Unsupervised Multilingual Grammar Induction

Benjamin Snyder, Tahira Naseem, and **Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{bsnyder, tahira, regina}@csail.mit.edu

Abstract

We investigate the task of unsupervised constituency parsing from bilingual parallel corpora. Our goal is to use bilingual cues to learn improved parsing models for each language and to evaluate these models on held-out monolingual test data. We formulate a generative Bayesian model which seeks to explain the observed parallel data through a combination of bilingual and monolingual parameters. To this end, we adapt a formalism known as *unordered tree alignment* to our probabilistic setting. Using this formalism, our model loosely binds parallel trees while allowing language-specific syntactic structure. We perform inference under this model using Markov Chain Monte Carlo and dynamic programming. Applying this model to three parallel corpora (Korean-English, Urdu-English, and Chinese-English) we find substantial performance gains over the CCM model, a strong monolingual baseline. On average, across a variety of testing scenarios, our model achieves an 8.8 absolute gain in F-measure. [1]

1 Introduction

In this paper we investigate the task of unsupervised constituency parsing when bilingual parallel text is available. Our goal is to improve parsing performance on monolingual test data for each language by using unsupervised bilingual cues at training time. Multilingual learning has been successful for other linguistic induction tasks such as lexicon acquisition, morphological segmentation, and part-of-speech tagging (Genzel, 2005; Snyder and Barzilay, 2008; Snyder et al., 2008; Snyder et al., 2009). We focus here on the unsupervised induction of unlabeled constituency brackets. This task has been extensively studied in a monolingual setting and has proven to be difficult (Charniak and Carroll, 1992; Klein and Manning, 2002).

The key premise of our approach is that ambiguous syntactic structures in one language may correspond to less uncertain structures in the other language. For instance, the English sentence *I saw* [*the student* [*from MIT*]] exhibits the classic problem of PP-attachment ambiguity. However, its Urdu translation, literally glossed as *I* [[*MIT of*] *student*] *saw*, uses a genitive phrase that may only be attached to the adjacent noun phrase. Knowing the correspondence between these sentences should help us resolve the English ambiguity.

One of the main challenges of unsupervised multilingual learning is to exploit cross-lingual patterns discovered in data, while still allowing a wide range of language-specific idiosyncrasies. To this end, we adapt a formalism known as *unordered tree alignment* (Jiang et al., 1995) to a probabilistic setting. Under this formalism, any two trees can be embedded in an *alignment tree*. This alignment tree allows arbitrary parts of the two trees to diverge in structure, permitting language-specific grammatical structure to be preserved. Additionally, a computational advantage of this formalism is that the marginalized probability over all possible alignments for any two trees can be efficiently computed with a dynamic program in linear time.

We formulate a generative Bayesian model which seeks to explain the observed parallel data through a combination of bilingual and monolingual parameters. Our model views each pair of sentences as having been generated as follows: First an alignment tree is drawn. Each node in this alignment tree contains either a solitary monolingual constituent or a pair of coupled bilingual constituents. For each solitary mono-

[1] Code and the outputs of our experiments are available at http://groups.csail.mit.edu/rbg/code/multiling_induction.

lingual constituent, a sequence of part-of-speech tags is drawn from a language-specific distribution. For each pair of coupled bilingual constituents, a pair of part-of-speech sequences are drawn jointly from a cross-lingual distribution. Word-level alignments are then drawn based on the tree alignment. Finally, parallel sentences are assembled from these generated part-of-speech sequences and word-level alignments.

To perform inference under this model, we use a Metropolis-Hastings within-Gibbs sampler. We sample pairs of trees and then compute marginalized probabilities over all possible alignments using dynamic programming.

We test the effectiveness of our bilingual grammar induction model on three corpora of parallel text: English-Korean, English-Urdu and English-Chinese. The model is trained using bilingual data with automatically induced word-level alignments, but is tested on purely monolingual data for each language. In all cases, our model outperforms a state-of-the-art baseline: the Constituent Context Model (CCM) (Klein and Manning, 2002), sometimes by substantial margins. On average, over all the testing scenarios that we studied, our model achieves an absolute increase in F-measure of 8.8 points, and a 19% reduction in error relative to a theoretical upper bound.

