Learning Better Monolingual Models with Unannotated Bilingual Text

David Burkett[†] **Slav Petrov**[‡] **John Blitzer**[†] **Dan Klein**[†]

[†]University of California, Berkeley [‡]Google Research
{dburkett,blitzer,klein}@cs.berkeley.edu slav@google.com

Abstract

This work shows how to improve state-of-the-art monolingual natural language processing models using unannotated bilingual text. We build a multiview learning objective that enforces agreement between monolingual and bilingual models. In our method the first, monolingual view consists of supervised predictors learned separately for each language. The second, bilingual view consists of log-linear predictors learned over both languages on bilingual text. Our training procedure estimates the parameters of the bilingual model using the output of the monolingual model, and we show how to combine the two models to account for dependence between views. For the task of named entity recognition, using bilingual predictors increases F_1 by 16.1% absolute over a supervised monolingual model, and retraining on bilingual predictions increases *monolingual* model F_1 by 14.6%. For syntactic parsing, our bilingual predictor increases F_1 by 2.1% absolute, and retraining a monolingual model on its output gives an improvement of 2.0%.

1 Introduction

Natural language analysis in one language can be improved by exploiting translations in another language. This observation has formed the basis for important work on syntax projection across languages (Yarowsky et al., 2001; Hwa et al., 2005; Ganchev et al., 2009) and unsupervised syntax induction in multiple languages (Snyder et al., 2009), as well as other tasks, such as cross-lingual named entity recognition (Huang and Vogel, 2002; Moore, 2003) and information retrieval (Si and Callan, 2005). In all of these cases, multilingual models yield increased accuracy because different languages present different ambiguities and therefore offer complementary constraints on the shared underlying labels.

In the present work, we consider a setting where we already possess supervised monolingual models, and wish to improve these models using *unannotated* bilingual parallel text (bitext). We cast this problem in the multiple-view (multiview) learning framework (Blum and Mitchell, 1998; Collins and Singer, 1999; Balcan and Blum, 2005; Ganchev et al., 2008). Our two views are a *monolingual* view, which uses the supervised monolingual models but not bilingual information, and a *bilingual* view, which exploits features that measure agreement across languages. The parameters of the bilingual view are trained to reproduce the output of the monolingual view. We show that by introducing *weakened* monolingual models into the bilingual view, we can optimize the parameters of the bilingual model to improve monolingual models. At prediction time, we automatically account for the between-view dependence introduced by the weakened monolingual models with a simple but effective view-combination heuristic.

We demonstrate the performance of this method on two problems. The first is named entity recognition (NER). For this problem, our method automatically learns (a variation on) earlier hand-designed rule-based bilingual NER predictors (Huang and Vogel, 2002; Moore, 2003), resulting in absolute performance gains of up to 16.1% F_1. The second task we consider is statistical parsing. For this task, we follow the setup of Burkett and Klein (2008), who improved Chinese and English monolingual parsers using parallel, hand-parsed text. We achieve nearly identical improvements using a purely *unlabeled* bitext. These results carry over to machine translation, where we can achieve slightly better BLEU improvements than the *supervised* model of Burkett and Klein (2008) since we are able to train our model directly on the parallel data where we perform rule extraction.

Finally, for both of our tasks, we use our bilingual model to generate additional automatically labeled *monolingual* training data. We compare

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 46–54,
Uppsala, Sweden, 15-16 July 2010.

this approach to monolingual self-training and show an improvement of up to 14.4% F_1 for entity recognition. Even for parsing, where the bilingual portion of the treebank is much smaller than the monolingual, our technique still can improve over purely monolingual self-training by 0.7% F_1.

2 Prior Work on Learning from Bilingual Text

Prior work in learning monolingual models from bitexts falls roughly into three categories: Unsupervised induction, cross-lingual projection, and bilingual constraints for supervised monolingual models. Two recent, successful unsupervised induction methods are those of Blunsom et al. (2009) and Snyder et al. (2009). Both of them estimate hierarchical Bayesian models and employ bilingual data to constrain the types of models that can be derived. Projection methods, on the other hand, were among the first applications of parallel text (after machine translation) (Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Hwa et al., 2005; Ganchev et al., 2009). They assume the existence of a good, monolingual model for one language but little or no information about the second language. Given a parallel sentence pair, they use the annotations for one language to heavily constrain the set of possible annotations for the other.

