Phylogenetic Grammar Induction

Taylor Berg-Kirkpatrick and **Dan Klein**
Computer Science Division
University of California, Berkeley
{tberg, klein}@cs.berkeley.edu

Abstract

We present an approach to multilingual grammar induction that exploits a phylogeny-structured model of parameter drift. Our method does not require any translated texts or token-level alignments. Instead, the phylogenetic prior couples languages at a parameter level. Joint induction in the multilingual model substantially outperforms independent learning, with larger gains both from more articulated phylogenies and as well as from increasing numbers of languages. Across eight languages, the multilingual approach gives error reductions over the standard monolingual DMV averaging 21.1% and reaching as high as 39%.

1 Introduction

Learning multiple languages together should be easier than learning them separately. For example, in the domain of syntactic parsing, a range of recent work has exploited the mutual constraint between two languages' parses of the same bitext (Kuhn, 2004; Burkett and Klein, 2008; Kuzman et al., 2009; Smith and Eisner, 2009; Snyder et al., 2009a). Moreover, Snyder et al. (2009b) in the context of unsupervised part-of-speech induction (and Bouchard-Côté et al. (2007) in the context of phonology) show that extending beyond two languages can provide increasing benefit. However, multitexts are only available for limited languages and domains. In this work, we consider unsupervised grammar induction without bitexts or multitexts. Without translation examples, multilingual constraints cannot be exploited at the sentence token level. Rather, we capture multilingual constraints at a *parameter* level, using a phylogeny-structured prior to tie together the various individual languages' learning problems. Our joint, hierarchical prior couples model parameters for different languages in a way that respects knowledge about how the languages evolved.

Aspects of this work are closely related to Cohen and Smith (2009) and Bouchard-Côté et al. (2007). Cohen and Smith (2009) present a model for jointly learning English and Chinese dependency grammars without bitexts. In their work, structurally constrained covariance in a logistic normal prior is used to couple parameters between the two languages. Our work, though also different in technical approach, differs most centrally in the extension to multiple languages and the use of a phylogeny. Bouchard-Côté et al. (2007) considers an entirely different problem, phonological reconstruction, but shares with this work both the use of a phylogenetic structure as well as the use of log-linear parameterization of local model components. Our work differs from theirs primarily in the task (syntax vs. phonology) and the variables governed by the phylogeny: in our model it is the grammar parameters that drift (in the prior) rather than individual word forms (in the likelihood model).

Specifically, we consider dependency induction in the DMV model of Klein and Manning (2004). Our data is a collection of standard dependency data sets in eight languages: English, Dutch, Danish, Swedish, Spanish, Portuguese, Slovene, and Chinese. Our focus is not the DMV model itself, which is well-studied, but rather the prior which couples the various languages' parameters. While some choices of prior structure can greatly complicate inference (Cohen and Smith, 2009), we choose a hierarchical Gaussian form for the drift term, which allows the gradient of the observed data likelihood to be easily computed using standard dynamic programming methods.

In our experiments, joint multilingual learning substantially outperforms independent monolingual learning. Using a limited phylogeny that only couples languages within linguistic families reduces error by 5.6% over the monolingual baseline. Using a flat, global phylogeny gives a greater reduction, almost 10%. Finally, a more articulated phylogeny that captures both inter- and intra-family effects gives an even larger average relative error reduction of 21.1%.

2 Model

We define our model over two kinds of random variables: dependency trees and parameters. For each language ℓ in a set L, our model will generate a collection $\mathbf{t}_\ell$ of dependency trees $\mathbf{t}_\ell^i$. We assume that these dependency trees are generated by the DMV model of Klein and Manning (2004), which we write as $\mathbf{t}_\ell^i \sim \text{DMV}(\theta_\ell)$. Here, θ_ℓ is a vector of the various model parameters for language ℓ. The prior is what couples the θ_ℓ parameter vectors across languages; it is the focus of this work. We first consider the likelihood model before moving on to the prior.

2.1 Dependency Model with Valence

A dependency parse is a directed tree $\mathbf{t}$ over tokens in a sentence $\mathbf{s}$. Each edge of the tree specifies a directed dependency from a head token to a dependent, or argument token. The DMV is a generative model for trees $\mathbf{t}$, which has been widely used for dependency parse induction. The observed data likelihood, used for parameter estimation, is the marginal probability of generating the observed sentences $\mathbf{s}$, which are simply the leaves of the trees $\mathbf{t}$. Generation in the DMV model involves two types of local conditional probabilities: CONTINUE distributions that capture valence and ATTACH distributions that capture argument selection.

First, the Bernoulli CONTINUE probability distributions $P^{\text{CONTINUE}}(c|h, dir, adj; \theta_\ell)$ model the fertility of a particular head type h. The outcome $c \in \{stop, continue\}$ is conditioned on the head type h, direction dir, and adjacency adj. If a head type's continue probability is low, tokens of this type will tend to generate few arguments.

Second, the ATTACH multinomial probability distributions $P^{\text{ATTACH}}(a|h, dir; \theta_\ell)$ capture attachment preferences of heads, where a and h are both token types. We take the same approach as previous work (Klein and Manning, 2004; Cohen and Smith, 2009) and use gold part-of-speech labels as tokens. Thus, the basic observed "word" types are actually word classes.

