University of Pennsylvania
ScholarlyCommons

Lab Papers (GRASP)

General Robotics, Automation, Sensing and Perception Laboratory

8-2-2009

Dependency Grammar Induction via Bitext Projection Constraints

Kuzman Ganchev
University of Pennsylvania, kuzman@seas.upenn.edu

Jennifer Gillenwater
University of Pennsylvania, jengi@seas.upenn.edu

Ben Taskar
University of Pennsylvania, taskar@central.cis.upenn.edu

Reprinted from:
Dependency Grammar Induction via Bitext Projection Constraints. Kuzman Ganchev, Jennifer Gillenwater and Ben Taskar. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, Aug. 2-7, 2009. pp. 369-377.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/grasp_papers/43
For more information, please contact repository@pobox.upenn.edu.

Dependency Grammar Induction via Bitext Projection Constraints

Kuzman Ganchev and **Jennifer Gillenwater** and **Ben Taskar**
Department of Computer and Information Science
University of Pennsylvania, Philadelphia PA, USA
{kuzman,jengi,taskar}@seas.upenn.edu

Abstract

Broad-coverage annotated treebanks necessary to train parsers do not exist for many resource-poor languages. The wide availability of parallel text and accurate parsers in English has opened up the possibility of grammar induction through partial transfer across bitext. We consider generative and discriminative models for dependency grammar induction that use word-level alignments and a source language parser (English) to constrain the space of possible target trees. Unlike previous approaches, our framework does not require full projected parses, allowing partial, approximate transfer through linear expectation constraints on the space of distributions over trees. We consider several types of constraints that range from generic dependency conservation to language-specific annotation rules for auxiliary verb analysis. We evaluate our approach on Bulgarian and Spanish CoNLL shared task data and show that we consistently outperform unsupervised methods and can outperform supervised learning for limited training data.

1 Introduction

For English and a handful of other languages, there are large, well-annotated corpora with a variety of linguistic information ranging from named entity to discourse structure. Unfortunately, for the vast majority of languages very few linguistic resources are available. This situation is likely to persist because of the expense of creating annotated corpora that require linguistic expertise (Abeillé, 2003). On the other hand, parallel corpora between many resource-poor languages and resource-rich languages are ample, motivating recent interest in transferring linguistic resources from one language to another via parallel text. For example, several early works (Yarowsky and Ngai, 2001; Yarowsky et al., 2001; Merlo et al., 2002) demonstrate transfer of shallow processing tools such as part-of-speech taggers and noun-phrase chunkers by using word-level alignment models (Brown et al., 1994; Och and Ney, 2000).

Alshawi et al. (2000) and Hwa et al. (2005) explore transfer of deeper syntactic structure: dependency grammars. Dependency and constituency grammar formalisms have long coexisted and competed in linguistics, especially beyond English (Mel'čuk, 1988). Recently, dependency parsing has gained popularity as a simpler, computationally more efficient alternative to constituency parsing and has spurred several supervised learning approaches (Eisner, 1996; Yamada and Matsumoto, 2003a; Nivre and Nilsson, 2005; McDonald et al., 2005) as well as unsupervised induction (Klein and Manning, 2004; Smith and Eisner, 2006). Dependency representation has been used for language modeling, textual entailment and machine translation (Haghighi et al., 2005; Chelba et al., 1997; Quirk et al., 2005; Shen et al., 2008), to name a few tasks.

Dependency grammars are arguably more robust to transfer since syntactic relations between aligned words of parallel sentences are better conserved in translation than phrase structure (Fox, 2002; Hwa et al., 2005). Nevertheless, several challenges to accurate training and evaluation from aligned bitext remain: (1) partial word alignment due to non-literal or distant translation; (2) errors in word alignments and source language parses, (3) grammatical annotation choices that differ across languages and linguistic theories (e.g., how to analyze auxiliary verbs, conjunctions).

In this paper, we present a flexible learning

framework for transferring dependency grammars via bitext using the posterior regularization framework (Graça et al., 2008). In particular, we address challenges (1) and (2) by avoiding commitment to an entire projected parse tree in the target language during training. Instead, we explore formulations of both generative and discriminative probabilistic models where projected syntactic relations are constrained to hold approximately and only in expectation. Finally, we address challenge (3) by introducing a very small number of language-specific constraints that disambiguate arbitrary annotation choices.

We evaluate our approach by transferring from an English parser trained on the Penn treebank to Bulgarian and Spanish. We evaluate our results on the Bulgarian and Spanish corpora from the CoNLL X shared task. We see that our transfer approach consistently outperforms unsupervised methods and, given just a few (2 to 7) language-specific constraints, performs comparably to a supervised parser trained on a very limited corpus (30 - 140 training sentences).

