Selective Sharing for Multilingual Dependency Parsing

Selective Sharing for Multilingual Dependency Parsing

Tahira Naseem
CSAIL, MIT
tahira@csail.mit.edu

Regina Barzilay
CSAIL, MIT
regina@csail.mit.edu

Amir Globerson
Hebrew University
gamir@cs.huji.ac.il

Abstract

We present a novel algorithm for multilingual dependency parsing that uses annotations from a diverse set of source languages to parse a new unannotated language. Our motivation is to broaden the advantages of multilingual learning to languages that exhibit significant differences from existing resource-rich languages. The algorithm learns which aspects of the source languages are relevant for the target language and ties model parameters accordingly. The model factorizes the process of generating a dependency tree into two steps: *selection* of syntactic dependents and their *ordering*. Being largely language-universal, the selection component is learned in a supervised fashion from all the training languages. In contrast, the ordering decisions are only influenced by languages with similar properties. We systematically model this cross-lingual sharing using typological features. In our experiments, the model consistently outperforms a state-of-the-art multilingual parser. The largest improvement is achieved on the non Indo-European languages yielding a gain of 14.4%.[1]

1 Introduction

Current top performing parsing algorithms rely on the availability of annotated data for learning the syntactic structure of a language. Standard approaches for extending these techniques to resource-lean languages either use parallel corpora or rely on annotated trees from other source languages. These techniques have been shown to work well for language families with many annotated resources (such as Indo-European languages). Unfortunately, for many languages there are no available parallel corpora or annotated resources in related languages. For such languages the only remaining option is to resort to unsupervised approaches, which are known to produce highly inaccurate results.

In this paper, we present a new multilingual algorithm for dependency parsing. In contrast to previous approaches, this algorithm can learn dependency structures using annotations from a diverse set of source languages, even if this set is not related to the target language. In our *selective sharing* approach, the algorithm learns which aspects of the source languages are relevant for the target language and ties model parameters accordingly. This approach is rooted in linguistic theory that characterizes the connection between languages at various levels of sharing. Some syntactic properties are universal across languages. For instance, nouns take adjectives and determiners as dependents, but not adverbs. However, the order of these dependents with respect to the parent is influenced by the typological features of each language.

To implement this intuition, we factorize generation of a dependency tree into two processes: selection of syntactic dependents and their ordering. The first component models the distribution of dependents for each part-of-speech tag, abstracting over their order. Being largely language-universal, this distribution can be learned in a supervised fashion from all the training languages. On the other hand,

[1]The source code for the work presented in this paper is available at http://groups.csail.mit.edu/rbg/code/unidep/

ordering of dependents varies greatly across languages and therefore should only be influenced by languages with similar properties. Furthermore, this similarity has to be expressed at the level of dependency types – i.e., two languages may share noun-adposition ordering, but differ in noun-determiner ordering. To systematically model this cross-lingual sharing, we rely on typological features that reflect ordering preferences of a given language. In addition to the known typological features, our parsing model embeds latent features that can capture cross-lingual structural similarities.

While the approach described so far supports a seamless transfer of shared information, it does not account for syntactic properties of the target language unseen in the training languages. For instance, in the CoNLL data, Arabic is the only language with the VSO ordering. To handle such cases, our approach augments cross-lingual sharing with unsupervised learning on the target languages.

We evaluated our selective sharing model on 17 languages from 10 language families. On this diverse set, our model consistently outperforms state-of-the-art multilingual dependency parsers. Performance gain, averaged over all the languages, is 5.9% when compared to the highest baseline. Our model achieves the most significant gains on non-Indo-European languages, where we see a 14.4% improvement. We also demonstrate that in the absence of observed typological information, a set of automatically induced latent features can effectively work as a proxy for typology.

2 Related Work

Traditionally, parallel corpora have been a mainstay of multilingual parsing (Wu, 1997; Kuhn, 2004; Smith and Smith, 2004; Hwa et al., 2005; Xi and Hwa, 2005; Burkett and Klein, 2008; Snyder et al., 2009). However, recent work in multilingual parsing has demonstrated the feasibility of transfer in the absence of parallel data. As a main source of guidance, these methods rely on the commonalities in dependency structure across languages. For instance, Naseem et al. (2010) explicitly encode these similarities in the form of universal rules which guide grammar induction in the target language. An alternative approach is to directly employ a non-lexicalized parser trained on one language to process a target language (Zeman and Resnik, 2008; McDonald et al., 2011; Søgaard, 2011). Since many unlexicalized dependencies are preserved across languages, these approaches are shown to be effective for related languages. For instance, when applied to the language pairs within the Indo-European family, such parsers outperform unsupervised monolingual techniques by a significant margin.