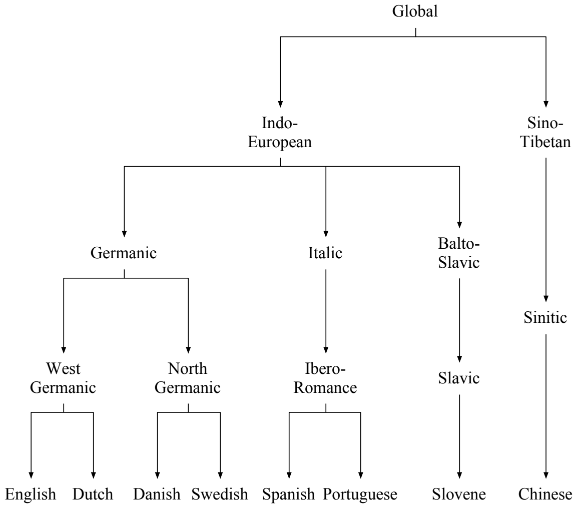

Figure 1: An example of a linguistically-plausible phylogenetic tree over the languages in our training data. Leaves correspond to (observed) modern languages, while internal nodes represent (unobserved) ancestral languages.

2.1.1 Log-Linear Parameterization

The DMV's local conditional distributions were originally given as simple multinomial distributions with one parameter per outcome. However, they can be re-parameterized to give the following log-linear form (Eisner, 2002; Bouchard-Côté et al., 2007; Berg-Kirkpatrick et al., 2010):

$$P^{\text{CONTINUE}}(c|h, dir, adj; \theta_\ell) = \frac{\exp\left[\theta_\ell{}^T \boldsymbol{f}_{\text{CONTINUE}}(c, h, dir, adj)\right]}{\sum_{c'} \exp\left[\theta_\ell{}^T \boldsymbol{f}_{\text{CONTINUE}}(c', h, dir, adj)\right]}$$

$$P^{\text{ATTACH}}(a|h, dir; \theta_\ell) = \frac{\exp\left[\theta_\ell{}^T \boldsymbol{f}_{\text{ATTACH}}(a, h, dir)\right]}{\sum_{a'} \exp\left[\theta_\ell{}^T \boldsymbol{f}_{\text{ATTACH}}(a', h, dir)\right]}$$

The parameters are weights θ_ℓ with one weight vector per language. In the case where the vector of feature functions $\boldsymbol{f}$ has an indicator for each possible conjunction of outcome and conditions, the original multinomial distributions are recovered. We refer to these full indicator features as the set of SPECIFIC features.

2.2 Phylogenetic Prior

The focus of this work is coupling each of the parameters θ_ℓ in a phylogeny-structured prior. Consider a phylogeny like the one shown in Figure 1, where each modern language ℓ in L is a leaf. We would like to say that the leaves' parameter vectors arise from a process which slowly drifts along each branch. A convenient choice is to posit additional parameter variables $\theta_{\ell+}$ at internal nodes $\ell^+ \in L^+$, a set of ancestral languages, and to assume that the conditional distribution $P(\theta_\ell | \theta_{\mathrm{par}(\ell)})$ at each branch in the phylogeny is a Gaussian centered on $\theta_{\mathrm{par}(\ell)}$, where $\mathrm{par}(\ell)$ is the parent of ℓ in the phylogeny and ℓ ranges over $L \cup L^+$. The variance structure of the Gaussian would then determine how much drift (and in what directions) is expected. Concretely, we assume that each drift distribution is an isotropic Gaussian with mean $\theta_{\mathrm{par}(\ell)}$ and scalar variance σ^2. The root is centered at zero. We have thus defined a joint distribution $P(\Theta|\sigma^2)$ where $\Theta = (\theta_\ell : \ell \in L \cup L^+)$. σ^2 is a hyperparameter for this prior which could itself be re-parameterized to depend on branch length or be learned; we simply set it to a plausible constant value.

Two primary challenges remain. First, inference under arbitrary priors can become complex. However, in the simple case of our diagonal covariance Gaussians, the gradient of the observed data likelihood can be computed directly using the DMV's expected counts and maximum-likelihood estimation can be accomplished by applying standard gradient optimization methods. Second, while the choice of diagonal covariance is efficient, it causes components of θ that correspond to features occurring in only one language to be marginally independent of the parameters of all other languages. In other words, only features which fire in more than one language are coupled by the prior. In the next section, we therefore increase the overlap between languages' features by using coarse projections of parts-of-speech.

2.3 Projected Features

With diagonal covariance in the Gaussian drift terms, each parameter evolves independently of the others. Therefore, our prior will be most informative when features activate in multiple languages. In phonology, it is useful to map phonemes to the International Phonetic Alphabet (IPA) in order to have a language-independent parameterization. We introduce a similarly neutral representation here by projecting language-specific parts-of-speech to a coarse, shared inventory.