2 Approach

At a high level our approach is illustrated in Figure 1(a). A parallel corpus is word-level aligned using an alignment toolkit (Graça et al., 2009) and the source (English) is parsed using a dependency parser (McDonald et al., 2005). Figure 1(b) shows an aligned sentence pair example where dependencies are perfectly conserved across the alignment. An edge from English parent p to child c is called conserved if word p aligns to word p' in the second language, c aligns to c' in the second language, and p' is the parent of c'. Note that we are not restricting ourselves to one-to-one alignments here; p, c, p', and c' can all also align to other words. After filtering to identify well-behaved sentences and high confidence projected dependencies, we learn a probabilistic parsing model using the posterior regularization framework (Graça et al., 2008). We estimate both generative and discriminative models by constraining the posterior distribution over possible target parses to approximately respect projected dependencies and other rules which we describe below. In our experiments we evaluate the learned models on dependency treebanks (Nivre et al., 2007).

Unfortunately the sentence in Figure 1(b) is highly unusual in its amount of dependency conservation. To get a feel for the typical case, we used off-the-shelf parsers (McDonald et al., 2005) for English, Spanish and Bulgarian on two bitexts (Koehn, 2005; Tiedemann, 2007) and compared several measures of dependency conservation. For the English-Bulgarian corpus, we observed that 71.9% of the edges we projected were edges in the corpus, and we projected on average 2.7 edges per sentence (out of 5.3 tokens on average). For Spanish, we saw conservation of 64.4% and an average of 5.9 projected edges per sentence (out of 11.5 tokens on average).

As these numbers illustrate, directly transferring information one dependency edge at a time is unfortunately error prone for two reasons. First, parser and word alignment errors cause much of the transferred information to be wrong. We deal with this problem by constraining groups of edges rather than a single edge. For example, in some sentence pair we might find 10 edges that have both end points aligned and can be transferred. Rather than requiring our target language parse to contain each of the 10 edges, we require that the expected number of edges from this set is at least 10η, where η is a strength parameter. This gives the parser freedom to have some uncertainty about which edges to include, or alternatively to choose to exclude some of the transferred edges.

A more serious problem for transferring parse information across languages are structural differences and grammar annotation choices between the two languages. For example dealing with auxiliary verbs and reflexive constructions. Hwa et al. (2005) also note these problems and solve them by introducing dozens of rules to transform the transferred parse trees. We discuss these differences in detail in the experimental section and use our framework introduce a very small number of rules to cover the most common structural differences.

3 Parsing Models

We explored two parsing models: a generative model used by several authors for unsupervised induction and a discriminative model used for fully supervised training.

The discriminative parser is based on the edge-factored model and features of the MST-Parser (McDonald et al., 2005). The parsing model defines a conditional distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$ over each projective parse tree $\mathbf{z}$ for a particular sentence $\mathbf{x}$, parameterized by a vector θ. The prob-

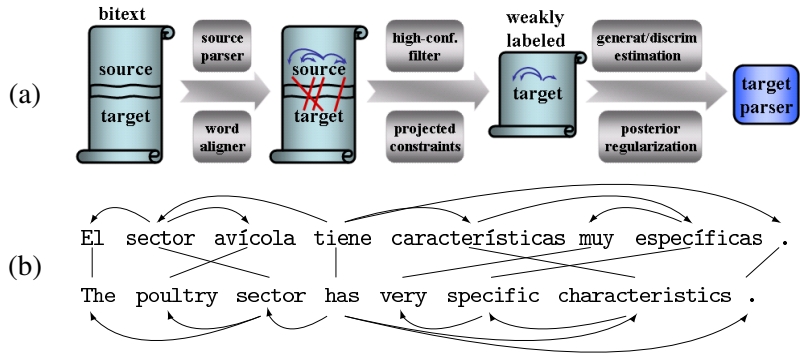

Figure 1: (a) Overview of our grammar induction approach via bitext: the source (English) is parsed and word-aligned with target; after filtering, projected dependencies define constraints over target parse tree space, providing weak supervision for learning a target grammar. (b) An example word-aligned sentence pair with perfectly projected dependencies.

ability of any particular parse is

$$p_\theta(\mathbf{z} \mid \mathbf{x}) \propto \prod_{z \in \mathbf{z}} e^{\theta \cdot \phi(z, \mathbf{x})}, \tag{1}$$

where z is a directed edge contained in the parse tree $\mathbf{z}$ and ϕ is a feature function. In the fully supervised experiments we run for comparison, parameter estimation is performed by stochastic gradient ascent on the conditional likelihood function, similar to maximum entropy models or conditional random fields. One needs to be able to compute expectations of the features $\phi(z, \mathbf{x})$ under the distribution $p_\theta(z \mid \mathbf{x})$. A version of the inside-outside algorithm (Lee and Choi, 1997) performs this computation. Viterbi decoding is done using Eisner's algorithm (Eisner, 1996).

