champollion Tool Kit,[13] then word aligned with the BerkeleyAligner.[14]

For German NER experiments, we evaluate using the standard CoNLL-03 NER corpus (Sang and Meulder, 2003). The labeled training set has 12k and 15k sentences, containing four entity types.[15] An English CRF model is also trained on the CoNLL-03 English data with the same entity types. For bitext, we used a randomly sampled set of 40k parallel sentences from the de-en portion of the *News Commentary* dataset.[16] The English CRF tagger trained on CoNLL-03 English training corpus gives F_1 score of 90.4% on the CoNLL-03 test set.

We report typed entity precision (P), recall (R) and F_1 score. Statistical significance tests are done using a paired bootstrap resampling method with 1000 iterations, averaged over 5 runs. We compare against three recently approaches that were introduced in Section 2. They are: semi-supervised learning method using factored bilingual models with Gibbs sampling (Wang et al., 2013a); bilingual NER using Integer Linear Programming (ILP) with bilingual constraints, by (Che et al., 2013); and constraint-driven bilingual-reranking approach (Burkett et al., 2010). The code from (Che et al., 2013) and (Wang et al., 2013a) are publicly available.[17] Code from (Burkett et al., 2010) is obtained through personal communications.

Since the objective function in Eqn. 2 is non-convex, we adopted the early stopping training scheme from (Turian et al., 2010) as the following: after each iteration in L-BFGS training, the model

---

[7] http://www-nlp.stanford.edu/ner

[8] LDC catalogue No.: LDC2011T03

[9] File numbers: chtb_0001-0325, ectb_1001-1078

[10] PERSON, LOCATION, ORGANIZATION and GPE.

[11] We did not adopt the commonly seen BIO tagging scheme (Ramshaw and Marcus, 1999), because when projected across swapping word alignments, the "B-" and "I-" tag distinction may not be well-preserved and may introduce additional noise.

[12] The FBIS corpus is a collection of radio news casts and contains translations of openly available news and information from media sources outside the United States. The LDC catalogue No. is LDC2003E14.

[13] champollion.sourceforge.net

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

[15] PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS.

[16] http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz

[17] https://github.com/stanfordnlp/CoreNLP

is evaluated against the development set; the training procedure is terminated if no improvements have been made in 20 iterations.

4.2 Weakly supervised results

Figure 2a and 2b show results of weakly supervised learning experiments. Quite remarkably, on Chinese test set, our proposed method (CLiPER) achieves a F_1 score of 64.4% with 80k bitext, when no labeled training data is used. In contrast, the supervised CRF baseline would require as much as 12k labeled sentences to attain the same accuracy. Results on the German test set is less striking. With no labeled data and 40k of bitext, CLiPER performs at F_1 of 60.0%, the equivalent of using 1.5k labeled examples in the supervised setting. When combined with 1k labeled examples, performance of CLiPER reaches 69%, a gain of over 5% absolute over supervised CRF. We also notice that supervised CRF model learns much faster in German than Chinese. This result is not too surprising, since it is well recognized that Chinese NER is more challenging than German or English. The best supervised results for Chinese is 10-20% (F_1 score) behind best German and English supervised results. Chinese NER relies more on lexicalized features, and therefore needs more labeled data to achieve good coverage. The results suggest that CLiPER seems to be very effective at transferring lexical knowledge from English to Chinese.

Figure 2c and 2d compares soft GE projection with hard GE projection and the "project-then-train" style CRF training scheme (*cf.* Section 3.2). We observe that both soft and hard GE projection significantly outperform the "project-then-train" style training scheme. The difference is especially pronounced on the Chinese results when fewer labeled examples are available. Soft projection gives better accuracy than hard projection when no labeled data is available, and also has a faster learning rate.

Incorporating source-side noise using the method described in Section 3.3 gives a small improvement on Chinese with supervised data, increasing F_1 score from 64.40% to 65.50%. This improvement is statistically significant at 92% confidence interval. However, on the German data, we observe a tiny decrease with no statistical significance in F_1 score, dropping from 59.88% to 59.66%. A likely explanation of the difference is that the English CRF model in the English-Chinese experiment, which is trained on OntoNotes data, has a much higher error rate (18.32%) than the English CRF model in the English-German experiment trained on CoNLL-03 (9.55%). Therefore, modeling noise in the English-Chinese case is likely to have a greater effect than the English-German case.