Indeed, we assume that each language has a distinct tagset, and so the basic configurational features will be language specific. For example, when an English VBZ takes a left argument headed by a NNS, a feature will activate specific to VBZ-NNS-LEFT. That feature will be used in the log-linear attachment probability for English. However, because that feature does not show up in any other language, it is not usefully controlled by the prior. Therefore, we also include coarser features which activate on more abstract, cross-linguistic configurations. In the same example, a feature will fire indicating a coarse, direction-free NOUN-VERB attachment. This feature will now occur in multiple languages and will contribute to each of those languages' attachment models. Although such cross-lingual features will have different weight parameters in each language, those weights will covary, being correlated by the prior.

SPECIFIC: Activate for only one conjunction of outcome and conditions:
$\mathbb{1}(c = \cdot, h = \cdot, dir = \cdot, adj = \cdot)$
SHARED: Activate for heads from multiple languages using cross-lingual POS projection $\pi(\cdot)$:
$\mathbb{1}(c = \cdot, \pi(h) = \cdot, dir = \cdot, adj = \cdot)$

CONTINUE distribution feature templates.

SPECIFIC: Activate for only one conjunction of outcome and conditions:
$\mathbb{1}(a = \cdot, h = \cdot, dir = \cdot)$
SHARED: Activate for heads and arguments from multiple languages using cross-lingual POS projection $\pi(\cdot)$:
$\mathbb{1}(\pi(a) = \cdot, \pi(h) = \cdot, dir = \cdot)$
$\mathbb{1}(\pi(a) = \cdot, h = \cdot, dir = \cdot)$
$\mathbb{1}(a = \cdot, \pi(h) = \cdot, dir = \cdot)$

ATTACH distribution feature templates.

Table 1: Feature templates for CONTINUE and ATTACH conditional distributions.

The coarse features are defined via a projection π from language-specific part-of-speech labels to coarser, cross-lingual word classes, and hence we refer to them as SHARED features. For each corpus used in this paper, we use the tagging annotation guidelines to manually define a fixed mapping from the corpus tagset to the following coarse tagset: noun, verb, adjective, adverb, conjunction, preposition, determiner, interjection, numeral, and pronoun. Parts-of-speech for which this coarse mapping is ambiguous or impossible are not mapped, and do not have corresponding SHARED features.

We summarize the feature templates for the CONTINUE and ATTACH conditional distributions in Table 1. Variants of all feature templates that ignore direction and/or adjacency are included. In practice, we found it beneficial for all language-independent features to ignore direction.

Again, only the coarse features occur in multiple languages, so all phylogenetic influence is through those. Nonetheless, the effect of the phylogeny turns out to be quite strong.

2.4 Learning

We now turn to learning with the phylogenetic prior. Since the prior couples parameters across languages, this learning problem requires parameters for all languages be estimated jointly. We seek to find $\Theta = (\theta_\ell : \ell \in L \cup L^+)$ which optimizes $\log P(\Theta|\mathbf{s})$, where $\mathbf{s}$ aggregates the observed leaves of all the dependency trees in all the languages. This can be written as

$$\log P(\Theta) + \log P(\mathbf{s}|\Theta) - \log P(\mathbf{s})$$

The third term is a constant and can be ignored. The first term can be written as

$$\log P(\Theta) = \sum_{\ell \in L \cup L^+} \frac{1}{2\sigma^2} \|\theta_\ell - \theta_{\mathrm{par}(\ell)}\|_2^2 + C$$

where C is a constant. The form of $\log P(\Theta)$ immediately shows how parameters are penalized for being different across languages, more so for languages that are near each other in the phylogeny. The second term

$$\log P(\mathbf{s}|\Theta) = \sum_{\ell \in L} \log P(\mathbf{s}_\ell|\theta_\ell)$$

is a sum of observed data likelihoods under the standard DMV models for each language, computable by dynamic programming (Klein and Manning, 2004). Together, this yields the following objective function:

$$l(\Theta) = \textstyle\sum_{\ell \in L \cup L^+} \frac{1}{2\sigma^2} \|\theta_\ell - \theta_{\mathrm{par}(\ell)}\|_2^2 + \sum_{\ell \in L} \log P(\mathbf{s}_\ell|\theta_\ell)$$

which can be optimized using gradient methods or (MAP) EM. Here we used L-BFGS (Liu et al., 1989). This requires computation of the gradient of the observed data likelihood $\log P(\mathbf{s}_\ell|\theta_\ell)$ which is given by:

$$\nabla \log P(\mathbf{s}_\ell|\theta_\ell) = \mathbb{E}_{\mathbf{t}_\ell|\mathbf{s}_\ell}\big[\nabla \log P(\mathbf{s}_\ell, \mathbf{t}_\ell|\theta_\ell)\big] =$$

$$\left[\begin{array}{l} \sum_{c,h,dir,adj} e_{c,h,dir,adj}(\mathbf{s}_\ell;\theta_\ell) \cdot \Big[\boldsymbol{f}_{\text{CONTINUE}}(c,h,dir,adj) - \\ \qquad \sum_{c'} P^{\text{CONTINUE}}(c'|h,dir,adj;\theta_\ell)\boldsymbol{f}_{\text{CONTINUE}}(c',h,dir,adj)\Big] \\ \\ \sum_{a,h,dir} e_{a,h,dir}(\mathbf{s}_\ell;\theta_\ell) \cdot \Big[\boldsymbol{f}_{\text{ATTACH}}(a,h,dir) - \\ \qquad \sum_{a'} P^{\text{ATTACH}}(a'|h,dir;\theta_\ell)\boldsymbol{f}_{\text{ATTACH}}(a',h,dir)\Big] \end{array}\right]$$

