Joint Parsing and Alignment with Weakly Synchronized Grammars

David Burkett **John Blitzer** **Dan Klein**
Computer Science Division
University of California, Berkeley
{dburkett,blitzer,klein}@cs.berkeley.edu

Abstract

Syntactic machine translation systems extract rules from bilingual, word-aligned, syntactically parsed text, but current systems for parsing and word alignment are at best cascaded and at worst totally independent of one another. This work presents a unified joint model for simultaneous parsing and word alignment. To flexibly model syntactic divergence, we develop a discriminative log-linear model over two parse trees and an ITG derivation which is encouraged but not forced to synchronize with the parses. Our model gives absolute improvements of 3.3 F_1 for English parsing, 2.1 F_1 for Chinese parsing, and 5.5 F_1 for word alignment over each task's independent baseline, giving the best reported results for both Chinese-English word alignment and joint parsing on the parallel portion of the Chinese treebank. We also show an improvement of 1.2 BLEU in downstream MT evaluation over basic HMM alignments.

1 Introduction

Current syntactic machine translation (MT) systems build synchronous context free grammars from aligned syntactic fragments (Galley et al., 2004; Zollmann et al., 2006). Extracting such grammars requires that bilingual word alignments and monolingual syntactic parses be compatible. Because of this, much recent work in both word alignment and parsing has focused on changing aligners to make use of syntactic information (DeNero and Klein, 2007; May and Knight, 2007; Fossum et al., 2008) or changing parsers to make use of word alignments (Smith and Smith, 2004; Burkett and Klein, 2008; Snyder et al., 2009). In the first case, however, parsers do not exploit bilingual information. In the second, word alignment is performed with a model that does not exploit syntactic information. This work presents a single, joint model for parsing and word alignment that allows both pieces to influence one another simultaneously.

While building a joint model seems intuitive, there is no easy way to characterize how word alignments and syntactic parses should relate to each other in general. In the ideal situation, each pair of sentences in a bilingual corpus could be syntactically parsed using a synchronous context-free grammar. Of course, real translations are almost always at least partially syntactically divergent. Therefore, it is unreasonable to expect perfect matches of any kind between the two sides' syntactic trees, much less expect that those matches be well explained at a word level. Indeed, it is sometimes the case that large pieces of a sentence pair are completely asynchronous and can only be explained monolingually.

Our model exploits synchronization where possible to perform more accurately on both word alignment and parsing, but also allows independent models to dictate pieces of parse trees and word alignments when synchronization is impossible. This notion of "weak synchronization" is parameterized and estimated from data to maximize the likelihood of the correct parses and word alignments. Weak synchronization is closely related to the quasi-synchronous models of Smith and Eisner (2006; 2009) and the bilingual parse reranking model of Burkett and Klein (2008), but those models assume that the word alignment of a sentence pair is known and fixed.

To simultaneously model both parses and align-

ments, our model loosely couples three separate combinatorial structures: monolingual trees in the source and target languages, and a synchronous ITG alignment that links the two languages (but is not constrained to match linguistic syntax). The model has no hard constraints on how these three structures must align, but instead contains a set of "synchronization" features that are used to propagate influence between the three component grammars. The presence of synchronization features couples the parses and alignments, but makes exact inference in the model intractable; we show how to use a variational mean field approximation, both for computing approximate feature expectations during training, and for performing approximate joint inference at test time.

We train our joint model on the parallel, gold word-aligned portion of the Chinese treebank. When evaluated on parsing and word alignment, this model significantly improves over independently-trained baselines: the monolingual parser of Petrov and Klein (2007) and the discriminative word aligner of Haghighi et al. (2009). It also improves over the discriminative, bilingual parsing model of Burkett and Klein (2008), yielding the highest joint parsing F_1 numbers on this data set. Finally, our model improves word alignment in the context of translation, leading to a 1.2 BLEU increase over using HMM word alignments.

2 Joint Parsing and Alignment

Given a source-language sentence, s, and a target-language sentence, s', we wish to predict a source tree t, a target tree t', and some kind of alignment a between them. These structures are illustrated in Figure 1.

To facilitate these predictions, we define a conditional distribution $\mathrm{P}(t, a, t'|s, s')$. We begin with a generic conditional exponential form:

$$\mathrm{P}(t, a, t'|s, s') \propto \exp \theta^\top \phi(t, a, t', s, s') \quad (1)$$

Unfortunately, a generic model of this form is intractable, because we cannot efficiently sum over all triples (t, a, t') without some assumptions about how the features $\phi(t, a, t', s, s')$ decompose.