Our work falls into the final category: We wish to use bilingual data to improve monolingual models which are already trained on large amounts of data and effective on their own (Huang and Vogel, 2002; Smith and Smith, 2004; Snyder and Barzilay, 2008; Burkett and Klein, 2008). Procedurally, our work is most closely related to that of Burkett and Klein (2008). They used an annotated bitext to learn parse reranking models for English and Chinese, exploiting features that examine pieces of parse trees in both languages. Our method can be thought of as the semi-supervised counterpart to their supervised model. Indeed, we achieve nearly the same results, but without annotated bitexts. Smith and Smith (2004) consider a similar setting for parsing both English and Korean, but instead of learning a joint model, they consider a fixed combination of two parsers and a word aligner. Our model learns parameters for combining two monolingual models and potentially thousands of bilingual features. The result is that our model significantly improves state-of-the-art results, for both parsing and NER.

3 A Multiview Bilingual Model

Given two input sentences $x = (x_1, x_2)$ that are word-aligned translations of each other, we consider the problem of predicting (structured) labels $y = (y_1, y_2)$ by estimating conditional models on pairs of labels from both languages, $p(y_1, y_2|x_1, x_2)$. Our model consists of two views, which we will refer to as monolingual and bilingual. The monolingual view estimates the joint probability as the product of independent marginal distributions over each language, $p_M(y|x) = p_1(y_1|x_1)p_2(y_2|x_2)$. In our applications, these marginal distributions will be computed by state-of-the-art statistical taggers and parsers trained on large monolingual corpora.

This work focuses on learning parameters for the bilingual view of the data. We parameterize the bilingual view using at most one-to-one matchings between *nodes* of structured labels in each language (Burkett and Klein, 2008). In this work, we use the term *node* to indicate a particular component of a label, such as a single (multi-word) named entity or a node in a parse tree. In Figure 2(a), for example, the nodes labeled NP_1 in both the Chinese and English trees are matched. Since we don't know a priori how the components relate to one another, we treat these matchings as hidden. For each matching a and pair of labels y, we define a feature vector $\boldsymbol{\phi}(y_1, a, y_2)$ which factors on edges in the matching. Our model is a conditional exponential family distribution over matchings and labels:

$$p_{\boldsymbol{\theta}}(y, a|x) = \exp\left[\boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - A(\boldsymbol{\theta}; x)\right] ,$$

where $\boldsymbol{\theta}$ is a parameter vector, and $A(\boldsymbol{\theta}; x)$ is the log partition function for a sentence pair x. We must approximate $A(\boldsymbol{\theta}; x)$ because summing over all at most one-to-one matchings a is #P-hard. We approximate this sum using the maximum-scoring matching (Burkett and Klein, 2008):

$$\tilde{A}(\boldsymbol{\theta}; x) = \log \sum_y \max_a \left(\exp\left[\boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2)\right]\right) .$$

In order to compute the distribution on labels y, we must marginalize over hidden alignments between nodes, which we also approximate by using the maximum-scoring matching:

$$q_{\boldsymbol{\theta}}(y|x) \stackrel{\text{def}}{=} \max_a \exp\left[\boldsymbol{\theta}^\top \boldsymbol{\phi}(y_1, a, y_2) - \tilde{A}(\boldsymbol{\theta}; x)\right] .$$

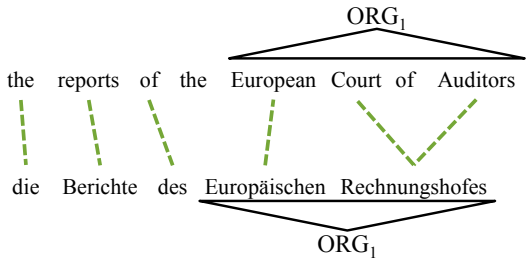

Figure 1: An example where English NER can be used to disambiguate German NER.

We further simplify inference in our model by working in a reranking setting (Collins, 2000; Charniak and Johnson, 2005), where we only consider the top k outputs from monolingual models in both languages, for a total of k^2 labels y. In practice, $k^2 \leq 10,000$ for our largest problem.

3.1 Including Weakened Models

Now that we have defined our bilingual model, we could train it to agree with the output of the monolingual model (Collins and Singer, 1999; Ganchev et al., 2008). As we will see in Section 4, however, the feature functions $\boldsymbol{\phi}(y_1, a, y_2)$ make no reference to the input sentences x, other than through a fixed word alignment. With such limited monolingual information, it is impossible for the bilingual model to adequately capture all of the information necessary for NER or parsing. As a simple example, a bilingual NER model will be perfectly happy to label two aligned person names as *ORG* instead of *PER*: both labelings agree equally well. We briefly illustrate how poorly such a basic bilingual model performs in Section 10.