The expected gradient of the log joint likelihood of sentences and parses is equal to the gradient of the log marginal likelihood of just sentences, or the observed data likelihood (Salakhutdinov et al., 2003). $e_{a,h,dir}(\mathbf{s}_\ell;\theta_\ell)$ is the expected count of the number of times head h is attached to a in direction dir given the observed sentences $\mathbf{s}_\ell$ and DMV parameters θ_ℓ. $e_{c,h,dir,adj}(\mathbf{s}_\ell;\theta_\ell)$ is defined similarly. Note that these are the same expected counts required to perform EM on the DMV, and are computable by dynamic programming.

The computation time is dominated by the computation of each sentence's posterior expected counts, which are independent given the parameters, so the time required per iteration is essentially the same whether training all languages jointly or independently. In practice, the total number of iterations was also similar.

3 Experimental Setup

3.1 Data

We ran experiments with the following languages: English, Dutch, Danish, Swedish, Spanish, Portuguese, Slovene, and Chinese. For all languages but English and Chinese, we used corpora from the 2006 CoNLL-X Shared Task dependency parsing data set (Buchholz and Marsi, 2006). We used the shared task training set to both train and test our models. These corpora provide hand-labeled part-of-speech tags (except for Dutch, which is automatically tagged) and provide dependency parses, which are either themselves hand-labeled or have been converted from hand-labeled parses of other kinds. For English and Chinese we use sections 2-21 of the Penn Treebank (PTB) (Marcus et al., 1993) and sections 1-270 of the Chinese Treebank (CTB) (Xue et al., 2002) respectively. Similarly, these sections were used for both training and testing. The English and Chinese data sets have hand-labeled constituency parses and part-of-speech tags, but no dependency parses. We used the Bikel Chinese head finder (Bikel and Chiang, 2000) and the Collins English head finder (Collins, 1999) to transform the gold constituency parses into gold dependency parses. None of the corpora are bitexts. For all languages, we ran experiments on all sentences of length 10 or less after punctuation has been removed.

When constructing phylogenies over the languages we made use of their linguistic classifications. English and Dutch are part of the West Ger-

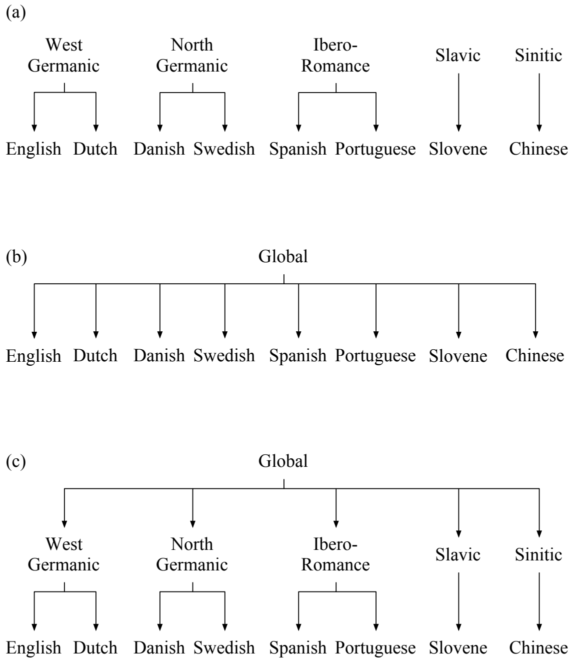

Figure 2: (a) Phylogeny for FAMILIES model. (b) Phylogeny for GLOBAL model. (c) Phylogeny for LINGUISTIC model.

manic family of languages, whereas Danish and Swedish are part of the North Germanic family. Spanish and Portuguese are both part of the Ibero-Romance family. Slovene is part of the Slavic family. Finally, Chinese is in the Sinitic family, and is not an Indo-European language like the others. We interchangeably speak of a language family and the ancestral node corresponding to that family's root language in a phylogeny.

3.2 Models Compared

We evaluated three phylogenetic priors, each with a different phylogenetic structure. We compare with two monolingual baselines, as well as an all-pairs multilingual model that does not have a phylogenetic interpretation, but which provides very similar capacity for parameter coupling.

3.2.1 Phylogenetic Models

The first phylogenetic model uses the shallow phylogeny shown in Figure 2(a), in which only languages within the same family have a shared parent node. We refer to this structure as FAMILIES. Under this prior, the learning task decouples into independent subtasks for each family, but no regularities across families can be captured.

The family-level model misses the constraints between distant languages. Figure 2(b) shows another simple configuration, wherein all languages share a common parent node in the prior, meaning that global regularities that are consistent across all languages can be captured. We refer to this structure as GLOBAL.