We also used a generative model based on dependency model with valence (Klein and Manning, 2004). Under this model, the probability of a particular parse $\mathbf{z}$ and a sentence with part of speech tags $\mathbf{x}$ is given by

$$\begin{aligned} p_\theta(\mathbf{z}, \mathbf{x}) = {} & p_{\text{root}}(r(\mathbf{x})) \cdot \\ & \Big(\prod_{z \in \mathbf{z}} p_{\neg\text{stop}}(z_p, z_d, v_z)\, p_{\text{child}}(z_p, z_d, z_c) \Big) \cdot \\ & \Big(\prod_{x \in \mathbf{x}} p_{\text{stop}}(x, \text{left}, v_l)\, p_{\text{stop}}(x, \text{right}, v_r) \Big) \end{aligned} \tag{2}$$

where $r(\mathbf{x})$ is the part of speech tag of the root of the parse tree $\mathbf{z}$, z is an edge from parent z_p to child z_c in direction z_d, either left or right, and v_z indicates valency—false if z_p has no other children further from it in direction z_d than z_c, true otherwise. The valencies v_r/v_l are marked as true if x has any children on the left/right in $\mathbf{z}$, false otherwise.

4 Posterior Regularization

Graça et al. (2008) introduce an estimation framework that incorporates side-information into unsupervised problems in the form of linear constraints on posterior expectations. In grammar transfer, our basic constraint is of the form: the expected proportion of conserved edges in a sentence pair is at least η (the exact proportion we used was 0.9, which was determined using unlabeled data as described in Section 5). Specifically, let $C_\mathbf{x}$ be the set of directed edges projected from English for a given sentence $\mathbf{x}$, then given a parse $\mathbf{z}$, the proportion of conserved edges is $f(\mathbf{x}, \mathbf{z}) = \frac{1}{|C_\mathbf{x}|} \sum_{z \in \mathbf{z}} \mathbf{1}(z \in C_\mathbf{x})$ and the expected proportion of conserved edges under distribution $p(\mathbf{z} \mid \mathbf{x})$ is

$$\mathbf{E}_p[f(\mathbf{x}, \mathbf{z})] = \frac{1}{|C_\mathbf{x}|} \sum_{z \in C_\mathbf{x}} p(z \mid \mathbf{x}).$$

The posterior regularization framework (Graça et al., 2008) was originally defined for generative unsupervised learning. The standard objective is to minimize the negative marginal log-likelihood of the data : $\widehat{\mathbf{E}}[-\log p_\theta(\mathbf{x})] = \widehat{\mathbf{E}}[-\log \sum_\mathbf{z} p_\theta(\mathbf{z}, \mathbf{x})]$ over the parameters θ (we use $\widehat{\mathbf{E}}$ to denote expectation over the sample sentences $\mathbf{x}$). We typically also add standard regularization term on θ, resulting from a parameter prior $-\log p(\theta) = R(\theta)$, where $p(\theta)$ is Gaussian for the MST-Parser models and Dirichlet for the valence model.

To introduce supervision into the model, we define a set $\mathcal{Q}_\mathbf{x}$ of distributions over the hidden variables $\mathbf{z}$ satisfying the desired posterior constraints in terms of linear equalities or inequalities on feature expectations (we use inequalities in this paper):

$$\mathcal{Q}_\mathbf{x} = \{q(\mathbf{z}) : \mathbf{E}[\mathbf{f}(\mathbf{x}, \mathbf{z})] \leq \mathbf{b}\}.$$

Basic Uni-gram Features	Basic Bi-gram Features	In Between POS Features
x_i-word, x_i-pos	x_i-word, x_i-pos, x_j-word, x_j-pos	x_i-pos, b-pos, x_j-pos
x_i-word	x_i-pos, x_j-word, x_j-pos	
x_i-pos	x_i-word, x_j-word, x_j-pos	**Surrounding Word POS Features**
x_j-word, x_j-pos	x_i-word, x_i-pos, x_j-pos	x_i-pos, x_i-pos+1, x_j-pos-1, x_j-pos
x_j-word	x_i-word, x_i-pos, x_j-word	x_i-pos-1, x_i-pos, x_j-pos-1, x_j-pos
x_j-pos	x_i-word, x_j-word	x_i-pos, x_i-pos+1, x_j-pos, x_j-pos+1
	x_i-pos, x_j-pos	x_i-pos-1, x_i-pos, x_j-pos, x_j-pos+1

Table 1: Features used by the MSTParser. For each edge (i, j), x_i-word is the parent word and x_j-word is the child word, analogously for POS tags. The +1 and -1 denote preceeding and following tokens in the sentence, while b denotes tokens between x_i and x_j.