One way to solve this problem is to include the output of the full monolingual models as features in the bilingual view. However, we are training the bilingual view to match the output of these same models, which can be trivially achieved by putting weight on only the monolingual model scores and never recruiting any bilingual features. Therefore, we use an intermediate approach: we introduce the output of deliberately weakened monolingual models as features in the bilingual view. A weakened model is from the same class as the full monolingual models, but is intentionally crippled in some way (by removing feature templates, for example). Crucially, the weakened models will make predictions that are roughly similar to the full models, but systematically worse. Therefore, model scores from the *weakened* models provide enough power for the bilingual view to make accurate predictions, but ensure that bilingual features will be required to optimize the training objective.

Feat. types	Examples	
Align Density Indicators	INSIDEBOTH=3 LBLMATCH=*true*	INENONLY=0 BIAS=*true*

Table 1: Sample features used for named entity recognition for the ORG entity in Figure 1.

Let $\ell_1^W = \log p_1^W(y_1|x_1)$, $\ell_2^W = \log p_2^W(y_2|x_2)$ be the log-probability scores from the weakened models. Our final approximation to the marginal distribution over labels y is:

$$q_{\lambda_1,\lambda_2,\boldsymbol{\theta}}(y|x) \stackrel{def}{=} \max_a \exp\Big[\lambda_1\ell_1^W + \lambda_2\ell_2^W + \boldsymbol{\theta}^\top\boldsymbol{\phi}(y_1,a,y_2) - \tilde{A}(\lambda_1,\lambda_2,\boldsymbol{\theta};x)\Big] \; . \quad (1)$$

Where

$$\tilde{A}(\lambda_1,\lambda_2,\boldsymbol{\theta};x) = \log\sum_y \max_a \exp\Big[\lambda_1\ell_1^W + \lambda_2\ell_2^W + \boldsymbol{\theta}^\top\boldsymbol{\phi}(y_1,a,y_2)\Big]$$

is the updated approximate log partition function.

4 NER and Parsing Examples

Before formally describing our algorithm for finding the parameters $[\lambda_1, \lambda_2, \boldsymbol{\theta}]$, we first give examples of our problems of named entity recognition and syntactic parsing, together with node alignments and features for each. Figure 1 depicts a correctly-labeled sentence fragment in both English and German. In English, the capitalization of the phrase *European Court of Auditors* helps identify the span as a named entity. However, in German, all nouns are capitalized, and capitalization is therefore a less useful cue. While a monolingual German tagger is likely to miss the entity in the German text, by exploiting the parallel English text and word alignment information, we can hope to improve the German performance, and correctly tag *Europäischen Rechnungshofes*.

The monolingual features are standard features for discriminative, state-of-the-art entity recognizers, and we can produce weakened monolingual models by simply limiting the feature set. The bilingual features, $\boldsymbol{\phi}(y_1, a, y_2)$, are over pairs of aligned nodes, where nodes of the labels y_1 and y_2 are simply the individual named entities. We use a small bilingual feature set consisting of two types of features. First, we use the word alignment density features from Burkett and Klein (2008), which measure how well the aligned entity pair matches up with alignments from an independent

Input: full and weakened monolingual models: $p_1(y_1|x_1)$, $p_2(y_2|x_2)$, $p_1^w(y_1|x_1)$, $p_2^w(y_2|x_2)$
unannotated bilingual data: U
Output: bilingual parameters: $\hat{\boldsymbol{\theta}}, \hat{\lambda}_1, \hat{\lambda}_2$

1. Label U with full monolingual models: $\forall x \in U,\ \hat{y}_M = \text{argmax}_y\, p_1(y_1|x_1)p_2(y_2|x_2)$.

2. Return $\text{argmax}_{\lambda_1,\lambda_2,\boldsymbol{\theta}} \prod_{x \in U} q_{\boldsymbol{\theta},\lambda_1,\lambda_2}(\hat{y}_M|x)$, where $q_{\boldsymbol{\theta},\lambda_1,\lambda_2}$ has the form in Equation 1.

Figure 3: Bilingual training with multiple views.

Input: full and weakened monolingual models: $p_1(y_1|x_1)$, $p_2(y_2|x_2)$, $p_1^w(y_1|x_1)$, $p_2^w(y_2|x_2)$
bilingual parameters: $\hat{\boldsymbol{\theta}}, \hat{\lambda}_1, \hat{\lambda}_2$
bilingual input: $x = (x_1, x_2)$
Output: bilingual label: y^*

Bilingual w/ Weak	Bilingual w/ Full
1a. $l_1 = \log\left(p_1^w(y_1|x_1)\right)$	**1b.** $l_1 = \log\left(p_1(y_1|x_1)\right)$
2a. $l_2 = \log\left(p_2^w(y_2|x_2)\right)$	**2b.** $l_2 = \log\left(p_2(y_2|x_2)\right)$

3. Return $\text{argmax}_y \max_a \hat{\lambda}_1 l_1 + \hat{\lambda}_2 l_2 + \hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(y_1, a, y_2)$

Figure 4: Prediction by combining monolingual and bilingual models.

word aligner. We also include two indicator features: a bias feature that allows the model to learn a general preference for matched entities, and a feature that is active whenever the pair of nodes has the same label. Figure 1 contains sample values for each of these features.