2 Related Work

The unsupervised grammar induction task has been studied extensively, mostly in a monolingual setting (Charniak and Carroll, 1992; Stolcke and Omohundro, 1994; Klein and Manning, 2002; Seginer, 2007). While PCFGs perform poorly on this task, the CCM model (Klein and Manning, 2002) has achieved large gains in performance and is among the state-of-the-art probabilistic models for unsupervised constituency parsing. We therefore use CCM as our basic model of monolingual syntax.

While there has been some previous work on bilingual CFG parsing, it has mainly focused on improving MT systems rather than monolingual parsing accuracy. Research in this direction was pioneered by (Wu, 1997), who developed Inversion Transduction Grammars to capture cross-lingual grammar variations such as phrase reorderings. More general formalisms for the same purpose were later developed (Wu and Wong, 1998; Chiang, 2005; Melamed, 2003; Eisner, 2003; Zhang and Gildea, 2005; Blunsom et al., 2008). We know of only one study which evaluates these bilingual grammar formalisms on the task of grammar induction itself (Smith and Smith, 2004). Both our model and even the monolingual CCM baseline yield far higher performance on the same Korean-English corpus.

Our approach is closer to the unsupervised bilingual parsing model developed by Kuhn (2004), which aims to improve monolingual performance. Assuming that trees induced over parallel sentences have to exhibit certain structural regularities, Kuhn manually specifies a set of rules for determining when parsing decisions in the two languages are inconsistent with GIZA++ word-level alignments. By incorporating these constraints into the EM algorithm he was able to improve performance over a monolingual unsupervised PCFG. Still, the performance falls short of state-of-the-art monolingual models such as the CCM.

More recently, there has been a body of work attempting to improve parsing performance by exploiting syntactically annotated parallel data. In one strand of this work, annotations are assumed only in a resource-rich language and are projected onto a resource-poor language using the parallel data (Hwa et al., 2005; Xi and Hwa, 2005). In another strand of work, syntactic annotations are assumed on both sides of the parallel data, and a model is trained to exploit the parallel data at test time as well (Smith and Smith, 2004; Burkett and Klein, 2008). In contrast to this work, our goal is to explore the benefits of multilingual grammar induction in a fully unsupervised setting.

We finally note a recent paper which uses parameter tying to improve unsupervised dependency parse induction (Cohen and Smith, 2009). While the primary performance gains occur when tying related parameters within a language, some additional benefit is observed through bilingual tying, even in the absence of a parallel corpus.

3 Model

We propose an unsupervised Bayesian model for learning bilingual syntactic structure using parallel corpora. Our key premise is that difficult-to-learn syntactic structures of one language may correspond to simpler or less uncertain structures in the other language. We treat the part-of-speech tag sequences of parallel sentences, as well as their

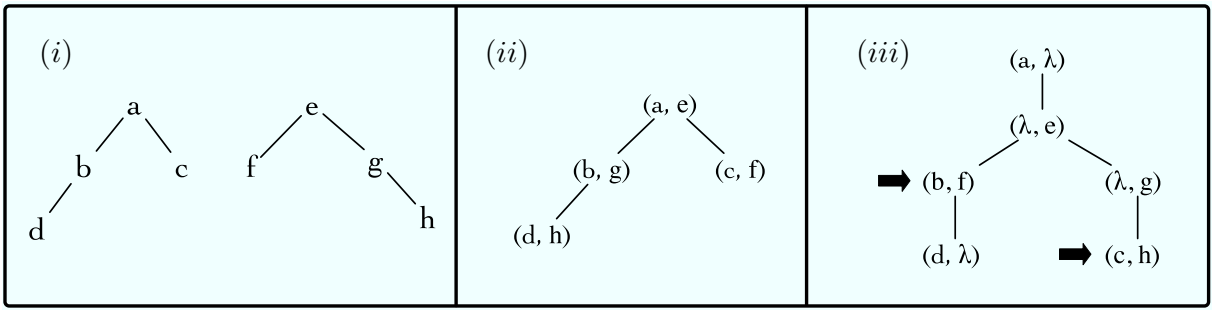

Figure 1: A pair of trees (i) and two possible alignment trees. In (ii), no empty spaces are inserted, but the order of one of the original tree's siblings has been reversed. In (iii), only two pairs of nodes have been aligned (indicated by arrows) and many empty spaces inserted.

word-level alignments, as observed data. We obtain these word-level alignments using GIZA++ (Och and Ney, 2003).