While the global model couples the parameters for all eight languages, it does so without sensitivity to the articulated structure of their descent. Figure 2(c) shows a more nuanced prior structure, LINGUISTIC, which groups languages first by family and then under a global node. This structure allows global regularities as well as regularities within families to be learned.

3.2.2 Parameterization and ALLPAIRS Model

Daumé III (2007) and Finkel and Manning (2009) consider a formally similar Gaussian hierarchy for domain adaptation. As pointed out in Finkel and Manning (2009), there is a simple equivalence between hierarchical regularization as described here and the addition of new tied features in a "flat" model with zero-meaned Gaussian regularization on all parameters. In particular, instead of parameterizing the objective in Section 2.4 in terms of multiple sets of weights, one at each node in the phylogeny (the *hierarchical parameterization*, described in Section 2.4), it is equivalent to parameterize this same objective in terms of a single set of weights on a larger of group features (the *flat parameterization*). This larger group of features contains a duplicate set of the features discussed in Section 2.3 for each node in the phylogeny, each of which is active only on the languages that are its descendants. A linear transformation between parameterizations gives equivalence. See Finkel and Manning (2009) for details.

In the flat parameterization, it seems equally reasonable to simply tie all pairs of languages by adding duplicate sets of features for each pair. This gives the ALLPAIRS setting, which we also compare to the tree-structured phylogenetic models above.

3.3 Baselines

To evaluate the impact of multilingual constraint, we compared against two monolingual baselines. The first baseline is the standard DMV with only SPECIFIC features, which yields the standard multinomial DMV (*weak baseline*). To facilitate comparison to past work, we used no prior for this monolingual model. The second baseline is the DMV with added SHARED features. This model includes a simple isotropic Gaussian prior on pa-

			Monolingual		**Multilingual**				
						Phylogenetic			
		Corpus Size	Baseline	Baseline w/ Shared	AllPairs	Families	BestPair	Global	Linguistic
West Germanic	English	6008	47.1	51.3	48.5	51.3	51.3 (Ch)	51.2	**62.3**
	Dutch	6678	36.3	36.0	44.0	36.1	36.2 (Sw)	44.0	**45.1**
North Germanic	Danish	1870	33.5	33.6	40.5	31.4	34.2 (Du)	39.6	**41.6**
	Swedish	3571	45.3	44.8	56.3	44.8	44.8 (Ch)	44.5	**58.3**
Ibero-Romance	Spanish	712	28.0	40.5	58.7	63.4	**63.8** (Da)	59.4	58.4
	Portuguese	2515	38.5	38.5	**63.1**	37.4	38.4 (Sw)	37.4	63.0
Slavic	Slovene	627	38.5	39.7	49.0	–	**49.6** (En)	49.4	48.4
Sinitic	Chinese	959	36.3	43.3	**50.7**	–	49.7 (Sw)	50.1	49.6
	Macro-Avg. Relative Error Reduction				17.1	5.6	8.5	9.9	**21.1**

Table 2: Directed dependency accuracy of monolingual and multilingual models, and relative error reduction over the monolingual baseline with SHARED features macro-averaged over languages. Multilingual models outperformed monolingual models in general, with larger gains from increasing numbers of languages. Additionally, more nuanced phylogenetic structures outperformed cruder ones.

rameters. This second baseline is the more direct comparison to the multilingual experiments here (*strong baseline*).

3.4 Evaluation

For each setting, we evaluated the directed dependency accuracy of the minimum Bayes risk (MBR) dependency parses produced by our models under maximum (posterior) likelihood parameter estimates. We computed accuracies separately for each language in each condition. In addition, for multilingual models, we computed the relative error reduction over the strong monolingual baseline, macro-averaged over languages.

3.5 Training

Our implementation used the flat parameterization described in Section 3.2.2 for both the phylogenetic and ALLPAIRS models. We originally did this in order to facilitate comparison with the non-phylogenetic ALLPAIRS model, which has no equivalent hierarchical parameterization. In practice, optimizing with the hierarchical parameterization also seemed to underperform.[1]

All models were trained by directly optimizing the observed data likelihood using L-BFGS (Liu et al., 1989). Berg-Kirkpatrick et al. (2010) suggest that directly optimizing the observed data likelihood may offer improvements over the more standard expectation-maximization (EM) optimization procedure for models such as the DMV, especially when the model is parameterized using features. We stopped training after 200 iterations in all cases. This fixed stopping criterion seemed to be adequate in all experiments, but presumably there is a potential gain to be had in fine tuning. To initialize, we used the harmonic initializer presented in Klein and Manning (2004). This type of initialization is deterministic, and thus we did not perform random restarts.

We found that for all models $\sigma^2 = 0.2$ gave reasonable results, and we used this setting in all experiments. For most models, we found that varying σ^2 in a reasonable range did not substantially affect accuracy. For some models, the directed accuracy was less flat with respect to σ^2. In these less-stable cases, there seemed to be an interaction between the variance and the choice between head conventions. For example, for some settings of σ^2, but not others, the model would learn that determiners head noun phrases. In particular, we observed that even when direct accuracy did fluctuate, undirected accuracy remained more stable.