Another natural setting where bilingual constraints can be exploited is syntactic parsing. Figure 2 shows an example English prepositional phrase attachment ambiguity that can be resolved bilingually by exploiting Chinese. The English monolingual parse mistakenly attaches *to* to the verb *increased*. In Chinese, however, this ambiguity does not exist. Instead, the word 对, which aligns to *to*, has strong selectional preference for attaching to a noun on the left.

In our parsing experiments, we use the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007), a split-merge latent variable parser, for our monolingual models. Our full model is the result of training the parser with five split-merge phases. Our weakened model uses only two. For the bilingual model, we use the same bilingual feature set as Burkett and Klein (2008). Table 2 gives some examples, but does not exhaustively enumerate those features.

5 Training Bilingual Models

Previous work in multiview learning has focused on the case of agreement regularization (Collins and Singer, 1999; Ganchev et al., 2008). If we had bilingual labeled data, together with our unlabeled data and monolingual labeled data, we could exploit these techniques. Because we do not possess bilingual labeled data, we must train the bilingual model in another way. Here we advocate training the bilingual model (consisting of the bilingual features and weakened monolingual models) to imitate the full monolingual models. In terms of agreement regularization, our procedure may be thought of as "regularizing" the bilingual model to be similar to the full monolingual models.

Our training algorithm is summarized in Figure 3. For each unlabeled point $x = (x_1, x_2)$, let $\hat{y}_M$ be the joint label which has the highest score from the independent monolingual models (line 1). We then find bilingual parameters $\hat{\boldsymbol{\theta}}, \hat{\lambda}_1, \hat{\lambda}_2$ that maximize $q_{\hat{\boldsymbol{\theta}},\hat{\lambda}_1,\hat{\lambda}_2}(\hat{y}_x|x)$ (line 2). This max-likelihood optimization can be solved by an EM-like procedure (Burkett and Klein, 2008). This procedure iteratively updates the parameter estimates by (a) finding the optimum alignments for each candidate label pair under the current parameters and then (b) updating the parameters to maximize a modified version of Equation 1, restricted to the optimal alignments. Because we restrict alignments to the set of at most one-to-one matchings, the (a) step is tractable using the Hungarian algorithm. With the alignments fixed, the (b) step just involves maximizing likelihood under a log-linear model with no latent variables – this problem is convex and can be solved efficiently using gradient-based methods. The procedure has no guarantees, but is observed in practice to converge to a local optimum.

6 Predicting with Monolingual and Bilingual Models

Once we have learned the parameters of the bilingual model, the standard method of bilingual prediction would be to just choose the y that is most likely under $q_{\hat{\boldsymbol{\theta}},\hat{\lambda}_1,\hat{\lambda}_2}$:

$$\hat{y} = \underset{y}{\text{argmax}}\, q_{\hat{\boldsymbol{\theta}},\hat{\lambda}_1,\hat{\lambda}_2}(y|x)\ . \qquad (2)$$

We refer to prediction under this model as "Bilingual w/ Weak," to evoke the fact that the model is making use of weakened monolingual models in its feature set.

Given that we have two views of the data, though, we should be able to leverage additional information in order to make better predictions. In