One natural solution is to restrict our candidate triples to those given by a synchronous context free grammar (SCFG) (Shieber and Schabes, 1990). Figure 1(a) gives a simple example of generation from a log-linearly parameterized synchronous grammar, together with its features. With the SCFG restriction, we can sum over the necessary structures using the $O(n^6)$ bitext inside-outside algorithm, making $\mathrm{P}(t, a, t'|s, s')$ relatively efficient to compute expectations under.

Unfortunately, an SCFG requires that *all* the constituents of each tree, from the root down to the words, are generated perfectly in tandem. The resulting inability to model any level of syntactic divergence prevents accurate modeling of the individual monolingual trees. We will consider the running example from Figure 2 throughout the paper. Here, for instance, the verb phrase *established in such places as Quanzhou, Zhangzhou, etc.* in English does not correspond to any single node in the Chinese tree. A synchronous grammar has no choice but to analyze this sentence incorrectly, either by ignoring this verb phrase in English or postulating an incorrect Chinese constituent that corresponds to it.

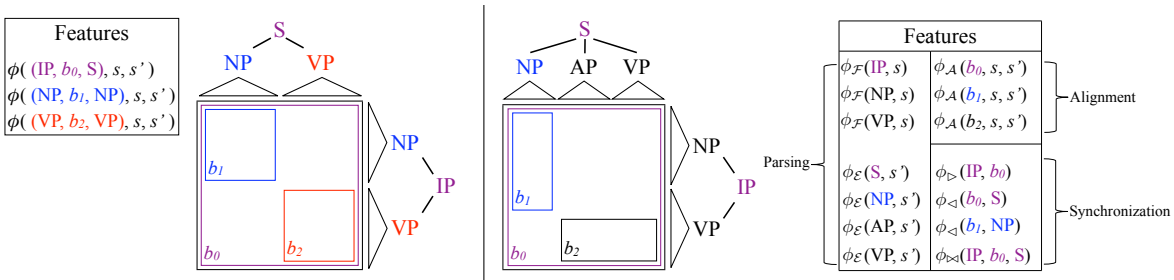
Therefore, instead of requiring strict synchronization, our model treats the two monolingual trees and the alignment as separate objects that can vary arbitrarily. However, the model rewards synchronization appropriately when the alignment brings the trees into correspondence.

3 Weakly Synchronized Grammars

We propose a joint model which still gives probabilities on triples (t, a, t'). However, instead of using SCFG rules to synchronously enforce the tree constraints on t and t', we only require that each of t and t' be well-formed under separate monolingual CFGs.

In order to permit efficient enumeration of all possible alignments a, we also restrict a to the set of unlabeled ITG bitrees (Wu, 1997), though again we do not require that a relate to t or t' in any particular way. Although this assumption does limit the space of possible word-level alignments, for the domain we consider (Chinese-English word alignment), the reduced space still contains almost all empirically observed alignments (Haghighi et al., 2009).[1] For

[1] See Section 8.1 for some new terminal productions required to make this true for the parallel Chinese treebank.



(a) Synchronous Rule

(b) Asynchronous Rule

Figure 1: Source trees, t (right), alignments, a (grid), and target trees, t' (top), and feature decompositions for synchronous (a) and weakly synchronous (b) grammars. Features always condition on bispans and/or anchored syntactic productions, but weakly synchronous grammars permit more general decompositions.

example, in Figure 2, the word alignment is ITG-derivable, and each of the colored rectangles is a bispan in that derivation.

There are no additional constraints beyond the independent, internal structural constraints on t, a, and t'. This decoupling permits derivations like that in Figure 1(b), where the top-level syntactic nodes align, but their children are allowed to diverge. With the three structures separated, our first model is a completely factored decomposition of (1).