[1] We noticed that the weights of features shared across languages had larger magnitude early in the optimization procedure when using the flat parameterization compared to using the hierarchical parameterization, perhaps indicating that cross-lingual influences had a larger effect on learning in its initial stages.

4 Results

Table 2 shows the overall results. In all cases, methods which coupled the languages in some way outperformed the independent baselines that considered each language independently.

4.1 Bilingual Models

The weakest of the coupled models was FAMILIES, which had an average relative error reduction of 5.6% over the strong baseline. In this case, most of the average improvement came from a single family: Spanish and Portuguese. The limited improvement of the family-level prior compared to other phylogenies suggests that there are important multilingual interactions that do not happen within families. Table 2 also reports the maximum accuracy achieved for each language when it was paired with another language (same family or otherwise) and trained together with a single common parent. These results appear in the column headed by BESTPAIR, and show the best accuracy for the language on that row over all possible pairings with other languages. When pairs of languages were trained together in isolation, the largest benefit was seen for languages with small training corpora, not necessarily languages with common ancestry. In our setup, Spanish, Slovene, and Chinese have substantially smaller training corpora than the rest of the languages considered. Otherwise, the patterns are not particularly clear; combined with subsequent results, it seems that pairwise constraint is fairly limited.

4.2 Multilingual Models

Models that coupled multiple languages performed better in general than models that only considered pairs of languages. The GLOBAL model, which couples all languages, if crudely, yielded an average relative error reduction of 9.9%. This improvement comes as the number of languages able to exert mutual constraint increases. For example, Dutch and Danish had large improvements, over and above any improvements these two languages gained when trained with a single additional language. Beyond the simplistic GLOBAL phylogeny, the more nuanced LINGUISTIC model gave large improvements for English, Swedish, and Portuguese. Indeed, the LINGUISTIC model is the only model we evaluated that gave improvements for *all* the languages we considered.

It is reasonable to worry that the improvements from these multilingual models might be partially due to having more total training data in the multilingual setting. However, we found that halving the amount of data used to train the English, Dutch, and Swedish (the languages with the most training data) monolingual models did not substantially affect their performance, suggesting that for languages with several thousand sentences or more, the increase in statistical support due to additional monolingual data was not an important effect (the DMV is a relatively low-capacity model in any case).

4.3 Comparison of Phylogenies

Recall the structures of the three phylogenies presented in Figure 2. These phylogenies differ in the correlations they can represent. The GLOBAL phylogeny captures only “universals,” while FAMILIES captures only correlations between languages that are known to be similar. The LINGUISTIC model captures both of these effects simultaneously by using a two layer hierarchy. Notably, the improvement due to the LINGUISTIC model is more than the sum of the improvements due to the GLOBAL and FAMILIES models.

4.4 Phylogenetic vs. ALLPAIRS

The phylogeny is capable of allowing appropriate influence to pass between languages at multiple levels. We compare these results to the ALLPAIRS model in order to see whether limitation to a tree structure is helpful. The ALLPAIRS model achieved an average relative error reduction of 17.1%, certainly outperforming both the simple phylogenetic models. However, the rich phylogeny of the LINGUISTIC model, which incorporates linguistic constraints, outperformed the freer ALLPAIRS model. A large portion of this improvement came from English, a language for which the LINGUISTIC model greatly outperformed all other models evaluated. We found that the improved English analyses produced by the LINGUISTIC model were more consistent with this model’s analyses of other languages. This consistency was not present for the English analyses produced by other models. We explore consistency in more detail in Section 5.

4.5 Comparison to Related Work

The likelihood models for both the strong monolingual baseline and the various multilingual models are the same, both expanding upon the standard DMV by adding coarse SHARED features. These coarse features, even in a monolingual setting, improved performance slightly over the weak baseline, perhaps by encouraging consistent treatment of the different finer-grained variants of parts-of-speech (Berg-Kirkpatrick et al., 2010).[2] The only difference between the multilingual systems and the strong baseline is whether or not cross-language influence is allowed through the prior.

While this progression of model structure is similar to that explored in Cohen and Smith (2009), Cohen and Smith saw their largest improvements from tying together parameters for the varieties of coarse parts-of-speech monolinugally, and then only moderate improvements from allowing cross-linguistic influence on top of monolingual sharing. When Cohen and Smith compared their best shared logistic-normal bilingual models to monolingual counter-parts for the languages they investigate (Chinese and English), they reported a relative error reduction of 5.3%. In comparison, with the LINGUISTIC model, we saw a much larger 16.9% relative error reduction over our strong baseline for these languages. Evaluating our LINGUISTIC model on the same test sets as (Cohen and Smith, 2009), sentences of length 10 or less in section 23 of PTB and sections 271-300 of CTB, we achieved an accuracy of 56.6 for Chinese and 60.3 for English. The best models of Cohen and Smith (2009) achieved accuracies of 52.0 and 62.0 respectively on these same test sets.

Our results indicate that the majority of our model's power beyond that of the standard DMV is derived from multilingual, and in particular, more-than-bilingual, interaction. These are, to the best of our knowledge, the first results of this kind for grammar induction without bitext.