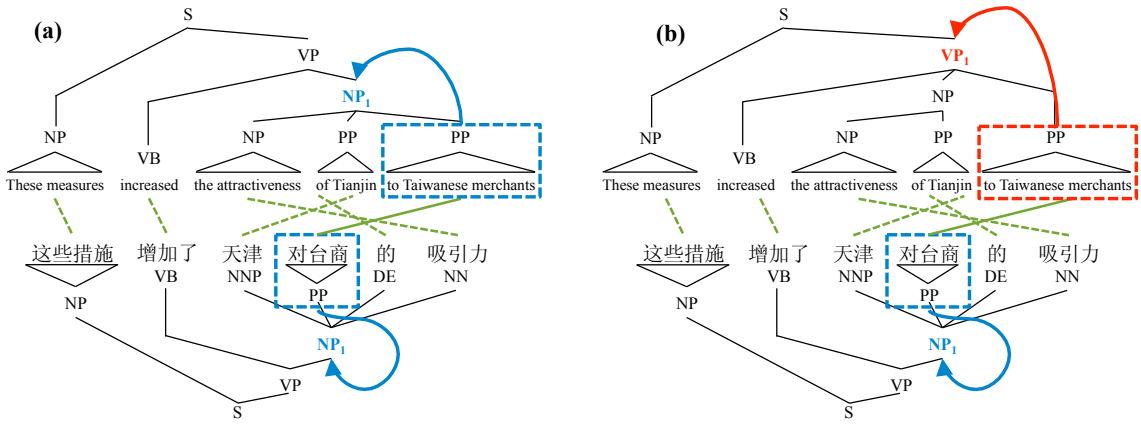

Figure 2: An example of PP attachment that is ambiguous in English, but simple in Chinese. In (a) the correct parses agree (low PP attachment), whereas in (b) the incorrect parses disagree.

Feature Types	Feature Templates	Examples	
		Correct	Incorrect
Alignment Density	INSIDEBOTH, INSIDEENONLY	INSIDEENONLY=0	INSIDEENONLY=1
Span Difference	ABSDIFFERENCE	ABSDIFFERENCE=3	ABSDIFFERENCE=4
Syntactic Indicators	LABEL⟨E,C⟩, NUMCHILDREN⟨E,C⟩	LABEL⟨NP,NP⟩=*true*	LABEL⟨VP,NP⟩=*true*

Table 2: Sample bilingual features used for parsing. The examples are features that would be extracted by aligning the parents of the PP nodes in Figure 2(a) (Correct) and Figure 2(b) (Incorrect).

particular, the monolingual view uses monolingual models that are known to be superior to the monolingual information available in the bilingual view. Thus, we would like to find some way to incorporate the full monolingual models into our prediction method. One obvious choice is to choose the labeling that maximizes the "agreement distribution" (Collins and Singer, 1999; Ganchev et al., 2008). In our setting, this amounts to choosing:

$$\hat{y} = \operatorname*{argmax}_{y} \; p_M(y|x) \; q_{\hat{\boldsymbol{\theta}},\hat{\lambda}_1\hat{\lambda}_2}(y|x) \; . \qquad (3)$$

This is the correct decision rule if the views are independent and the labels y are uniformly distributed a priori,[1] but we have deliberately introduced between-view dependence in the form of the weakened monolingual models. Equation 3 implicitly double-counts monolingual information.

One way to avoid this double-counting is to simply discard the weakened monolingual models when making a joint prediction:

$$\hat{y} = \operatorname*{argmax}_{y} \; \max_{a} \; p_M(y|x) \; \exp\left[\hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(y_1, a, y_2)\right] \; . \qquad (4)$$

[1]See, e.g. Ando & Zhang(Ando and Zhang, 2007) for a derivation of the decision rule from Equation 3 under these assumptions.

This decision rule uniformly combines the two monolingual models and the bilingual model. Note, however, that we have already learned non-uniform weights for the weakened monolingual models. Our final decision rule uses these weights as weights for the *full* monolingual models:

$$\hat{y} = \operatorname*{argmax}_{y} \max_{a} \exp\Big[\hat{\lambda}_1 \log\big(p_1(y_1|x_1)\big) + \hat{\lambda}_2 \log\big(p_2(y_2|x_2)\big) + \hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(y_1, a, y_2)\Big] \; . \qquad (5)$$

As we will show in Section 10, this rule for combining the monolingual and bilingual views performs significantly better than the alternatives, and comes close to the optimal weighting for the bilingual and monolingual models.

We will refer to predictions made with Equation 5 as "Bilingual w/ Full", to evoke the use of the full monolingual models alongside our bilingual features. Prediction using "Bilingual w/ Weak" and "Bilingual w/ Full" is summarized in Figure 4.

7 Retraining Monolingual Models

Although bilingual models have many direct applications (e.g. in machine translation), we also wish to be able to apply our models on purely monolingual data. In this case, we can still take

Input: annotated monolingual data: L_1, L_2
unannotated bilingual data: U
monolingual models: $p_1(y_1|x_1)$, $p_2(y_2|x_2)$
bilingual parameters: $\hat{\boldsymbol{\theta}}, \hat{\lambda}_1, \hat{\lambda}_2$

Output: retrained monolingual models:
$p_1^r(y_1|x_1)$, $p_2^r(y_2|x_2)$

$\forall x = (x_1, x_2) \in U$:

Self-Retrained	Bilingual-Retrained
1a. $\hat{y}_{x_1} = \text{argmax}_{y_1} p_1(y_1|x_1)$ $\hat{y}_{x_2} = \text{argmax}_{y_2} p_2(y_2|x_2)$	**1b.** Pick $\hat{y}_x$, Fig. 4 (Bilingual w/ Full)