Formally, we represent a source tree t as a set of nodes $\{n\}$, each node representing a labeled span. Likewise, a target tree t' is a set of nodes $\{n'\}$.[2] We represent alignments a as sets of *bispans* $\{b\}$, indicated by rectangles in Figure 1.[3] Using this notation, the initial model has the following form:

$$\mathrm{P}(t, a, t' | s, s') \propto \exp \left[\sum_{n \in t} \theta^\top \phi_{\mathcal{F}}(n, s) + \sum_{b \in a} \theta^\top \phi_{\mathcal{A}}(b, s, s') + \sum_{n' \in t'} \theta^\top \phi_{\mathcal{E}}(n', s') \right] \quad (2)$$

Here $\phi_{\mathcal{F}}(n, s)$ indicates a vector of source node features, $\phi_{\mathcal{E}}(n', s')$ is a vector of target node features, and $\phi_{\mathcal{A}}(b, s, s')$ is a vector of alignment bispan features. Of course, this model is completely asynchronous so far, and fails to couple the trees and alignments at all. To permit soft constraints between the three structures we are modeling, we add a set of *synchronization* features.

For $n \in t$ and $b \in a$, we say that $n \triangleright b$ if n and b both map onto the same span of s. We define $b \triangleleft n'$ analogously for $n' \in t'$. We now consider three different types of synchronization features. Source-alignment synchronization features $\phi_{\triangleright}(n, b)$ are extracted whenever $n \triangleright b$. Similarly, target-alignment features $\phi_{\triangleleft}(b, n')$ are extracted if $b \triangleleft n'$. These features capture phenomena like that of bispan b_7 in Figure 2. Here the Chinese noun 地 synchronizes with the ITG derivation, but the English projection of b_7 is a distituent. Finally, we extract source-target features $\phi_{\bowtie}(n, b, n')$ whenever $n \triangleright b \triangleleft n'$. These features capture complete bispan synchrony (as in bispan b_8) and can be expressed over triples (n, b, n') which happen to align, allowing us to reward synchrony, but not requiring it. All of these licensing conditions are illustrated in Figure 1(b).

With these features added, the final form of the model is:

$$\mathrm{P}(t, a, t' | s, s') \propto \exp \left[\sum_{n \in t} \theta^\top \phi_{\mathcal{F}}(n, s) + \sum_{b \in a} \theta^\top \phi_{\mathcal{A}}(b, s, s') + \sum_{n' \in t'} \theta^\top \phi_{\mathcal{E}}(n', s') + \sum_{n \triangleright b} \theta^\top \phi_{\triangleright}(n, b) + \sum_{b \triangleleft n'} \theta^\top \phi_{\triangleleft}(b, n') + \sum_{n \triangleright b \triangleleft n'} \theta^\top \phi_{\bowtie}(n, b, n') \right] \quad (3)$$

[2] For expositional clarity, we describe n and n' as labeled spans only. However, in general, features that depend on n or n' are permitted to depend on the entire rule, and do in our final system.

[3] Alignments a link arbitrary spans of s and s' (including non-constituents and individual words). We discuss the relation to word-level alignments in Section 4.

We emphasize that because of the synchronization features, this final form does *not* admit any known efficient dynamic programming for the exact computation of expectations. We will therefore turn to a variational inference method in Section 6.

4 Features

With the model's locality structure defined, we just need to specify the actual feature function, ϕ. We divide the features into three types: parsing features ($\phi_{\mathcal{F}}(n, s)$ and $\phi_{\mathcal{E}}(n', s')$), alignment features ($\phi_{\mathcal{A}}(b, s, s')$) and synchronization features ($\phi_{\triangleright}(n, b)$, $\phi_{\triangleleft}(b, n')$, and $\phi_{\bowtie}(n, b, n')$). We detail each of these in turn here.

4.1 Parsing

The monolingual parsing features we use are simply parsing model scores under the parser of Petrov and Klein (2007). While that parser uses heavily refined PCFGs with rule probabilities defined at the refined symbol level, we interact with its posterior distribution via posterior marginal probabilities over unrefined symbols. In particular, to each unrefined anchored production ${}_i\mathrm{A}_j \rightarrow {}_i\mathrm{B}_k\mathrm{C}_j$, we associate a single feature whose value is the marginal quantity $\log \mathrm{P}({}_i\mathrm{B}_k\mathrm{C}_j|{}_i\mathrm{A}_j, s)$ under the monolingual parser. These scores are the same as the variational rule scores of Matsuzaki et al. (2005).[4]

4.2 Alignment

We begin with the same set of alignment features as Haghighi et al. (2009), which are defined only for terminal bispans. In addition, we include features on nonterminal bispans, including a bias feature, features that measure the difference in size between the source and target spans, features that measure the difference in relative sentence position between the source and target spans, and features that measure the density of word-to-word alignment posteriors under a separate unsupervised word alignment model.