5 Analysis

By examining the proposed parses we found that the LINGUISTIC and ALLPAIRS models produced analyses that were more consistent across languages than those of the other models. We also observed that the most common errors can be summarized succinctly by looking at attachment counts between coarse parts-of-speech. Figure 3 shows matrix representations of dependency counts. The area of a square is proportional to the number of order-collapsed dependencies where the column label is the head and the row label is the argument in the parses from each system. For ease of comprehension, we use the cross-lingual projections and only show counts for selected interesting classes.

Comparing Figure 3(c), which shows dependency counts proposed by the LINGUISTIC model, to Figure 3(a), which shows the same for the strong monolingual baseline, suggests that the analyses proposed by the LINGUISTIC model are more consistent across languages than are the analyses proposed by the monolingual model. For example, the monolingual learners are divided as to whether determiners or nouns head noun phrases. There is also confusion about which labels head whole sentences. Dutch has the problem that verbs modify pronouns more often than pronouns modify verbs, and pronouns are predicted to head sentences as often as verbs are. Spanish has some confusion about conjunctions, hypothesizing that verbs often attach to conjunctions, and conjunctions frequently head sentences. More subtly, the monolingual analyses are inconsistent in the way they head prepositional phrases. In the monolingual Portuguese hypotheses, prepositions modify nouns more often than nouns modify prepositions. In English, nouns modify prepositions, and prepositions modify verbs. Both the Dutch and Spanish models are ambivalent about the attachment of prepositions.

As has often been observed in other contexts (Liang et al., 2008), promoting agreement can improve accuracy in unsupervised learning. Not only are the analyses proposed by the LINGUISTIC model more consistent, they are also more in accordance with the gold analyses. Under the LINGUISTIC model, Dutch now attaches pronouns to verbs, and thus looks more like English, its sister in the phylogenetic tree. The LINGUISTIC model has also chosen consistent analyses for prepositional phrases and noun phrases, calling prepositions and nouns the heads of each, respectively. The problem of conjunctions heading Spanish sentences has also been corrected.

Figure 3(b) shows dependency counts for the GLOBAL multilingual model. Unsurprisingly, the analyses proposed under global constraint appear somewhat more consistent than those proposed under no multi-lingual constraint (now three lan-

[2] Coarse features that only tie nouns and verbs are explored in Berg-Kirkpatrick et al. (2010). We found that these were very effective for English and Chinese, but gave worse performance for other languages.