2. Add $(x_1, \hat{y}_{x_1})$ to L_1 and add $(x_2, \hat{y}_{x_2})$ to L_2.

3. Return full monolingual models $p_1^r(y_1|x_1)$, $p_2^r(y_2|x_2)$ trained on newly enlarged L_1, L_2.

Figure 5: Retraining monolingual models.

advantage of parallel corpora by using our bilingual models to generate new training data for the monolingual models. This can be especially useful when we wish to use our monolingual models in a domain for which we lack annotated data, but for which bitexts are plentiful.[2]

Our retraining procedure is summarized in Figure 5. Once we have trained our bilingual parameters and have a "Bilingual w/ Full" predictor (using Equation 5), we can use that predictor to annotate a large corpus of parallel data (line 1b). We then retrain the full monolingual models on a concatenation of their original training data and the newly annotated data (line 3). We refer to the new *monolingual* models retrained on the output of the bilingual models as "Bilingual-Retrained," and we tested such models for both NER and parsing. For comparison, we also retrained monolingual models directly on the output of the original full monolingual models, using the same unannotated bilingual corpora for self-training (line 1a). We refer to these models as "Self-Retrained".

We evaluated our retrained monolingual models on the same test sets as our bilingual models, but using only monolingual data at test time. The texts used for retraining overlapped with the bitexts used for training the bilingual model, but both sets were disjoint from the test sets.

8 NER Experiments

We demonstrate the utility of multiview learning for named entity recognition (NER) on English/German sentence pairs. We built both our full and weakened monolingual English and German models from the CoNLL 2003 shared task training data. The bilingual model parameters were trained on 5,000 parallel sentences extracted from the Europarl corpus. For the retraining experiments, we added an additional 5,000 sentences, for 10,000 in all. For testing, we used the Europarl 2006 development set and the 2007 newswire test set. Neither of these data sets were annotated with named entities, so we manually annotated 200 sentences from each of them.

We used the Stanford NER tagger (Finkel et al., 2005) with its default configuration as our full monolingual model for each language. We weakened both the English and German models by removing several non-lexical and word-shape features. We made one more crucial change to our monolingual German model. The German entity recognizer has extremely low recall (44 %) when out of domain, so we chose $\hat{y}_x$ from Figure 3 to be the label in the top five which had the largest number of named entities.

Table 3 gives results for named entity recognition. The first two rows are the full and weakened monolingual models alone. The second two are the multiview trained bilingual models. We first note that for English, using the full bilingual model yields only slight improvements over the baseline full monolingual model, and in practice the predictions were almost identical. For this problem, the monolingual German model is much worse than the monolingual English model, and so the bilingual model doesn't offer significant improvements in English. The bilingual model does show significant German improvements, however, including a 16.1% absolute gain in F_1 over the baseline for parliamentary proceedings.

The last two rows of Table 3 give results for *monolingual* models which are trained on data that was automatically labeled using the our models. English results were again mixed, due to the relatively weak English performance of the bilingual model. For German, though, the "Bilingual-Retrained" model improves 14.4% F_1 over the "Self-Retrained" baseline.

9 Parsing Experiments

Our next set of experiments are on syntactic parsing of English and Chinese. We trained both our full and weakened monolingual English models on the Penn Wall Street Journal corpus (Marcus et al., 1993), as described in Section 4. Our full and weakened Chinese models were trained on

[2]Of course, unannotated *monolingual* data is even more plentiful, but as we will show, with the same amount of data, our method is more effective than simple monolingual self-training.

	Eng Parliament			Eng Newswire			Ger Parliament			Ger Newswire		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Monolingual Models (Baseline)												
Weak Monolingual	52.6	65.9	58.5	67.7	83.0	74.6	71.3	36.4	48.2	80.0	51.5	62.7
Full Monolingual	65.7	71.4	68.4	80.1	88.7	84.2	69.8	44.0	54.0	73.0	56.4	63.7
Multiview Trained Bilingual Models												
Bilingual w/ Weak	56.2	70.8	62.7	71.4	86.2	78.1	70.1	66.3	68.2	76.5	76.1	**76.3**
Bilingual w/ Full	65.4	**72.4**	**68.7**	**80.6**	**88.7**	**84.4**	70.1	**70.1**	**70.1**	74.6	**77.3**	75.9
Retrained Monolingual Models												
Self-Retrained	**71.7**	**74.0**	**72.9**	79.9	87.4	83.5	70.4	44.0	54.2	79.3	58.9	67.6
Bilingual-Retrained	68.6	70.8	69.7	**80.7**	**89.3**	**84.8**	**74.5**	**63.6**	**68.6**	77.9	**69.3**	**73.4**

Table 3: NER Results. Rows are grouped by data condition. We bold all entries that are best in their group and beat the strongest monolingual baseline.