In this paper, for example, we use the conserved-edge-proportion constraint as defined above. The marginal log-likelihood objective is then modified with a penalty for deviation from the desired set of distributions, measured by KL-divergence from the set $\mathcal{Q}_{\mathbf{x}}$, $\mathrm{KL}(\mathcal{Q}_{\mathbf{x}}||p_\theta(\mathbf{z}|\mathbf{x})) = \min_{q\in\mathcal{Q}_{\mathbf{x}}} \mathrm{KL}(q(\mathbf{z})||p_\theta(\mathbf{z}|\mathbf{x}))$. The generative learning objective is to minimize:

$$\widehat{\mathbf{E}}[-\log p_\theta(\mathbf{x})] + R(\theta) + \widehat{\mathbf{E}}[\mathrm{KL}(\mathcal{Q}_{\mathbf{x}}||p_\theta(\mathbf{z} \mid \mathbf{x}))].$$

For discriminative estimation (Ganchev et al., 2008), we do not attempt to model the marginal distribution of $\mathbf{x}$, so we simply have the two regularization terms:

$$R(\theta) + \widehat{\mathbf{E}}[\mathrm{KL}(\mathcal{Q}_{\mathbf{x}}||p_\theta(\mathbf{z} \mid \mathbf{x}))].$$

Note that the idea of regularizing moments is related to generalized expectation criteria algorithm of Mann and McCallum (2007), as we discuss in the related work section below. In general, the objectives above are not convex in θ. To optimize these objectives, we follow an Expectation Maximization-like scheme. Recall that standard EM iterates two steps. An E-step computes a probability distribution over the model's hidden variables (posterior probabilities) and an M-step that updates the model's parameters based on that distribution. The posterior-regularized EM algorithm leaves the M-step unchanged, but involves projecting the posteriors onto a constraint set after they are computed for each sentence $\mathbf{x}$:

$$\begin{aligned} &\arg\min_q \mathrm{KL}(q(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \\ &\quad \text{s.t. } \mathbf{E}_q[\mathbf{f}(\mathbf{x}, \mathbf{z})] \leq \mathbf{b}, \end{aligned} \tag{3}$$

where $p_\theta(\mathbf{z}|\mathbf{x})$ are the posteriors. The new posteriors $q(\mathbf{z})$ are used to compute sufficient statistics for this instance and hence to update the model's parameters in the M-step for either the generative or discriminative setting.

The optimization problem in Equation 3 can be efficiently solved in its dual formulation:

$$\arg\min_{\lambda\geq 0} \mathbf{b}^\top\lambda + \log \sum_{\mathbf{z}} p_\theta(\mathbf{z} \mid \mathbf{x}) \exp\{-\lambda^\top \mathbf{f}(\mathbf{x}, \mathbf{z})\}. \tag{4}$$

Given λ, the primal solution is given by: $q(\mathbf{z}) = p_\theta(\mathbf{z} \mid \mathbf{x}) \exp\{-\lambda^\top \mathbf{f}(\mathbf{x}, \mathbf{z})\}/Z$, where Z is a normalization constant. There is one dual variable per expectation constraint, and we can optimize them by projected gradient descent, similar to log-linear model estimation. The gradient with respect to λ is given by: $\mathbf{b} - \mathbf{E}_q[\mathbf{f}(\mathbf{x}, \mathbf{z})]$, so it involves computing expectations under the distribution $q(\mathbf{z})$. This remains tractable as long as features factor by edge, $f(\mathbf{x}, \mathbf{z}) = \sum_{z\in\mathbf{z}} f(\mathbf{x}, z)$, because that ensures that $q(\mathbf{z})$ will have the same form as $p_\theta(\mathbf{z} \mid \mathbf{x})$. Furthermore, since the constraints are per instance, we can use incremental or online version of EM (Neal and Hinton, 1998), where we update parameters θ after posterior-constrained E-step on each instance $\mathbf{x}$.