The challenge, however, is to enable dependency transfer for target languages that exhibit structural differences from source languages. In such cases, the extent of multilingual transfer is determined by the relation between source and target languages. Berg-Kirkpatrick and Klein (2010) define such a relation in terms of phylogenetic trees, and use this distance to selectively tie the parameters of monolingual syntactic models. Cohen et al. (2011) do not use a predefined linguistic hierarchy of language relations, but instead learn the contribution of source languages to the training mixture based on the likelihood of the target language. Søgaard (2011) proposes a different measure of language relatedness based on perplexity between POS sequences of source and target languages. Using this measure, he selects a subset of training source sentences that are closer to the target language. While all of the above techniques demonstrate gains from modeling language relatedness, they still underperform when the source and target languages are unrelated.

Our model differs from the above approaches in its emphasis on the selective information sharing driven by language relatedness. This is further combined with monolingual unsupervised learning. As our evaluation demonstrates, this layered approach broadens the advantages of multilingual learning to languages that exhibit significant differences from the languages in the training mix.

3 Linguistic Motivation

Language-Independent Dependency Properties Despite significant syntactic differences, human languages exhibit striking similarity in dependency patterns. For a given part-of-speech tag, the set of tags that can occur as its dependents is largely consistent across languages. For instance, adverbs and nouns are likely to be dependents of verbs, while adjectives

are not. Thus, these patterns can be freely transferred across languages.

Shared Dependency Properties Unlike dependent selection, the ordering of dependents in a sentence differs greatly across languages. In fact, cross-lingual syntactic variations are primarily expressed in different ordering of dependents (Harris, 1968; Greenberg, 1963). Fortunately, the dimensions of these variations have been extensively studied in linguistics and are documented in the form of typological features (Comrie, 1989; Haspelmath et al., 2005). For instance, most languages are either dominantly prepositional like English or post-positional like Urdu. Moreover, a language may be close to different languages for different dependency types. For instance, Portuguese is a prepositional language like English, but the order of its noun-adjective dependency is different from English and matches that of Arabic. Therefore, we seek a model that can express parameter sharing at the level of dependency types and can benefit from known language relations.

Language-specific Dependency Variations Not every aspect of syntactic structure is shared across languages. This is particularly true given a limited number of supervised source languages; it is quite likely that a target language will have previously unseen syntactic phenomena. In such a scenario, the raw text in the target language might be the only source of information about its unique aspects.

4 Model

We propose a probabilistic model for generating dependency trees that facilitates parameter sharing across languages. We assume a setup where dependency tree annotations are available for a set of source languages and we want to use these annotations to infer a parser for a target language. Syntactic trees for the target language are not available during training. We also assume that both source and target languages are annotated with a coarse parts-of-speech tagset which is shared across languages. Such tagsets are commonly used in multilingual parsing (Zeman and Resnik, 2008; McDonald et al., 2011; Søgaard, 2011; Naseem et al., 2010).

The key feature of our model is a two-tier approach that separates the *selection* of dependents from their *ordering*:

1. *Selection Component:* Determines the dependent tags given the parent tag.

2. *Ordering Component:* Determines the position of each dependent tag with respect to its parent (right or left) and the order within the right and left dependents.

This factorization constitutes a departure from traditional parsing models where these decisions are tightly coupled. By separating the two, the model is able to support different degrees of cross-lingual sharing on each level.

For the selection component, a reasonable approximation is to assume that it is the same for all languages. This is the approach we take here.