Our model seeks to explain this observed data through a generative process whereby two aligned parse trees are produced jointly. Though they are aligned, arbitrary parts of the two trees are permitted to diverge, accommodating language-specific grammatical structure. In effect, our model loosely binds the two trees: node-to-node alignments need only be used where repeated bilingual patterns can be discovered in the data.

3.1 Tree Alignments

We achieve this loose binding of trees by adapting *unordered tree alignment* (Jiang et al., 1995) to a probabilistic setting. Under this formalism, any two trees can be aligned using an *alignment tree*. The alignment tree embeds the original two trees within it: each node is labeled by a pair (x, y), (λ, y), or (x, λ) where x is a node from the first tree, y is a node from the second tree, and λ is an empty space. The individual structure of each tree must be preserved under the embedding with the exception of sibling order (to allow variations in phrase and word order).

The flexibility of this formalism can be demonstrated by two extreme cases: (1) an alignment between two trees may actually align *none* of their individual nodes, instead inserting an empty space λ for each of the original two trees' nodes. (2) if the original trees are isomorphic to one another, the alignment may match their nodes exactly, without inserting any empty spaces. See Figure 1 for an example.

3.2 Model overview

As our basic model of syntactic structure, we adopt the Constituent-Context Model (CCM) of Klein and Manning (2002). Under this model, the part-of-speech sequence of each span in a sentence is generated either as a *constituent yield* — if it is dominated by a node in the tree — or otherwise as a *distituent yield*. For example, in the bracketed sentence [*John*/NNP [*climbed*/VB [*the*/DT *tree*/NN]]], the sequence VB DT NN is generated as a constituent yield, since it constitutes a complete bracket in the tree. On the other hand, the sequence VB DT is generated as a distituent, since it does not. Besides these yields, the *contexts* (two surrounding POS tags) of constituents and distituents are generated as well. In this example, the context of the constituent VB DT NN would be (NNP, #), while the context of the distituent VB DT would be (NNP, NN). The CCM model employs separate multinomial distributions over constituents, distituents, constituent contexts, and distituent contexts. While this model is deficient — each observed subsequence of part-of-speech tags is generated many times over — its performance is far higher than that of unsupervised PCFGs.

Under our bilingual model, each pair of sentences is assumed to have been generated jointly in the following way: First, an unlabeled *alignment tree* is drawn uniformly from the set of all such trees. This alignment tree specifies the structure of each of the two individual trees, as well as the pairs of nodes which are aligned and those which are not aligned (i.e. paired with a λ).

For each pair of aligned nodes, a corresponding pair of constituents and contexts are jointly drawn from a bilingual distribution. For unaligned nodes (i.e. nodes paired with a λ in the alignment

tree), a single constituent and context are drawn, from language-specific distributions. Distituents and their contexts are also drawn from language-specific distributions. Finally, word-level alignments are drawn based on the structure of the alignment tree.

In the next two sections, we describe our model in more formal detail by specifying the parameters and generative process by which sentences are formed.

3.3 Parameters

Our model employs a number of multinomial distributions:

- π_i^C : over constituent yields of language i,
- π_i^D : over distituent yields of language i,
- ϕ_i^C : over constituent contexts of language i,
- ϕ_i^D : over distituent contexts of language i,
- ω : over *pairs* of constituent yields, one from the first language and the other from the second language,
- Gz_{pair} : over a finite set of integer values $\{-m, \ldots, -2, -1, 0, 1, 2, \ldots, m\}$, measuring the *Giza-score* of aligned tree node pairs (see below),
- Gz_{node} : over a finite set of integer values $\{-m, \ldots, -2, -1, 0\}$, measuring the *Giza-score* of unaligned tree nodes (see below).

The first four distributions correspond exactly to the parameters of the CCM model. Parameter ω is a "coupling parameter" which measures the compatibility of tree-aligned constituent yield pairs. The final two parameters measure the compatibility of syntactic alignments with the observed lexical GIZA++ alignments. Intuitively, aligned nodes should have a high density of word-level alignments between them, and unaligned nodes should have few lexical alignments.