5 Experiments

We conducted experiments on two languages: Bulgarian and Spanish, using each of the parsing models. The Bulgarian experiments transfer a parser from English to Bulgarian, using the OpenSubtitles corpus (Tiedemann, 2007). The Spanish experiments transfer from English to Spanish using the Spanish portion of the Europarl corpus (Koehn, 2005). For both corpora, we performed word alignments with the open source PostCAT (Graça et al., 2009) toolkit. We used the Tokyo tagger (Tsuruoka and Tsujii, 2005) to POS tag the English tokens, and generated parses using the first-order model of McDonald et al. (2005) with projective decoding, trained on sections 2-21 of the Penn treebank with dependencies extracted using the head rules of Yamada and Matsumoto (2003b). For Bulgarian we trained the Stanford POS tagger (Toutanova et al., 2003) on the Bul-

	Discriminative model					Generative model				
	Bulgarian			Spanish		Bulgarian			Spanish	
	no rules	2 rules	7 rules	no rules	3 rules	no rules	2 rules	7 rules	no rules	3 rules
Baseline	63.8	72.1	72.6	67.6	69.0	66.5	69.1	**71.0**	68.2	71.3
Post.Reg.	**66.9**	**77.5**	**78.3**	**70.6**	**72.3**	**67.8**	**70.7**	70.8	**69.5**	**72.8**

Table 2: Comparison between transferring a single tree of edges and transferring all possible projected edges. The transfer models were trained on 10k sentences of length up to 20, all models tested on CoNLL train sentences of up to 10 words. Punctuation was stripped at train time.

gtreebank corpus from CoNLL X. The Spanish Europarl data was POS tagged with the FreeLing language analyzer (Atserias et al., 2006). The discriminative model used the same features as MST-Parser, summarized in Table 1.

In order to evaluate our method, we a baseline inspired by Hwa et al. (2005). The baseline constructs a full parse tree from the incomplete and possibly conflicting transferred edges using a simple random process. We start with no edges and try to add edges one at a time verifying at each step that it is possible to complete the tree. We first try to add the transferred edges in random order, then for each orphan node we try all possible parents (both in random order). We then use this full labeling as supervision for a parser. Note that this baseline is very similar to the first iteration of our model, since for a large corpus the different random choices made in different sentences tend to smooth each other out. We also tried to create rules for the adoption of orphans, but the simple rules we tried added bias and performed worse than the baseline we report. Table 2 shows attachment accuracy of our method and the baseline for both language pairs under several conditions. By attachment accuracy we mean the fraction of words assigned the correct parent. The experimental details are described in this section. Link-left baselines for these corpora are much lower: 33.8% and 27.9% for Bulgarian and Spanish respectively.

5.1 Preprocessing

Preliminary experiments showed that our word alignments were not always appropriate for syntactic transfer, even when they were correct for translation. For example, the English "bike/V" could be translated in French as "aller/V en vélo/N", where the word "bike" would be aligned with "vélo". While this captures some of the semantic shared information in the two languages, we have no expectation that the noun "vélo" will have a similar syntactic behavior to the verb "bike". To prevent such false transfer, we filter out alignments between incompatible POS tags. In both language pairs, filtering out noun-verb alignments gave the biggest improvement.

Both corpora also contain sentence fragments, either because of question responses or fragmented speech in movie subtitles or because of voting announcements and similar formulaic sentences in the parliamentary proceedings. We overcome this problem by filtering out sentences that do not have a verb as the English root or for which the English root is not aligned to a verb in the target language. For the subtitles corpus we also remove sentences that end in an ellipsis or contain more than one comma. Finally, following (Klein and Manning, 2004) we strip out punctuation from the sentences. For the discriminative model this did not affect results significantly but improved them slightly in most cases. We found that the generative model gets confused by punctuation and tends to predict that periods at the end of sentences are the parents of words in the sentence.

Our basic model uses constraints of the form: the expected proportion of conserved edges in a sentence pair is at least $\eta = 90\%$.[1]

5.2 No Language-Specific Rules

We call the generic model described above "no-rules" to distinguish it from the language-specific constraints we introduce in the sequel. The no rules columns of Table 2 summarize the performance in this basic setting. Discriminative models outperform the generative models in the majority of cases. The left panel of Table 3 shows the most common errors by child POS tag, as well as by true parent and guessed parent POS tag.

Figure 2 shows that the discriminative model continues to improve with more transfer-type data

[1] We chose η in the following way: we split the unlabeled parallel text into two portions. We trained a models with different η on one portion and ran it on the other portion. We chose the model with the highest fraction of conserved constraints on the second portion.

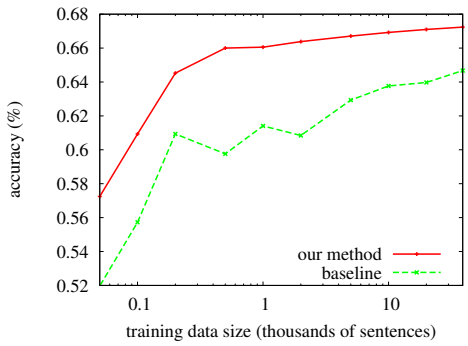

Figure 2: Learning curve of the discriminative no-rules transfer model on Bulgarian bitext, testing on CoNLL train sentences of up to 10 words.