	Chinese	English
Monolingual Models (Baseline)		
Weak Monolingual	78.3	67.6
Full Monolingual	84.2	75.4
Multiview Trained Bilingual Models		
Bilingual w/ Weak	80.4	70.8
Bilingual w/ Full	**85.9**	**77.5**
Supervised Trained Bilingual Models		
Burkett and Klein (2008)	86.1	78.2
Retrained Monolingual Models		
Self-Retrained	83.6	76.7
Bilingual-Retrained	83.9	**77.4**

Table 4: Parsing results. Rows are grouped by data condition. We bold entries that are best in their group and beat the the Full Monolingual baseline.

the Penn Chinese treebank (Xue et al., 2002) (articles 400-1151), excluding the bilingual portion. The bilingual data consists of the parallel part of the Chinese treebank (articles 1-270), which also includes manually parsed English translations of each Chinese sentence (Bies et al., 2007). Only the Chinese sentences and their English translations were used to train the bilingual models – the gold trees were ignored. For retraining, we used the same data, but weighted it to match the sizes of the original monolingual treebanks. We tested on the standard Chinese treebank development set, which also includes English translations.

Table 4 gives results for syntactic parsing. For comparison, we also show results for the *supervised* bilingual model of Burkett and Klein (2008). This model uses the same features at prediction time as the multiview trained "Bilingual w/ Full" model, but it is trained on hand-annotated parses. We first examine the first four rows of Table 4. The "Bilingual w/ Full" model significantly improves performance in both English and Chinese relative to the monolingual baseline. Indeed, it performs only slightly worse than the supervised model.

Phrase-Based System	
Moses (No Parser)	18.8
Syntactic Systems	
Monolingual Parser	18.7
Supervised Bilingual (Treebank Bi-trees)	21.1
Multiview Bilingual (Treebank Bitext)	20.9
Multiview Bilingual (Domain Bitext)	**21.2**

Table 5: Machine translation results.

The last two rows of Table 4 are the results of *monolingual* parsers trained on automatically labeled data. In general, gains in English, which is out of domain relative to the Penn Treebank, are larger than those in Chinese, which is in domain. We also emphasize that, unlike our NER data, this bitext was fairly small relative to the annotated monolingual data. Therefore, while we still learn good bilingual model parameters which give a sizable agreement-based boost when doing bilingual prediction, we don't expect retraining to result in a coverage-based boost in monolingual performance.

9.1 Machine Translation Experiments

Although we don't have hand-labeled data for our largest Chinese-English parallel corpora, we can still evaluate our parsing results via our performance on a downstream machine translation (MT) task. Our experimental setup is as follows: first, we used the first 100,000 sentences of the English-Chinese bitext from Wang et al. (2007) to train Moses (Koehn et al., 2007), a phrase-based MT system that we use as a baseline. We then used the same sentences to extract tree-to-string transducer rules from target-side (English) trees (Galley et al., 2004). We compare the single-reference BLEU scores of syntactic MT systems that result from using different parsers to generate these trees.

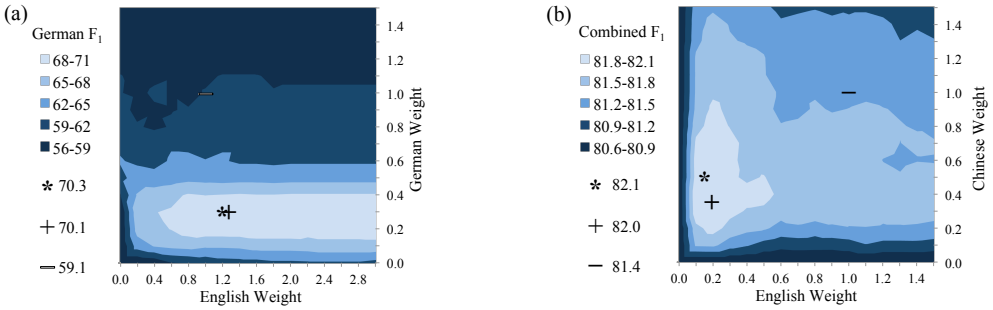

Figure 6: (a) NER and (b) parsing results for different values of λ_1 and λ_2 (see Equation 6). '*' shows optimal weights, '+' shows our learned weights, and '-' shows uniform combination weights.