[4]Of course the structure of our model permits any of the additional rule-factored monolingual parsing features that have been described in the parsing literature, but in the present work we focus on the contributions of joint modeling.

4.3 Synchronization

Our synchronization features are indicators for the syntactic types of the participating nodes. We determine types at both a coarse (more collapsed than Treebank symbols) and fine (Treebank symbol) level. At the coarse level, we distinguish between phrasal nodes (e.g. S, NP), synthetic nodes introduced in the process of binarizing the grammar (e.g. S′, NP′), and part-of-speech nodes (e.g. NN, VBZ). At the fine level, we distinguish all nodes by their exact label. We use coarse and fine types for both partially synchronized (source-alignment or target-alignment) features and completely synchronized (source-alignment-target) features. The inset of Figure 2 shows some sample features. Of course, we could devise even more sophisticated features by using the input text itself. As we shall see, however, our model gives significant improvements with these simple features alone.

5 Learning

We learn the parameters of our model on the parallel portion of the Chinese treebank. Although our model assigns probabilities to entire synchronous derivations of sentences, the parallel Chinese treebank gives alignments only at the word level (1 by 1 bispans in Figure 2). This means that our alignment variable a is not fully observed. Because of this, given a particular word alignment w, we maximize the marginal probability of the set of derivations $\mathcal{A}(w)$ that are consistent with w (Haghighi et al., 2009).[5]

$$\mathcal{L}(\theta) = \log \sum_{a \in \mathcal{A}(w_i)} \mathrm{P}(t_i, a, t'_i | s_i, s'_i)$$

We maximize this objective using standard gradient methods (Nocedal and Wright, 1999). As with fully visible log-linear models, the gradient for the i^{th} sentence pair with respect to θ is a difference of feature expectations:

$$\begin{aligned}\nabla\mathcal{L}(\theta) = &\mathrm{E}_{\mathrm{P}(a|t_i,w_i,t'_i,s_i,s'_i)}\left[\phi(t_i, a, t'_i, s_i, s'_i)\right] \\ &- \mathrm{E}_{\mathrm{P}(t,a,t'|s_i,s'_i)}\left[\phi(t, a, t', s_i, s'_i)\right]\end{aligned} \quad (4)$$

[5]We also learn from non-ITG alignments by maximizing the marginal probability of the set of minimum-recall error alignments in the same way as Haghighi et al. (2009)