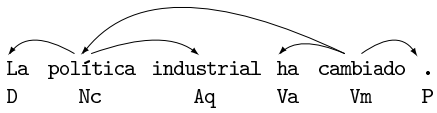

Figure 3: A Spanish example where an auxiliary verb dominates the main verb.

up to at least 40 thousand sentences.

5.3 Annotation guidelines and constraints

Using the straightforward approach outlined above is a dramatic improvement over the standard link-left baseline (and the unsupervised generative model as we discuss below), however it doesn't have any information about the annotation guidelines used for the testing corpus. For example, the Bulgarian corpus has an unusual treatment of nonfinite clauses. Figure 4 shows an example. We see that the "да" is the parent of both the verb and its object, which is different than the treatment in the English corpus.

We propose to deal with these annotation dissimilarities by creating very simple rules. For Spanish, we have three rules. The first rule sets main verbs to dominate auxiliary verbs. Specifically, whenever an auxiliary precedes a main verb the main verb becomes its parent and adopts its children; if there is only one main verb it becomes the root of the sentence; main verbs also become

Figure 4: An example where transfer fails because of different handling of reflexives and nonfinite clauses. The alignment links provide correct glosses for Bulgarian words. "Бях" is a past tense marker while "се" is a reflexive marker.

parents of pronouns, adverbs, and common nouns that directly preceed auxiliary verbs. By adopting children we mean that we change the parent of transferred edges to be the adopting node. The second Spanish rule states that the first element of an adjective-noun or noun-adjective pair dominates the second; the first element also adopts the children of the second element. The third and final Spanish rule sets all prepositions to be children of the first main verb in the sentence, unless the preposition is a "de" located between two noun phrases. In this later case, we set the closest noun in the first of the two noun phrases as the preposition's parent.

For Bulgarian the first rule is that "да" should dominate all words until the next verb and adopt their noun, preposition, particle and adverb children. The second rule is that auxiliary verbs should dominate main verbs and adopt their children. We have a list of 12 Bulgarian auxiliary verbs. The "seven rules" experiments add rules for 5 more words similar to the rule for "да", specifically "че", "ли", "какво", "не", "за". Table 3 compares the errors for different linguistic rules. When we train using the "да" rule and the rules for auxiliary verbs, the model learns that main verbs attach to auxiliary verbs and that "да" dominates its nonfinite clause. This causes an improvement in the attachment of verbs, and also drastically reduces words being attached to verbs instead of particles. The latter is expected because "да" is analyzed as a particle in the Bulgarian POS tagset. We see an improvement in root/verb confusions since "да" is sometimes errenously attached to a the following verb rather than being the root of the sentence.

The rightmost panel of Table 3 shows similar analysis when we also use the rules for the five other closed-class words. We see an improvement in attachments in all categories, but no qualitative change is visible. The reason for this is probably that these words are relatively rare, but by encouraging the model to add an edge, it also rules out incorrect edges that would cross it. Consequently we are seeing improvements not only directly from the constraints we enforce but also indirectly as types of edges that tend to get ruled out.

5.4 Generative parser

The generative model we use is a state of the art model for unsupervised parsing and is our only

No Rules				
child POS			parent POS	
	acc(%)	errors		errors
V	65.2	2237	T/V	2175
N	73.8	1938	V/V	1305
P	58.5	1705	N/V	1112
R	70.3	961	root/V	555

Two Rules				
child POS			parent POS	
	acc(%)	errors		errors
N	78.7	1572	N/V	938
P	70.2	1224	V/V	734
V	84.4	1002	V/N	529
R	79.3	670	N/N	376

Seven Rules				
child POS			parent POS	
	acc(%)	errors		errors
N	79.3	1532	N/V	1116
P	75.7	998	V/V	560
R	69.3	993	V/N	507
V	86.2	889	N/N	450

Table 3: Top 4 discriminative parser errors by child POS tag and true/guess parent POS tag in the Bulgarian CoNLL train data of length up to 10. Training with no language-specific rules (left); two rules (center); and seven rules (right). POS meanings: V verb, N noun, P pronoun, R preposition, T particle. Accuracies are by child or parent truth/guess POS tag.