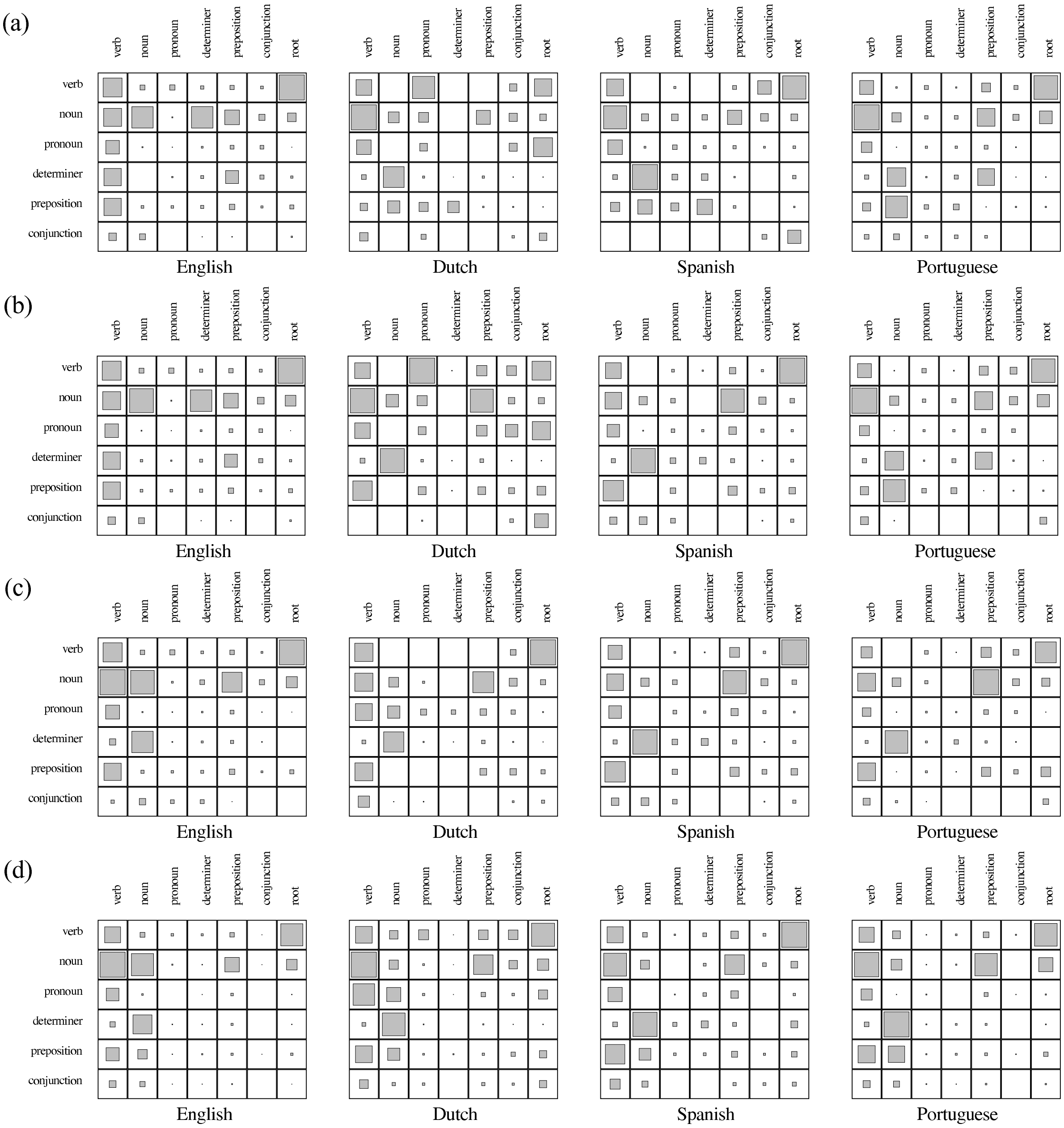

Figure 3: Dependency counts in proposed parses. Row label modifies column label. (a) Monolingual baseline with SHARED features. (b) GLOBAL model. (c) LINGUISTIC model. (d) Dependency counts in hand-labeled parses. Analyses proposed by monolingual baseline show significant inconsistencies across languages. Analyses proposed by LINGUISTIC model are more consistent across languages than those proposed by either the monolingual baseline or the GLOBAL model.

guages agree that prepositional phrases are headed by prepositions), but not as consistent as those proposed by the LINGUISTIC model.

Finally, Figure 3(d) shows dependency counts in the hand-labeled dependency parses. It appears that even the very consistent LINGUISTIC parses do not capture the non-determinism of prepositional phrase attachment to both nouns and verbs.

6 Conclusion

Even without translated texts, multilingual constraints expressed in the form of a phylogenetic prior on parameters can give substantial gains in grammar induction accuracy over treating languages in isolation. Additionally, articulated phylogenies that are sensitive to evolutionary structure can outperform not only limited flatter priors but also unconstrained all-pairs interactions.

7 Acknowledgements

This project is funded in part by the NSF under grant 0915265 and DARPA under grant N10AP20007.

References

T. Berg-Kirkpatrick, A. Bouchard-Côté, J. DeNero, and D. Klein. 2010. Painless unsupervised learning with features. In *North American Chapter of the Association for Computational Linguistics*.

D. M. Bikel and D. Chiang. 2000. Two statistical parsing models applied to the Chinese treebank. In *Second Chinese Language Processing Workshop*.

A. Bouchard-Côté, P. Liang, D. Klein, and T. L. Griffiths. 2007. A probabilistic approach to diachronic phonology. In *Empirical Methods in Natural Language Processing*.

S. Buchholz and E. Marsi. 2006. Computational Natural Language Learning-X shared task on multilingual dependency parsing. In *Conference on Computational Natural Language Learning*.

D. Burkett and D. Klein. 2008. Two languages are better than one (for syntactic parsing). In *Empirical Methods in Natural Language Processing*.

S. B. Cohen and N. A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *North American Chapter of the Association for Computational Linguistics*.

M. Collins. 1999. Head-driven statistical models for natural language parsing. In *Ph.D. thesis, University of Pennsylvania, Philadelphia*.

H. Daumé III. 2007. Frustratingly easy domain adaptation. In *Association for Computational Linguistics*.

J. Eisner. 2002. Parameter estimation for probabilistic finite-state transducers. In *Association for Computational Linguistics*.

J. R. Finkel and C. D. Manning. 2009. Hierarchical bayesian domain adaptation. In *North American Chapter of the Association for Computational Linguistics*.

D. Klein and C. D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Association for Computational Linguistics*.

J. Kuhn. 2004. Experiments in parallel-text based grammar induction. In *Association for Computational Linguistics*.

G. Kuzman, J. Gillenwater, and B. Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Association for Computational Linguistics/International Joint Conference on Natural Language Processing*.

P. Liang, D. Klein, and M. I. Jordan. 2008. Agreement-based learning. In *Advances in Neural Information Processing Systems*.

D. C. Liu, J. Nocedal, and C. Dong. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*.

R. Salakhutdinov, S. Roweis, and Z. Ghahramani. 2003. Optimization with EM and expectation-conjugate-gradient. In *International Conference on Machine Learning*.

D. A. Smith and J. Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Empirical Methods in Natural Language Processing*.

B. Snyder, T. Naseem, and R. Barzilay. 2009a. Unsupervised multilingual grammar induction. In *Association for Computational Linguistics/International Joint Conference on Natural Language Processing*.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2009b. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *North American Chapter of the Association for Computational Linguistics*.

N. Xue, F-D Chiou, and M. Palmer. 2002. Building a large-scale annotated Chinese corpus. In *International Conference on Computational Linguistics*.