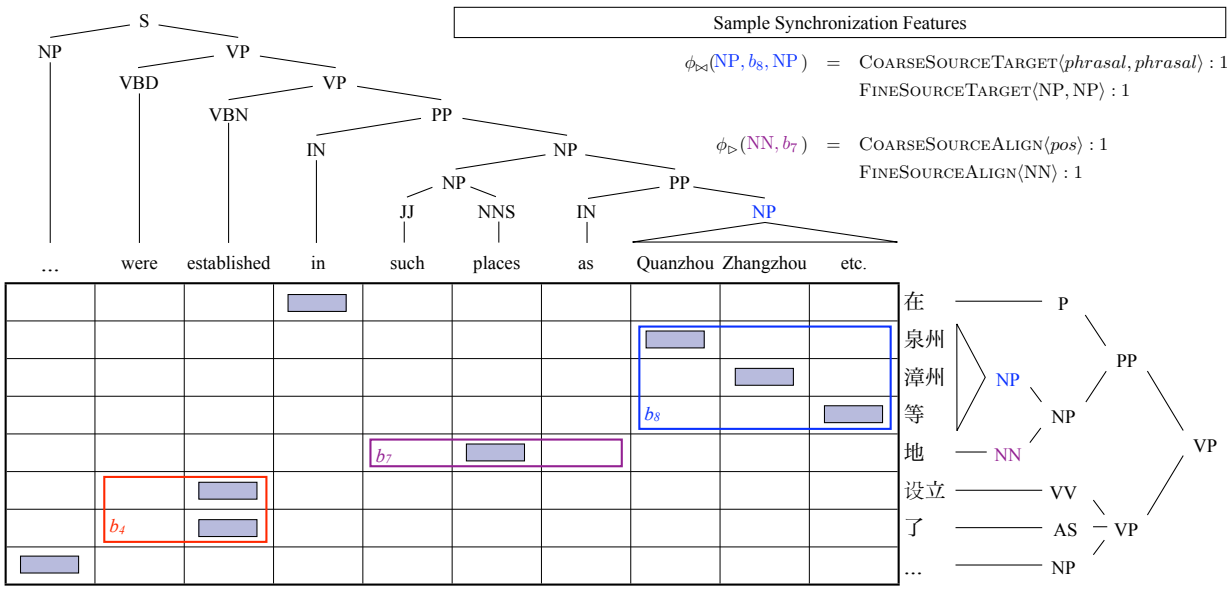


Figure 2: An example of a Chinese-English sentence pair with parses, word alignments, and a subset of the full optimal ITG derivation, including one totally unsynchronized bispan (b_4), one partially synchronized bispan (b_7), and and fully synchronized bispan (b_8). The inset provides some examples of active synchronization features (see Section 4.3) on these bispans. On this example, the monolingual English parser erroneously attached the lower PP to the VP headed by *established*, and the non-syntactic ITG word aligner misaligned 等 to *such* instead of to *etc.* Our joint model corrected both of these mistakes because it was rewarded for the synchronization of the two NPs joined by b_8.

We cannot efficiently compute the model expectations in this equation exactly. Therefore we turn next to an approximate inference method.

6 Mean Field Inference

Instead of computing the model expectations from (4), we compute the expectations for each sentence pair with respect to a simpler, fully factored distribution $\text{Q}(t, a, t') = q(t)q(a)q(t')$. Rewriting Q in log-linear form, we have:

$$\text{Q}(t, a, t') \propto \exp\left[\sum_{n \in t} \psi_n + \sum_{b \in a} \psi_b + \sum_{n' \in t'} \psi_{n'}\right]$$

Here, the ψ_n, ψ_b and $\psi_{n'}$ are variational parameters which we set to best approximate our weakly synchronized model from (3):

$$\psi^* = \operatorname*{argmin}_{\psi} \text{KL}\left(\text{Q}_\psi || P_\theta(t, a, t' | s, s')\right)$$

Once we have found Q, we compute an approximate gradient by replacing the model expectations with expectations under Q:

$$\text{E}_{\text{Q}(a|w_i)}\left[\phi(t_i, a, t'_i, s_i, s'_i)\right] - \text{E}_{\text{Q}(t,a,t'|s_i,s'_i)}\left[\phi(t, a, t', s_i, s'_i)\right]$$

Now, we will briefly describe how we compute Q. First, note that the parameters ψ of Q factor along individual source nodes, target nodes, and bispans. The combination of the KL objective and our particular factored form of Q make our inference procedure a structured mean field algorithm (Saul and Jordan, 1996). Structured mean field techniques are well-studied in graphical models, and our adaptation in this section to multiple grammars follows standard techniques (see e.g. Wainwright and Jordan, 2008).

Rather than derive the mean field updates for ψ, we describe the algorithm (shown in Figure 3) procedurally. Similar to block Gibbs sampling, we iteratively optimize each component (source parse, target parse, and alignment) of the model in turn, conditioned on the others. Where block Gibbs sampling conditions on fixed trees or ITG derivations, our mean field algorithm maintains uncertainty in

Input: sentence pair (s, s')
parameter vector θ

Output: variational parameters ψ

1. Initialize

$\psi_n^0 \leftarrow \theta^\top \phi_{\mathcal{F}}(n, s)$
$\psi_b^0 \leftarrow \theta^\top \phi_{\mathcal{A}}(b, s, s')$
$\psi_{n'}^0 \leftarrow \theta^\top \phi_{\mathcal{E}}(n', s')$

$\mu_n^0 \leftarrow \sum_t q_{\psi^0}(t) I(n \in t)$, etc for μ_b^0, $\mu_{n'}^0$

2. While not converged, for each n, n', b in the monolingual and ITG charts

$$\psi_n^i \leftarrow \theta^\top \Big(\phi_{\mathcal{F}}(n, s) + \textstyle\sum_{b, n \triangleright b} \mu_b^{i-1} \phi_{\triangleright}(n, b) + \sum_{b, n \triangleright b} \sum_{n', b \triangleleft n'} \mu_b^{i-1} \mu_{n'}^{i-1} \phi_{\bowtie}(n, b, n')\Big)$$