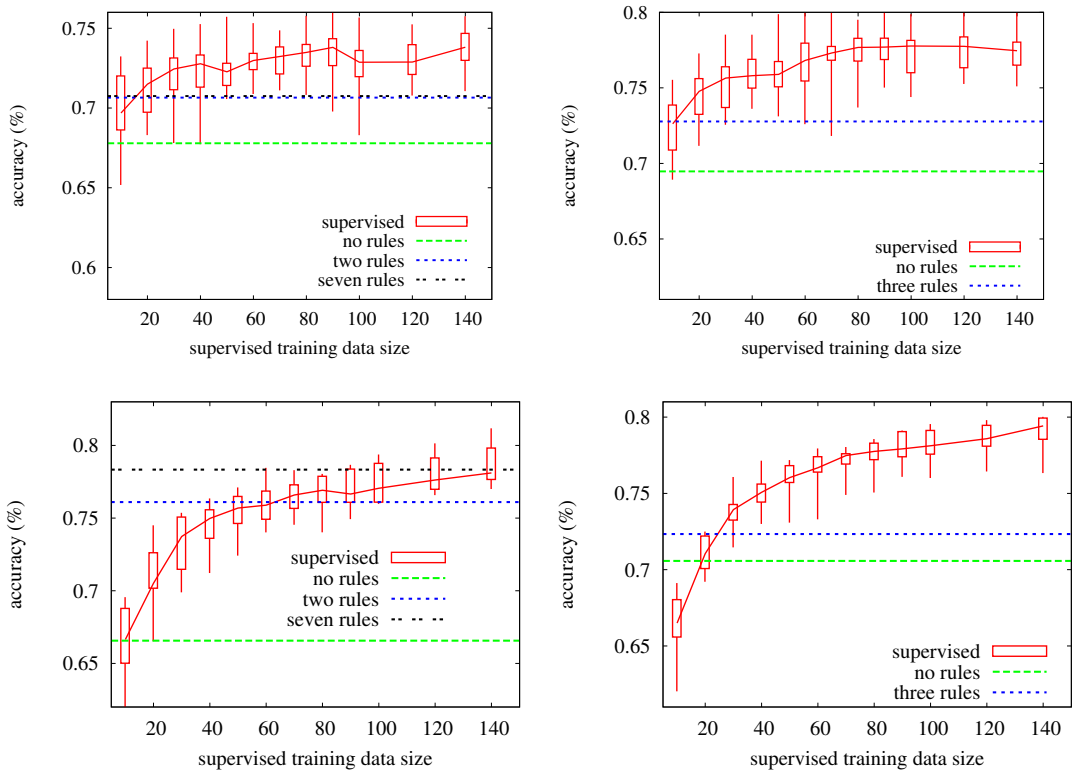

Figure 5: Comparison to parsers with supervised estimation and transfer. Top: Generative. Bottom: Discriminative. Left: Bulgarian. Right: Spanish. The transfer models were trained on 10k sentences all of length at most 20, all models tested on CoNLL train sentences of up to 10 words. The x-axis shows the number of examples used to train the supervised model. Boxes show first and third quartile, whiskers extend to max and min, with the line passing through the median. Supervised experiments used 30 random samples from CoNLL train.

fully unsupervised baseline. As smoothing we add a very small backoff probability of 4.5×10^{-5} to each learned paramter. Unfortunately, we found generative model performance was disappointing overall. The maximum unsupervised accuracy it achieved on the Bulgarian data is 47.6% with initialization from Klein and Manning (2004) and this result is not stable. Changing the initialization parameters, training sample, or maximum sentence length used for training drastically affected the results, even for samples with several thousand sentences. When we use the transferred information to constrain the learning, EM stabilizes and achieves much better performance. Even setting all parameters equal at the outset does not prevent the model from learning the dependency structure of the aligned language. The top panels in Figure 5 show the results in this setting. We see that performance is still always below the accuracy achieved by supervised training on 20 annotated sentences. However, the improvement in stability makes the algorithm much more usable. As we shall see below, the discriminative parser performs even better than the generative model.

5.5 Discriminative parser

We trained our discriminative parser for 100 iterations of online EM with a Gaussian prior variance of 100. Results for the discriminative parser are shown in the bottom panels of Figure 5. The

supervised experiments are given to provide context for the accuracies. For Bulgarian, we see that without any hints about the annotation guidelines, the transfer system performs better than an unsupervised parser, comparable to a supervised parser trained on 10 sentences. However, if we specify just the two rules for "да" and verb conjugations performance jumps to that of training on 60-70 fully labeled sentences. If we have just a little more prior knowledge about how closed-class words are handled, performance jumps above 140 fully labeled sentence equivalent.

We observed another desirable property of the discriminative model. While the generative model can get confused and perform poorly when the training data contains very long sentences, the discriminative parser does not appear to have this drawback. In fact we observed that as the maximum training sentence length increased, the parsing performance also improved.

6 Related Work

Our work most closely relates to Hwa et al. (2005), who proposed to learn generative dependency grammars using Collins' parser (Collins, 1999) by constructing full target parses via projected dependencies and completion/transformation rules. Hwa et al. (2005) found that transferring dependencies directly was not sufficient to get a parser with reasonable performance, even when both the source language parses and the word alignments are performed by hand. They adjusted for this by introducing on the order of one or two dozen language-specific transformation rules to complete target parses for unaligned words and to account for diverging annotation rules. Transferring from English to Spanish in this way, they achieve 72.1% and transferring to Chinese they achieve 53.9%.