For our syntactic baseline, we used the monolingual English parser. For our remaining experiments, we parsed both English and Chinese simultaneously. The supervised model and the first multiview trained model are the same Chinese treebank trained models for which we reported parsing results. We also used our multiview method to train an additional bilingual model on part of the bitext we used to extract translation rules.

The results are shown in Table 5. Once again, our multiview trained model yields comparable results to the supervised model. Furthermore, while the differences are small, our best performance comes from the model trained on in-domain data, for which no gold trees exist.

10 Analyzing Combined Prediction

In this section, we explore combinations of the full monolingual models, $p_1(y_1|x_1)$ and $p_2(y_2|x_2)$, and the bilingual model, $\max_a \hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(y_1, a, y_2)$. For parsing, the results in this section are for *combined* F_1. This simply computes F_1 over all of the sentences in both the English and Chinese test sets. For NER, we just use German F_1, since English is relatively constant across runs.

We begin by examining how poorly our model performs if we do not consider monolingual information in the bilingual view. For parsing, the combined Chinese and English F_1 for this model is 78.7%. When we combine this model uniformly with the full monolingual model, as in Equation 4, combined F_1 improves to 81.2%, but is still well below our best combined score of 82.1%. NER results for a model trained without monolingual information show an even larger decline.

Now let us consider decision rules of the form:

$$\hat{y} = \operatorname*{argmax}_y \max_a \exp[\lambda_1 \log\left(p_1(y_1|x_1)\right) + \lambda_2 \log\left(p_2(y_2|x_2)\right) + \hat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(y_1, a, y_2)] \,.$$

Note that when $\lambda_1 = \lambda_2 = 1$, this is exactly the uniform decision rule (Equation 4). When $\lambda_1 = \hat{\lambda}_1$ and $\lambda_2 = \hat{\lambda}_2$, this is the "Bilingual w/ Full" decision rule (Equation 5). Figure 6 is a contour plot of F_1 with respect to the parameters λ_1 and λ_2. Our decision rule "Bilingual w/ Full" (Equation 5, marked with a '+') is near the optimum ('*'), while the uniform decision rule ('-') performs quite poorly. This is true for both NER (Figure 6a) and parsing (Figure 6b).

There is one more decision rule which we have yet to consider: the "conditional independence" decision rule from Equation 3. While this rule cannot be shown on the plots in Figure 6 (because it uses both the full and weakened monolingual models), we note that it also performs poorly in both cases (80.7% F_1 for parsing, for example).

11 Conclusions

We show for the first time that state-of-the-art, discriminative monolingual models can be significantly improved using unannotated bilingual text. We do this by first building bilingual models that are trained to agree with pairs of independently-trained monolingual models. Then we combine the bilingual and monolingual models to account for dependence across views. By automatically annotating unlabeled bitexts with these bilingual models, we can train new *monolingual* models that do not rely on bilingual data at test time, but still perform substantially better than models trained using only monolingual resources.

Acknowledgements

This project is funded in part by NSF grants 0915265 and 0643742, an NSF graduate research fellowship, the DNI under grant HM1582-09-1-0021, and BBN under DARPA contract HR0011-06-C-0022.

References

Rie Kubota Ando and Tong Zhang. 2007. Two-view feature generation model for semi-supervised learning. In *ICML*.

Maria-Florina Balcan and Avrim Blum. 2005. A pac-style model for learning from labeled and unlabeled data. In *COLT*.

Ann Bies, Martha Palmer, Justin Mott, and Colin Warner. 2007. English chinese translation treebank v 1.0. Web download. LDC2007T02.

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

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2009. Bayesian synchronous grammar induction. In *NIPS*.

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

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL*.

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

Michael Collins. 2000. Discriminative reranking for natural language parsing. In *ICML*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL*.

Kuzman Ganchev, Joao Graca, John Blitzer, and Ben Taskar. 2008. Multi-view learning over structured and non-identical outputs. In *UAI*.

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

Fei Huang and Stephan Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *ICMI*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Special Issue of the Journal of Natural Language Engineering on Parallel Texts*, 11(3):311–325.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Robert Moore. 2003. Learning translations of named-entity phrases from parallel corpora. In *EACL*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL*.

Luo Si and Jamie Callan. 2005. Clef 2005: Multilingual retrieval by combining multiple multilingual ranked lists. In *CLEF*.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: using english to parse korean. In *EMNLP*.

Benjamin Snyder and Regina Barzilay. 2008. Cross-lingual propagation for morphological analysis. In *AAAI*.

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

Wen Wang, Andreas Stolcke, and Jing Zheng. 2007. Reranking machine translation hypotheses with structured and web-based language models. In *IEEE ASRU Workshop*.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *COLING*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Human Language Technologies*.