More formally, consider a tree-aligned node pair (n_1, n_2) with corresponding yields (y_1, y_2). We call a word-level alignment *good* if it aligns a word in y_1 with a word in y_2. We call a word-level alignment *bad* if it aligns a word in y_1 with a word outside y_2, or vice versa. The *Giza-score* for (n_1, n_2) is the number of *good* word alignments minus the number of *bad* word alignments. For example, suppose the constituent *my long name* is node-aligned to its Urdu translation *mera lamba naam*. If only the word-pairs *my/mera* and *name/naam* are aligned, then the Giza-score for this node-alignment would be 2. If however, the English word *long* were (incorrectly) aligned under GIZA++ to some Urdu word outside the corresponding constituent, then the score would drop to 1. This score could even be negative if the number of *bad* alignments exceeds those that are *good*. Distribution Gz_{pair} provides a probability for these scores (up to some fixed absolute value).

For an unaligned node n with corresponding yield y, only *bad* GIZA++ alignments are possible, thus the Giza-score for these nodes will always be zero or negative. Distribution Gz_{node} provides a probability for these scores (down to some fixed value). We want our model to find tree alignments such that both aligned node pairs and unaligned nodes have high *Giza-score*.

3.4 Generative Process

Now we describe the stochastic process whereby the observed parallel sentences and their word-level alignments are generated, according to our model.

As the first step in the Bayesian generative process, all the multinomial parameters listed in the previous section are drawn from their conjugate priors — Dirichlet distributions of appropriate dimension. Then, each pair of word-aligned parallel sentences is generated through the following process:

1. A pair of binary trees T_1 and T_2 along with an alignment tree A are drawn according to $P(T_1, T_2, A)$. A is an alignment tree for T_1 and T_2 if it can be obtained by the following steps: First insert blank nodes (labeled by λ) into T_1 and T_2. Then permute the order of sibling nodes such that the two resulting trees T_1' and T_2' are identical in structure. Finally, overlay T_1' and T_2' to obtain A. We additionally require that A contain no extraneous nodes – that is no nodes with two blank labels (λ, λ). See Figure 1 for an example.

   We define the distribution $P(T_1, T_2, A)$ to be uniform over all pairs of binary trees and their alignments.

2. For each node in A of the form (n_1, λ) (i.e. nodes in T_1 left unaligned by A), draw

   (i) a constituent yield according to π_1^C,

(ii) a constituent context according to ϕ_1^C,

(iii) a Giza-score according to Gz_{node}.

3. For each node in A of the form (λ, n_2) (i.e. nodes in T_2 left unaligned by A), draw

   (i) a constituent yield according to π_2^C,

   (ii) a constituent context according to ϕ_2^C,

   (iii) a Giza-score according to Gz_{node}.

4. For each node in A of the form (n_1, n_2) (i.e. tree-aligned node pairs), draw

   (i) a pair of constituent yields (y_1, y_2) according to:

   $$\frac{\phi_1^C(y_1) \cdot \phi_2^C(y_2) \cdot \omega(y_1, y_2)}{Z} \quad (1)$$

   which is a product of experts combining the language specific context-yield distributions as well as the coupling distribution ω with normalization constant Z,

   (ii) a pair of contexts according to the appropriate language-specific parameters,

   (iii) a Giza-score according to Gz_{pair}.

5. For each span in T_i *not* dominated by a node (for each language $i \in \{1, 2\}$), draw a distituent yield according to π_i^D and a distituent context according to ϕ_i^D.

6. Draw actual word-level alignments consistent with the Giza-scores, according to a uniform distribution.

In the next section we turn to the problem of inference under this model when only the part-of-speech tag sequences of parallel sentences and their word-level alignments are observed.

3.5 Inference

Given a corpus of paired part-of-speech tag sequences $(\mathbf{s_1}, \mathbf{s_2})$ and their GIZA++ alignments $\mathbf{g}$, we would ideally like to predict the set of tree pairs $(\mathbf{T_1}, \mathbf{T_2})$ which have highest probability when conditioned on the observed data: $P(\mathbf{T_1}, \mathbf{T_2} | \mathbf{s_1}, \mathbf{s_2}, \mathbf{g})$. We could rewrite this by explicitly integrating over the yield, context, coupling, Giza-score parameters as well as the alignment trees. However, since maximizing this integral directly would be intractable, we resort to standard Markov chain sampling techniques. We use Gibbs sampling (Hastings, 1970) to draw trees for each sentence conditioned on those drawn for all other sentences. The samples form a Markov chain which is guaranteed to converge to the true joint distribution over all sentences.