4.3 Semi-supervised results

In the semi-supervised experiments, we let the CRF model use the full set of labeled examples in addition to the unlabeled bitext. Results on the test set are shown in Table 2. All semi-supervised baselines are tested with the same number of unlabeled bitext as CLiPER in each language. The "project-then-train" semi-supervised training scheme severely hurts performance on Chinese, but gives a small improvement on German. Moreover, on Chinese it learns to achieve high precision but at a significant loss in recall. On German its behavior is the opposite. Such drastic and erratic imbalance suggest that this method is not robust or reliable. The other three semi-supervised baselines (row 3-5) all show improvements over the CRF baseline, consistent with their reported results. CLiPER$_s$ gives the best results on both Chinese and German, yielding statistically significant improvements over all baselines except for CWD13 on German. The hard projection version of CLiPER also gives sizable gain over CRF. However, in comparison, CLiPER$_s$ is superior.

The improvements of CLiPER$_s$ over CRF on Chinese test set is over 2.8% in absolute F_1. The improvement over CRF on German is almost a percent. To our knowledge, these are the best reported numbers on the OntoNotes Chinese and CoNLL-03 German datasets.

4.4 Efficiency

Another advantage of our proposed approach is efficiency. Because we eliminated the previous multi-stage "uptraining" paradigm, but instead integrating the semi-supervised and supervised objective into one joint objective, we are able to attain significant speed improvements over all methods except CRF$_{ptt}$. Table 3 shows the required training time.
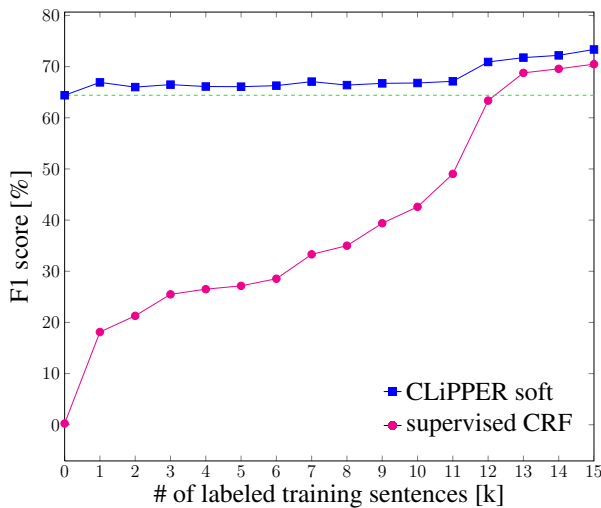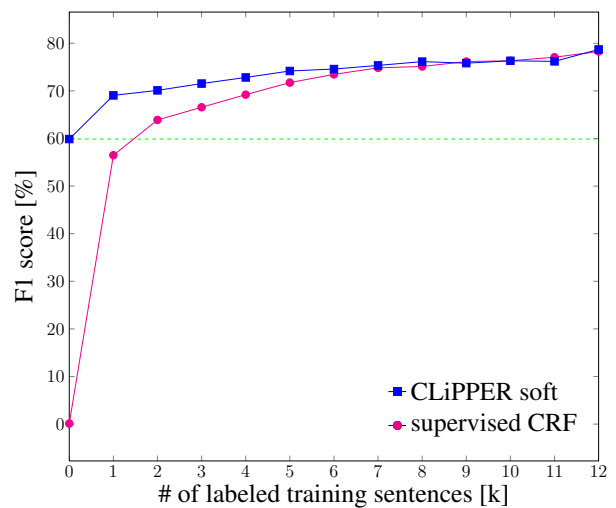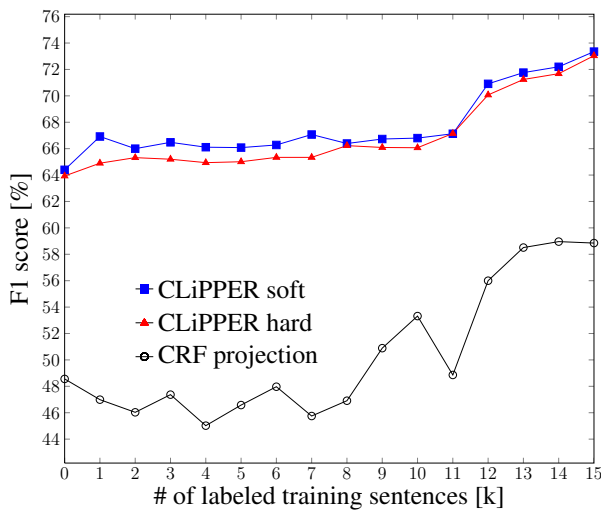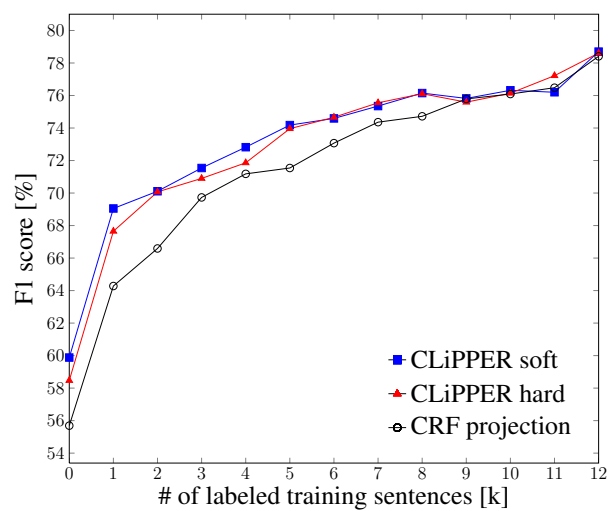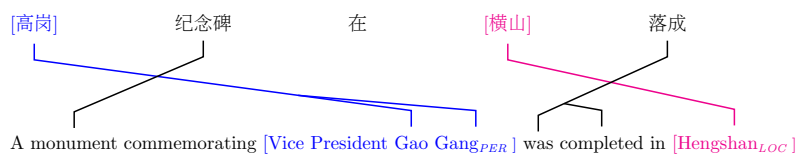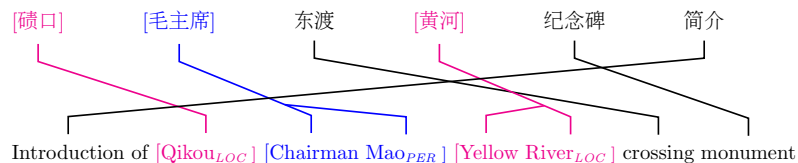
62