$\mu_n^i \leftarrow \sum_t q_\psi(t) I(n \in t)$ (inside-outside)

$$\psi_b^i \leftarrow \theta^\top \Big(\phi_{\mathcal{A}}(b, s, s') + \textstyle\sum_{n, n \triangleright b} \mu_n^{i-1} \phi_{\triangleright}(n, b) + \sum_{n', b \triangleleft n'} \mu_{n'}^{i-1} \phi_{\triangleleft}(b, n') + \sum_{n, n \triangleright b} \sum_{n', b \triangleleft n'} \mu_n^{i-1} \mu_{n'}^{i-1} \phi_{\bowtie}(n, b, n')\Big)$$

$\mu_b \leftarrow \sum_a q_\psi(a) I(b \in a)$ (bitext inside-outside)

updates for $\psi_{n'}^i$, $\mu_{n'}^i$ analogous to ψ_n^i, μ_n^i

3. Return variational parameters ψ

Figure 3: Structured mean field inference for the weakly synchronized model. $I(n \in t)$ is an indicator value for the presence of node n in source tree t.

the form of monolingual parse forests or ITG forests. The key components to this uncertainty are the expected counts of particular source nodes, target nodes, and bispans under the mean field distribution:

$$\mu_n = \sum_t q_\psi(t) I(n \in t)$$
$$\mu_{n'} = \sum_{t'} q_\psi(t') I(n' \in t')$$
$$\mu_b = \sum_a q_\psi(a) I(b \in a)$$

Since dynamic programs exist for summing over each of the individual factors, these expectations can be computed in polynomial time.

6.1 Pruning

Although we can approximate the expectations from (4) in polynomial time using our mean field distribution, in practice we must still prune the ITG forests and monolingual parse forests to allow tractable inference. We prune our ITG forests using the same basic idea as Haghighi et al. (2009), but we employ a technique that allows us to be more aggressive. Where Haghighi et al. (2009) pruned bispans based on how many unsupervised HMM alignments were violated, we first train a maximum-matching word aligner (Taskar et al., 2005) using our supervised data set, which has only half the precision errors of the unsupervised HMM. We then prune every bispan which violates at least three alignments from the maximum-matching aligner. When compared to pruning the bitext forest of our model with Haghighi et al. (2009)'s HMM technique, this new technique allows us to maintain the same level of accuracy while cutting the number of bispans in half.

In addition to pruning the bitext forests, we also prune the syntactic parse forests using the monolingual parsing model scores. For each unrefined anchored production ${}_i\mathrm{A}_j \rightarrow {}_i\mathrm{B}_k \mathrm{C}_j$, we compute the marginal probability $\mathrm{P}({}_i\mathrm{A}_j, {}_i\mathrm{B}_k, {}_k\mathrm{C}_j | s)$ under the monolingual parser (these are equivalent to the maxrule scores from Petrov and Klein 2007). We only include productions where this probability is greater than 10^{-20}. Note that at training time, we are not guaranteed that the gold trees will be included in the pruned forest. Because of this, we replace the gold trees t_i, t_i' with oracle trees from the pruned forest, which can be found efficiently using a variant of the inside algorithm (Huang, 2008).

7 Testing

Once the model has been trained, we still need to determine how to use it to predict parses and word alignments for our test sentence pairs. Ideally, given the sentence pair (s, s'), we would find:

$$\begin{aligned}(t^*, w^*, t'^*) &= \operatorname*{argmax}_{t,w,t'} \mathrm{P}(t, w, t' | s, s') \\ &= \operatorname*{argmax}_{t,w,t'} \sum_{a \in \mathcal{A}(w)} \mathrm{P}(t, a, t' | s, s')\end{aligned}$$

Of course, this is also intractable, so we once again resort to our mean field approximation. This yields the approximate solution:

$$(t^*, w^*, t'^*) = \operatorname*{argmax}_{t,w,t'} \sum_{a \in \mathcal{A}(w)} \mathrm{Q}(t, a, t')$$

However, recall that Q incorporates the model's mutual constraint into the variational parameters, which

factor into $q(t)$, $q(a)$, and $q(t')$. This allows us to simplify further, and find the maximum a posteriori assignments under the variational distribution. The trees can be found quickly using the Viterbi inside algorithm on their respective qs. However, the sum for computing w^* under q is still intractable.