In the monolingual setting, there is a well-known tree sampling algorithm (Johnson et al., 2007). This algorithm proceeds in top-down fashion by sampling individual split points using the marginal probabilities of all possible subtrees. These marginals can be efficiently pre-computed and form the "inside" table of the famous Inside-Outside algorithm. However, in our setting, trees come in pairs, and their joint probability crucially depends on their alignment.

For the i^{th} parallel sentence, we wish to jointly sample the pair of trees $(T_1, T_2)_i$ together with their alignment A_i. To do so directly would involve simultaneously marginalizing over all possible subtrees as well as all possible alignments between such subtrees when sampling upper-level split points. We know of no obvious algorithm for computing this marginal. We instead first sample the pair of trees $(T_1, T_2)_i$ from a simpler *proposal distribution* Q. Our proposal distribution assumes that *no* nodes of the two trees are aligned and therefore allows us to use the recursive top-down sampling algorithm mentioned above. After a new tree pair $T^* = (T_1^*, T_2^*)_i$ is drawn from Q, we accept the pair with the following probability:

$$\min\left\{1, \frac{P(T^*|\mathbf{T}_{-i}, \mathbf{A}_{-i})\ Q(T|\mathbf{T}_{-i}, \mathbf{A}_{-i})}{P(T|\mathbf{T}_{-i}, \mathbf{A}_{-i})\ Q(T^*|\mathbf{T}_{-i}, \mathbf{A}_{-i})}\right\}$$

where T is the previously sampled tree-pair for sentence i, P is the true model probability, and Q is the probability under the proposal distribution. This use of a tractable proposal distribution and acceptance ratio is known as the Metropolis-Hastings algorithm and it preserves the convergence guarantee of the Gibbs sampler (Hastings, 1970). To compute the terms $P(T^*|\mathbf{T}_{-i}, \mathbf{A}_{-i})$ and $P(T|\mathbf{T}_{-i}, \mathbf{A}_{-i})$ in the acceptance ratio above, we need to marginalize over all possible alignments between tree pairs.

Fortunately, for any given pair of trees T_1 and T_2 this marginalization can be computed using a dynamic program in time $O(|T_1||T_2|)$. Here we provide a very brief sketch. For every pair of nodes $n_1 \in T_1, n_2 \in T_2$, a table stores the marginal probability of the subtrees rooted at n_1 and n_2, respectively. A dynamic program builds this table from the bottom up: For each node pair n_1, n_2, we sum the probabilities of all local alignment configurations, each multiplied by the appro-

priate marginals already computed in the table for lower-level node pairs. This algorithm is an adaptation of the dynamic program presented in (Jiang et al., 1995) for finding minimum cost alignment trees (Fig. 5 of that publication).

Once a pair of trees (T_1, T_2) has been sampled, we can proceed to sample an alignment tree $A|T_1, T_2$.[2] We sample individual alignment decisions from the top down, at each step using the alignment marginals for the remaining subtrees (already computed using the afore-mentioned dynamic program). Once the triple (T_1, T_2, A) has been sampled, we move on to the next parallel sentence.

We avoid directly sampling parameter values, instead using the marginalized closed forms for multinomials with Dirichlet conjugate-priors using counts and hyperparameter pseudo-counts (Gelman et al., 2004). Note that in the case of yield pairs produced according to Distribution 1 (in step 4 of the generative process) conjugacy is technically broken, since the yield pairs are no longer produced by a single multinomial distribution. Nevertheless, we count the produced yields as if they had been generated separately by each of the distributions involved in the numerator of Distribution 1.

4 Experimental setup

We test our model on three corpora of bilingual parallel sentences: English-Korean, English-Urdu, and English-Chinese. Though the model is trained using parallel data, during testing it has access only to monolingual data. This set-up ensures that we are testing our model's ability to learn better parameters at training time, rather than its ability to exploit parallel data at test time. Following (Klein and Manning, 2002), we restrict our model to binary trees, though we note that the alignment trees do not follow this restriction.