Figure 2: Top four figures show performance curves of CLiPER with varying amounts of available labeled training data in a weakly supervised setting. Vertical axes show the F_1 score on the test set. Performance curves of supervised CRF and "project-then-train" CRF are plotted for comparison. Bottom two figures are examples of aligned sentence pairs in Chinese and English.

	Chinese			German		
	P	R	F_1	P	R	F_1
CRF	79.09	63.59	70.50	86.69	71.30	78.25
CRF$_{ptt}$	**84.01**	45.29	58.85	81.50	**75.56**	78.41
BPBK10	79.25	65.67	71.83	84.00	72.17	77.64
CWD13	81.31	65.50	72.55	85.99	72.98	78.95
WCD13a	80.31	65.78	72.33	85.98	72.37	78.59
WCD13b	78.55	66.54	72.05	85.19	72.98	78.62
CLiPER$_h$	83.67	64.80	73.04§‡	86.52	72.02	78.61*
CLiPER$_s$	82.57	**65.99**	**73.35**$^{\S\dagger\star}_{\diamond *}$	**87.11**	72.56	**79.17**$^{\ddagger\star}_{*\S}$

Table 2: Test set Chinese, German NER results. Best number of each column is highlighted in bold. CRF is the supervised baseline. CRF$_{ptt}$ is the "project-then-train" semi-supervised scheme for CRF. BPBK10 is (Burkett et al., 2010), WCD13 is (Wang et al., 2013a), CWD13A is (Che et al., 2013), and WCD13B is (Wang et al., 2013b) . CLiPER$_s$ and CLiPER$_h$ are the soft and hard projections. § indicates F_1 scores that are statistically significantly better than CRF baseline at 99.5% confidence level; ⋆ marks significance over CRF$_{ptt}$ with 99.5% confidence; † and ‡ marks significance over WCD13 with 99.9% and 94% confidence; and ◇ marks significance over CWD13 with 99.7% confidence; ∗ marks significance over BPBK10 with 99.9% confidence.