As we cannot find the maximum probability word alignment, we provide two alternative approaches for finding w^*. The first is to just find the Viterbi ITG derivation $a^* = \text{argmax}_a\, q(a)$ and then set w^* to contain exactly the 1x1 bispans in a^*. The second method, posterior thresholding, is to compute posterior marginal probabilities under q for each 1x1 cell beginning at position i, j in the word alignment grid:

$$m(i,j) = \sum_a q(a) I((i, i+1, j, j+1) \in a)$$

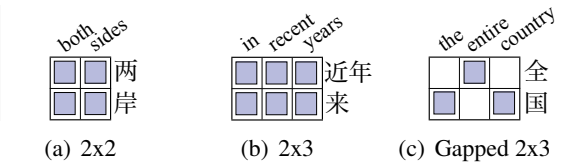
We then include $w(i,j)$ in w^* if $m(w(i,j)) > \tau$, where τ is a threshold chosen to trade off precision and recall. For our experiments, we found that the Viterbi alignment was uniformly worse than posterior thresholding. All the results from the next section use the threshold $\tau = 0.25$.

8 Experiments

We trained and tested our model on the translated portion of the Chinese treebank (Bies et al., 2007), which includes hand annotated Chinese and English parses and word alignments. We separated the data into three sets: *train*, *dev*, and *test*, according to the standard Chinese treebank split. To speed up training, we only used training sentences of length ≤ 50 words, which left us with 1974 of 2261 sentences. We measured the results in two ways. First, we directly measured F_1 for English parsing, Chinese parsing, and word alignment on a held out section of the hand annotated corpus used to train the model. Next, we further evaluated the quality of the word alignments produced by our model by using them as input for a machine translation system.

8.1 Dataset-specific ITG Terminals

The Chinese treebank gold word alignments include significantly more many-to-many word alignments than those used by Haghighi et al. (2009). We are able to produce some of these many-to-many alignments by including new many-to-many terminals in our ITG word aligner, as shown in Figure 4. Our terminal productions sometimes capture non-literal translation like *both sides* or *in recent years*. They also can allow us to capture particular, systematic changes in the annotation standard. For example, the gapped pattern from Figure 4 captures the standard that English word *the* is always aligned to the Chinese head noun in a noun phrase. We featurize these non-terminals with features similar to those of Haghighi et al. (2009), and all of the alignment results we report in Section 8.2 (both joint and ITG) employ these features.



(a) 2x2 (b) 2x3 (c) Gapped 2x3

Figure 4: Examples of phrasal alignments that can be represented by our new ITG terminal bispans.

8.2 Parsing and Word Alignment

To compute features that depend on external models, we needed to train an unsupervised word aligner and monolingual English and Chinese parsers. The unsupervised word aligner was a pair of jointly trained HMMs (Liang et al., 2006), trained on the FBIS corpus. We used the Berkeley Parser (Petrov and Klein, 2007) for both monolingual parsers, with the Chinese parser trained on the full Chinese treebank, and the English parser trained on a concatenation of the Penn WSJ corpus (Marcus et al., 1993) and the English side of *train*.[6]

We compare our parsing results to the monolingual parsing models and to the English-Chinese bilingual reranker of Burkett and Klein (2008), trained on the same dataset. The results are in Table 1. For word alignment, we compare to

[6]To avoid overlap in the data used to train the monolingual parsers and the joint model, at training time, we used a separate version of the Chinese parser, trained only on articles 400-1151 (omitting articles in *train*). For English parsing, we deemed it insufficient to entirely omit the Chinese treebank data from the monolingual parser's training set, as otherwise the monolingual parser would be trained entirely on out-of-domain data. Therefore, at training time we used two separate English parsers: to compute model scores for the *first* half of *train*, we used a parser trained on a concatenation of the WSJ corpus and the *second* half of *train*, and vice versa for the remaining sentences.

	Test Results		
	Ch F_1	Eng F_1	Tot F_1
Monolingual	83.6	81.2	82.5
Reranker	**86.0**	83.8	84.9
Joint	85.7	**84.5**	**85.1**

Table 1: Parsing results. Our joint model has the highest reported F_1 for English-Chinese bilingual parsing.