Data The Penn Korean Treebank (Han et al., 2002) consists of 5,083 Korean sentences translated into English for the purposes of language training in a military setting. Both the Korean and English sentences are annotated with syntactic trees. We use the first 4,000 sentences for training and the last 1,083 sentences for testing. We note that in the Korean data, a separate tag is given for each morpheme. We simply concatenate all the morpheme tags given for each word and treat the concatenation as a single tag. This procedure results in 199 different tags. The English-Urdu parallel corpus[3] consists of 4,325 sentences from the first three sections of the Penn Treebank and their Urdu translations annotated at the part-of-speech level. The Urdu side of this corpus does not provide tree annotations so here we can test parse accuracy only on English. We use the remaining sections of the Penn Treebank for English testing. The English-Chinese treebank (Bies et al., 2007) consists of 3,850 Chinese newswire sentences translated into English. Both the English and Chinese sentences are annotated with parse trees. We use the first 4/5 for training and the final 1/5 for testing.

During preprocessing of the corpora we remove all punctuation marks and special symbols, following the setup in previous grammar induction work (Klein and Manning, 2002). To obtain lexical alignments between the parallel sentences we employ GIZA++ (Och and Ney, 2003). We use intersection alignments, which are one-to-one alignments produced by taking the intersection of one-to-many alignments in each direction. These one-to-one intersection alignments tend to have higher precision.

We initialize the trees by making uniform split decisions recursively from the top down for sentences in both languages. Then for each pair of parallel sentences we randomly sample an initial alignment tree for the two sampled trees.

Baseline We implement a Bayesian version of the CCM as a baseline. This model uses the same inference procedure as our bilingual model (Gibbs sampling). In fact, our model reduces to this Bayesian CCM when it is assumed that no nodes between the two parallel trees are ever aligned and when word-level alignments are ignored. We also reimplemented the original EM version of CCM and found virtually no difference in performance when using EM or Gibbs sampling. In both cases our implementation achieves F-measure in the range of 69-70% on WSJ10, broadly in line with the performance reported by Klein and Manning (2002).

Hyperparameters Klein (2005) reports using smoothing pseudo-counts of 2 for constituent

[2] Sampling the alignment tree is important, as it provides us with counts of aligned constituents for the coupling parameter.

[3] http://www.crulp.org

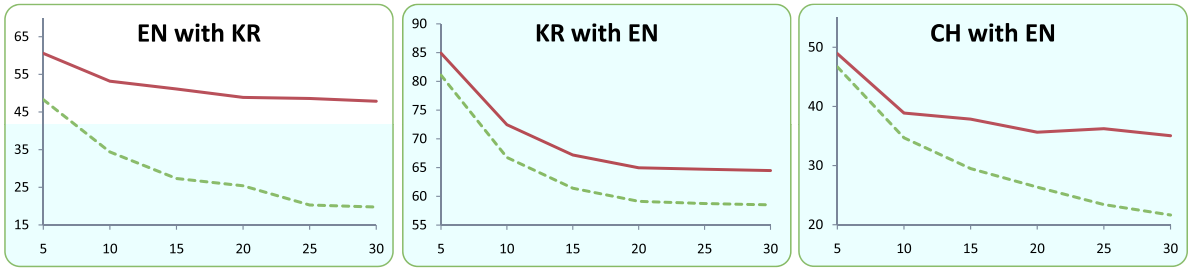

Figure 2: The F-measure of the CCM baseline (dotted line) and bilingual model (solid line) plotted on the y-axis, as the maximum sentence length in the test set is increased (x-axis). Results are averaged over all training scenarios given in Table 1.

yields and contexts and 8 for distituent yields and contexts. In our Bayesian model, these similar smoothing counts occur as the parameters of the Dirichlet priors. For Korean we found that the baseline performed well using these values. However, on our English and Chinese data, we found that somewhat higher smoothing values worked best, so we utilized values of 20 and 80 for constituent and distituent smoothing counts, respectively.