5 Discussions

Figure 2e and 2f give two examples of cross-lingual projection methods in action. Both examples have a named entity that immediately proceeds the word "纪念碑" (monument) in the Chinese sentence. In Figure 2e, the word "高岗" has literal meaning of *a hillock located at a high position*, which also happens to be the name of a former vice president of China. Without having previously observed this word as a person name in the labeled training data, the CRF model does not have enough evidence to believe that this is a PERSON, instead of LOCATION. But the aligned words in English ("Gao Gang") are clearly part of a person name as they were preceded by a title ("Vice President"). The English model has high expectation that the aligned Chinese word of "Gao Gang" is also a PERSON. Therefore, projecting the English expectations to Chinese provides a strong clue to help disambiguating this word. Figure 2f gives another example: the word "黄河"(Huang He, the Yellow River of China) can

	Chinese	German
CRF	19m30s	7m15s
CRF$_{ptt}$	34m2s	12m45s
WCD13	3h17m	1h1m
CWD13a	16h42m	4h49m
CWD13b	16h42m	4h49m
BPBK10	6h16m	2h42m
CLiPER$_h$	1h28m	16m30s
CLiPER$_s$	1h40m	18m51s

Table 3: Timing stats during model training.

be confused with a person name since "黄"(Huang or Hwang) is also a common Chinese last name.[18]. Again, knowing the translation in English, which has the indicative word "River" in it, helps disambiguation.

The CRF$_{ptt}$ and CLiPER$_h$ methods successfully labeled these two examples correctly, but failed to produce the correct label for the example in Figure 1. On the other hand, a model trained with the CLiPER$_s$ method does correctly label both entities in Figure 1, demonstrating the merits of the soft projection method.

6 Conclusion

We introduced a domain and language independent semi-supervised method for training discriminative models by projecting expectations across bitext. Experiments on Chinese and German NER show that our method, learned over bitext alone, can rival performance of supervised models trained with thousands of labeled examples. Furthermore, applying our method in a setting where all labeled examples are available also shows improvements over state-of-the-art supervised methods. Our experiments also showed that soft expectation projection is more favorable to hard projection. This technique can be generalized to all sequence labeling tasks, and can be extended to include more complex constraints. For future work, we plan to apply this method to more language pairs and also explore data selection strategies and modeling alignment uncertainties.

---

[18]In fact, a people search of the name 黄河 on the most popular Chinese social network (renren.com) returns over 13,000 matches.

Acknowledgments

References

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Head-transducer models for speech translation and their automatic acquisition from bilingual data. *Machine Translation*, 15.

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In *Proceedings of ACL*.

Kedar Bellare, Gregory Druck, and Andrew McCallum. 2009. Alternating projections for learning with expectation constraints. In *Proceedings of UAI*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of EMNLP*.

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

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of WSDM*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Proceedings of ACL*.

Wanxiang Che, Mengqiu Wang, and Christopher D. Manning. 2013. Named entity recognition with bilingual constraints. In *Proceedings of NAACL*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of EMNLP*.

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

Gregory Druck and Andrew McCallum. 2010. High-performance semi-supervised learning using discriminatively constrained generative models. In *Proceedings of ICML*.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2007. Leveraging existing resources using generalized expectation criteria. In *Proceedings of NIPS Workshop on Learning Problem Design*.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of EMNLP*.

Gregory Druck. 2011. *Generalized Expectation Criteria for Lightly Supervised Learning*. Ph.D. thesis, University of Massachusetts Amherst.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS*.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. 2006. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of EMNLP*.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *Proceedings of IJCNLP*.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proceedings of EMNLP*.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of ACL*.

Kuzman Ganchev, Jo ao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *JMLR*, 10:2001–2049.

Andrew B. Goldberg. 2010. *New Directions in Semi-supervised Learning*. Ph.D. thesis, University of Wisconsin-Madison.

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

Dan Klein. 2005. *The Unsupervised Learning of Natural Language Structure*. Ph.D. thesis, Stanford University.

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 ICML*.

Zhifei Li and Jason Eisner. 2009. First- and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proceedings of EMNLP*.

Shen Li, Jo ao Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of EMNLP-CoNLL*.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Gideon Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *JMLR*, 11:955–984.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of NAACL-HLT*.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *JAIR*, 36:1076–9757.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of EMNLP*.

Lance A. Ramshaw and Mitchell P. Marcus. 1999. Text chunking using transformation-based learning. *Natural Language Processing Using Very Large Corpora*, 11:157–176.

Rajhans Samdani, Ming-Wei Chang, and Dan Roth. 2012. Unified expectation maximization. In *Proceedings of NAACL*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of CoNLL*.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views, International Conference on Machine Learning*.

Noah A. Smith. 2006. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. Ph.D. thesis, Johns Hopkins University.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of ACL*.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of ACL*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. In *Proceedings of ACL*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013a. Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *Proceedings of AAAI*.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013b. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of ACL*.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of HLT-EMNLP*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*.

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