Our learning method is very closely related to the work of (Mann and McCallum, 2007; Mann and McCallum, 2008) who concurrently developed the idea of using penalties based on posterior expectations of features not necessarily in the model in order to guide learning. They call their method generalized expectation constraints or alternatively expectation regularization. In this volume (Druck et al., 2009) use this framework to train a dependency parser based on constraints stated as corpus-wide expected values of linguistic rules. The rules select a class of edges (e.g. auxiliary verb to main verb) and require that the expectation of these be close to some value. The main difference between this work and theirs is the source of the information (a linguistic informant vs. cross-lingual projection). Also, we define our regularization with respect to inequality constraints (the model is not penalized for exceeding the required model expectations), while they require moments to be close to an estimated value. We suspect that the two learning methods could perform comparably when they exploit similar information.

7 Conclusion

In this paper, we proposed a novel and effective learning scheme for transferring dependency parses across bitext. By enforcing projected dependency constraints approximately and in expectation, our framework allows robust learning from noisy partially supervised target sentences, instead of committing to entire parses. We show that discriminative training generally outperforms generative approaches even in this very weakly supervised setting. By adding easily specified language-specific constraints, our models begin to rival strong supervised baselines for small amounts of data. Our framework can handle a wide range of constraints and we are currently exploring richer syntactic constraints that involve conservation of multiple edge constructions as well as constraints on conservation of surface length of dependencies.

Acknowledgments

This work was partially supported by an Integrative Graduate Education and Research Traineeship grant from National Science Foundation (NSF-IGERT 0504487) and by ARO MURI SUBTLE W911NF-07-1-0216.

References

A. Abeillé. 2003. *Treebanks: Building and Using Parsed Corpora*. Springer.

H. Alshawi, S. Bangalore, and S. Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1).

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proc. LREC*, Genoa, Italy.

P. F. Brown, S. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1994. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

C. Chelba, D. Engle, F. Jelinek, V. Jimenez, S. Khudanpur, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, A. Stolcke, and D. Wu. 1997. Structure and performance of a dependency language model. In *Proc. Eurospeech*.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

G. Druck, G. Mann, and A. McCallum. 2009. Semi-supervised learning of dependency parsers using generalized expectation criteria. In *Proc. ACL*.

J. Eisner. 1996. Three new probabilistic models for dependency parsing: an exploration. In *Proc. CoLing*.

H. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proc. EMNLP*, pages 304–311.

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

J. Graça, K. Ganchev, and B. Taskar. 2008. Expectation maximization and posterior constraints. In *Proc. NIPS*.

J. Graça, K. Ganchev, and B. Taskar. 2009. Postcat - posterior constrained alignment toolkit. In *The Third Machine Translation Marathon*.

A. Haghighi, A. Ng, and C. Manning. 2005. Robust textual inference via graph matching. In *Proc. EMNLP*.

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

D. Klein and C. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. of ACL*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

S. Lee and K. Choi. 1997. Reestimation and best-first parsing algorithm for probabilistic dependency grammar. In *In WVLC-5*, pages 41–55.

G. Mann and A. McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proc. ICML*.

G. Mann and A. McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proc. ACL*, pages 870 – 878.

R. McDonald, K. Crammer, and F. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. ACL*, pages 91–98.

I. Mel'čuk. 1988. *Dependency syntax: theory and practice*. SUNY. inci.

P. Merlo, S. Stevenson, V. Tsang, and G. Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proc. ACL*.

R. M. Neal and G. E. Hinton. 1998. A new view of the EM algorithm that justifies incremental, sparse and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer.

J. Nivre and J. Nilsson. 2005. Pseudo-projective dependency parsing. In *Proc. ACL*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. EMNLP-CoNLL*.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proc. ACL*.

C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proc. ACL*.

L. Shen, J. Xu, and R. Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proc. of ACL*.

N. Smith and J. Eisner. 2006. Annealing structural bias in multilingual weighted grammar induction. In *Proc. ACL*.

J. Tiedemann. 2007. Building a multilingual parallel subtitle corpus. In *Proc. CLIN*.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. HLT-NAACL*.

Y. Tsuruoka and J. Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proc. HLT/EMNLP*.

H. Yamada and Y. Matsumoto. 2003a. Statistical dependency analysis with support vector machines. In *Proc. IWPT*, pages 195–206.

H. Yamada and Y. Matsumoto. 2003b. Statistical dependency analysis with support vector machines. In *Proc. IWPT*.

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

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. HLT*.