Our model additionally requires hyperparameter values for ω (the coupling distribution for aligned yields), Gz_{pair} and Gz_{node} (the distributions over Giza-scores for aligned nodes and unaligned nodes, respectively). For ω we used a symmetric Dirichlet prior with parameter 1. For Gz_{pair} and Gz_{node}, in order to create a strong bias towards high Giza-scores, we used non-symmetric Dirichlet priors. In both cases, we capped the absolute value of the scores at 3, to prevent count sparsity. In the case of Gz_{pair} we gave pseudo-counts of 1,000 for negative values and zero, and pseudo-counts of 1,000,000 for positive scores. For Gz_{node} we gave a pseudo-count of 1,000,000 for a score of zero, and 1,000 for all negative scores. This very strong prior bias encodes our intuition that syntactic alignments which respect lexical alignments should be preferred. Our method is not sensitive to these exact values and any reasonably strong bias gave similar results.

In all our experiments, we consider the hyperparameters fixed and observed values.

Testing and evaluation As mentioned above, we test our model only on monolingual data, where the parallel sentences are not provided to the model. To predict the bracketings of these monolingual test sentences, we take the smoothed counts accumulated in the final round of sampling over the training data and perform a maximum likelihood estimate of the monolingual CCM parameters. These parameters are then used to produce the highest probability bracketing of the test set.

To evaluate both our model as well as the baseline, we use (unlabeled) bracket precision, recall, and F-measure (Klein and Manning, 2002). Following previous work, we include the whole-sentence brackets but ignore single-word brackets. We perform experiments on different subsets of training and testing data based on the sentence-length. In particular we experimented with sentence length limits of 10, 20, and 30 for both the training and testing sets. We also report the upper bound on F-measure for binary trees. We average the results over 10 separate sampling runs.

5 Results

Table 1 reports the full results of our experiments. In all testing scenarios the bilingual model outperforms its monolingual counterpart in terms of both precision and recall. On average, the bilingual model gains 10.2 percentage points in precision, 7.7 in recall, and 8.8 in F-measure. The gap between monolingual performance and the binary tree upper bound is reduced by over 19%.

The extent of the gain varies across pairings. For instance, the smallest improvement is observed for English when trained with Urdu. The Korean-English pairing results in substantial improvements for Korean and quite large improvements for English, for which the absolute gain reaches 28 points in F-measure. In the case of Chinese and English, the gains for English are fairly minimal whereas those for Chinese are quite sub-

	Max Sent. Length		Monolingual			Bilingual			Upper Bound
	Test	Train	Precision	Recall	F1	Precision	Recall	F1	F1
EN with KR	10	10	52.74	39.53	45.19	57.76	43.30	49.50	85.6
		20	41.87	31.38	35.87	61.66	46.22	52.83	85.6
		30	33.43	25.06	28.65	64.41	48.28	**55.19**	85.6
	20	20	35.12	25.12	29.29	56.96	40.74	47.50	83.3
		30	26.26	18.78	21.90	60.07	42.96	**50.09**	83.3
	30	30	23.95	16.81	19.76	58.01	40.73	**47.86**	82.4
KR with EN	10	10	71.07	62.55	66.54	75.63	66.56	70.81	93.6
		20	71.35	62.79	66.80	77.61	68.30	72.66	93.6
		30	71.37	62.81	66.82	77.87	68.53	**72.91**	93.6
	20	20	64.28	54.73	59.12	70.44	59.98	64.79	91.9
		30	64.29	54.75	59.14	70.81	60.30	**65.13**	91.9
	30	30	63.63	54.17	58.52	70.11	59.70	**64.49**	91.9
EN with CH	10	10	50.09	34.18	40.63	37.46	25.56	30.39	81.0
		20	58.86	40.17	47.75	50.24	34.29	40.76	81.0
		30	64.81	44.22	52.57	68.24	46.57	**55.36**	81.0
	20	20	41.90	30.52	35.31	38.64	28.15	32.57	84.3
		30	52.83	38.49	44.53	58.50	42.62	**49.31**	84.3
	30	30	46.35	33.67	39.00	51.40	37.33	**43.25**	84.1
CH with EN	10	10	39.87	27.71	32.69	40.62	28.23	33.31	81.9
		20	43.44	30.19	35.62	47.54	33.03	38.98	81.9
		30	43.63	30.32	35.77	54.09	37.59	**44.36**	81.9
	20	20	29.80	23.46	26.25	36.93	29.07	32.53	88.0
		30	30.05	23.65	26.47	43.99	34.63	**38.75**	88.0
	30	30	24.46	19.41	21.64	39.61	31.43	**35.05**	88.4
EN with UR	10	10	57.98	45.68	51.10	73.43	57.85	64.71	88.1
		20	70.57	55.60	62.20	80.24	63.22	**70.72**	88.1
		30	75.39	59.40	66.45	79.04	62.28	69.67	88.1
	20	20	57.78	43.86	49.87	67.26	51.06	**58.05**	86.3
		30	63.12	47.91	54.47	64.45	48.92	55.62	86.3
	30	30	57.36	43.02	49.17	57.97	43.48	**49.69**	85.7