	Test Results			
	Precision	Recall	AER	F_1
HMM	86.0	58.4	30.0	69.5
ITG	**86.8**	73.4	20.2	79.5
Joint	85.5	**84.6**	**14.9**	**85.0**

Table 2: Word alignment results. Our joint model has the highest reported F_1 for English-Chinese word alignment.

the baseline unsupervised HMM word aligner and to the English-Chinese ITG-based word aligner of Haghighi et al. (2009). The results are in Table 2.

As can be seen, our model makes substantial improvements over the independent models. For parsing, we improve absolute F_1 over the monolingual parsers by 2.1 in Chinese, and by 3.3 in English. For word alignment, we improve absolute F_1 by 5.5 over the non-syntactic ITG word aligner. In addition, our English parsing results are better than those of the Burkett and Klein (2008) bilingual reranker, the current top-performing English-Chinese bilingual parser, despite ours using a much simpler set of synchronization features.

8.3 Machine Translation

We further tested our alignments by using them to train the Joshua machine translation system (Li and Khudanpur, 2008). Table 3 describes the results of our experiments. For all of the systems, we tuned on 1000 sentences of the NIST 2004 and 2005 machine translation evaluations, and tested on 400 sentences of the NIST 2006 MT evaluation. Our training set consisted of 250k sentences of newswire distributed with the GALE project, all of which were sub-sampled to have high Ngram overlap with the tune and test sets. All of our sentences were of length at most 40 words. When building the translation grammars, we used Joshua's default "tight" phrase extraction option. We ran MERT for 4 iterations, optimizing 20 weight vectors per iteration on a 200-best list.

	Rules	Tune	Test
HMM	1.1M	29.0	29.4
ITG	1.5M	**29.9**	30.4†
Joint	1.5M	29.6	**30.6**

Table 3: Tune and test BLEU results for machine translation systems built with different alignment tools. † indicates a statistically significant difference between a system's test performance and the one above it.

Table 3 gives the results. On the test set, we also ran the approximate randomization test suggested by Riezler and Maxwell (2005). We found that our joint parsing and alignment system significantly outperformed the HMM aligner, but the improvement over the ITG aligner was not statistically significant.

9 Conclusion

The quality of statistical machine translation models depends crucially on the quality of word alignments and syntactic parses for the bilingual training corpus. Our work presented the first joint model for parsing and alignment, demonstrating that we can improve results on both of these tasks, as well as on downstream machine translation, by allowing parsers and word aligners to simultaneously inform one another. Crucial to this improved performance is a notion of weak synchronization, which allows our model to learn when pieces of a grammar are synchronized and when they are not. Although exact inference in the weakly synchronized model is intractable, we developed a mean field approximate inference scheme based on monolingual and bitext parsing, allowing for efficient inference.

Acknowledgements

We thank Adam Pauls and John DeNero for their help in running machine translation experiments. We also thank the three anonymous reviewers for their helpful comments on an earlier draft of this paper. This project is funded in part by NSF grants 0915265 and 0643742, an NSF graduate research fellowship, the CIA under grant HM1582-09-1-0021, and BBN under DARPA contract HR0011-06-C-0022.

References

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

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

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL*.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using syntax to improve word alignment for syntax-based statistical machine translation. In *ACL MT Workshop*.

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

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

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *ACL*.

Zhifei Li and Sanjeev Khudanpur. 2008. A scalable decoder for parsing-based machine translation with equivalent language model state maintenance. In *ACL SSST*.

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

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.

Takuya Matsuzaki, Yusuki Miyao, and Jun'ichi Tsujii. 2005. Probabilistic CFG with latent annotations. In *ACL*.

Jon May and Kevin Knight. 2007. Syntactic realignment models for machine translation. In *EMNLP*.

Jorge Nocedal and Stephen J. Wright. 1999. *Numerical Optimization*. Springer.

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

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Workshop on Intrinsic and Extrinsic Evaluation Methods for MT and Summarization, ACL*.

Lawrence Saul and Michael Jordan. 1996. Exploiting tractable substructures in intractable networks. In *NIPS*.

Stuart M. Shieber and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *ACL*.

David A. Smith and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *HLT-NAACL*.

David A. Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *EMNLP*.

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

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

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *EMNLP*.

Martin J Wainwright and Michael I Jordan. 2008. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Andreas Zollmann, Ashish Venugopal, Stephan Vogel, and Alex Waibel. 2006. The CMU-AKA syntax augmented machine translation system for IWSLT-06. In *IWSLT*.