Table 1: Unlabeled precision, recall and F-measure for the monolingual baseline and the bilingual model on several test sets. We report results for different combinations of maximum sentence length in both the training and test sets. The right most column, in all cases, contains the maximum F-measure achievable using binary trees. The best performance for each test-length is highlighted in bold.

stantial. This asymmetry should not be surprising, as Chinese on its own seems to be quite a bit more difficult to parse than English.

We also investigated the impact of sentence length for both the training and testing sets. For our model, adding sentences of greater length to the training set leads to increases in parse accuracy for short sentences. For the baseline, however, adding this additional training data degrades performance in the case of English paired with Korean. Figure 2 summarizes the performance of our model for different sentence lengths on several of the test-sets. As shown in the figure, the largest improvements tend to occur at longer sentence lengths.

6 Conclusion

We have presented a probabilistic model for bilingual grammar induction which uses raw parallel text to learn tree pairs and their alignments. Our formalism loosely binds the two trees, using bilingual patterns when possible, but allowing substantial language-specific variation. We tested our model on three test sets and showed substantial improvement over a state-of-the-art monolingual baseline.[4]

[4]The authors acknowledge the support of the NSF (CAREER grant IIS-0448168, grant IIS-0835445, and grant IIS-0835652). Thanks to Amir Globerson and members of the MIT NLP group for their helpful suggestions. Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the funding organizations

References

Ann Bies, Martha Palmer, Justin Mott, and Colin Warner. 2007. *English Chinese translation treebank v 1.0*. LDC2007T02.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. Bayesian synchronous grammar induction. In *Proceedings of NIPS*.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of EMNLP*, pages 877–886.

Eugene Charniak and Glen Carroll. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, pages 1–13.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL*, pages 263–270.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of the NAACL/HLT*.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *The Companion Volume to the Proceedings of the ACL*, pages 205–208.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

Dmitriy Genzel. 2005. Inducing a multilingual dictionary from a parallel multitext in related languages. In *Proceedings of EMNLP/HLT*, pages 875–882.

C. Han, N.R. Han, E.S. Ko, H. Yi, and M. Palmer. 2002. Penn Korean Treebank: Development and evaluation. In *Proc. Pacific Asian Conf. Language and Comp*.

W. K. Hastings. 1970. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering*, 11(3):311–325.

T. Jiang, L. Wang, and K. Zhang. 1995. Alignment of trees – an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148.

M. Johnson, T. Griffiths, and S. Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of the NAACL/HLT*, pages 139–146.

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the ACL*, pages 128–135.

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

Jonas Kuhn. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of the ACL*, pages 470–477.

I. Dan Melamed. 2003. Multitext grammars and synchronous parsers. In *Proceedings of the NAACL/HLT*, pages 79–86.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the ACL*, pages 384–391.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceeding of EMNLP*, pages 49–56.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the ACL/HLT*, pages 737–745.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of EMNLP*, pages 1041–1050.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *Proceedings of the NAACL/HLT*.

Andreas Stolcke and Stephen M. Omohundro. 1994. Inducing probabilistic grammars by Bayesian model merging. In *Proceedings of ICGI*, pages 106–118.

Dekai Wu and Hongsing Wong. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of the ACL/COLING*, pages 1408–1415.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of EMNLP*, pages 851 – 858.

Hao Zhang and Daniel Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the ACL*, pages 475–482.