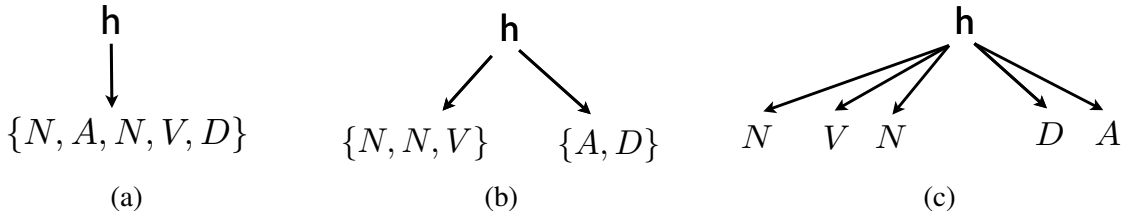
As mentioned in Section 3, the ordering of dependents is largely determined by the typological features of the language. We assume that we have a set of such features for every language l, and denote this feature vector by $\mathbf{v}_l$. We also experiment with a variant of our model where typological features are not observed. Instead, the model captures structural variations across languages by means of a small set of binary latent features. The values of these features are language dependent. We denote the set of latent features for language l by $\mathbf{b}_l$.

Finally, based on the well known fact that long distance dependencies are less likely (Eisner and Smith, 2010), we bias our model towards short dependencies. This is done by imposing a corpus-level soft constraint on dependency lengths using the posterior regularization framework (Graça et al., 2007).

4.1 Generative Process

Our model generates dependency trees one fragment at a time. A *fragment* is defined as a subtree comprising the immediate dependents of any node in the tree. The process recursively generates fragments in a head outwards manner, where the distribution over fragments depends on the head tag. If the generated fragment is not empty then the process continues for each child tag in the fragment, drawing new fragments from the distribution associated with the tag. The process stops when there are no more non-empty fragments.

A fragment with head node h is generated in language l via the following stages:



(a) (b) (c)

Figure 1: The steps of the generative process for a fragment with head h. In step (a), the unordered set of dependents is chosen. In step (b) they are partitioned into left and right unordered sets. Finally, each set is ordered in step (c).

- Generate the set of dependents of h via a distribution $P_{sel}(S|h)$. Here S is an unordered set of POS tags. Note that this part is universal (i.e., it does not depend on the language l).
- For each element in S decide whether it should go to the right or left of h as follows: for every $a \in S$, draw its direction from the distribution $P_{ord}(d|a, h, l)$, where $d \in \{R, L\}$. This results in two unordered sets S_R, S_L, the right and left dependents of h. This part *does* depend on the language l, since the relative ordering of dependents is not likely to be universal.
- Order the sets S_R, S_L. For simplicity, we assume that the order is drawn uniformly from all the possible unique permutations over S_R and S_L. We denote the number of such unique permutations of S_R by $n(S_R)$.[2] Thus the probability of each permutation of S_R is $\frac{1}{n(S_R)}$ [3].

Figure 1 illustrates the generative process. The first step constitutes the *selection* component and the last two steps constitute the *ordering* component. Given this generation scheme, the probability $P(D)$ of generating a given fragment D with head h will be:

$$P_{sel}(\{D\}|h) \prod_{a \in D} P_{ord}(d_D(a)|a, h, l) \frac{1}{n(D_R)n(D_L)} \quad (1)$$

Where we use the following notations:

- D_R, D_L denote the parts of the fragment that are to the left and right of h.
- $\{D\}$ is the unordered set of tags in D.
- $d_D(a)$ is the position (either R or L) of the dependent a w.r.t. the head of D.

In what follows we discuss the parameterizations of the different distributions.

4.1.1 Selection Component

The selection component draws an unordered set of tags S given the head tag h. We assume that the process is carried out in two steps. First the number of dependents n is drawn from a distribution:

$$P_{size}(n|h) = \theta_{size}(n|h) \quad (2)$$

where $\theta_{size}(n|h)$ is a parameter for each value of n and h. We restrict the maximum value of n to four, since this is a reasonable bound on the total number of dependents for a single parent node in a tree. These parameters are non-negative and satisfy $\sum_n \theta_{size}(n|h) = 1$. In other words, the size is drawn from a categorical distribution that is fully parameterized.

Next, given the size n, a set S with $|S| = n$ is drawn according to the following log-linear model:

$$P_{set}(S|h, n) = \frac{1}{Z_{set}(h, n)} e^{\sum_{S_i \in S} \theta_{sel}(S_i|h)}$$
$$Z_{set}(h, n) = \sum_{S:|S|=n} e^{\sum_{S_i \in S} \theta_{sel}(S_i|h)}$$

In the above, S_i is the i^{th} POS tag in the unordered set S, and $\theta_{sel}(S_i|h)$ are parameters. Thus, large values of $\theta_{sel}(S_i|h)$ indicate that POS S_i is more likely to appear in the subset with parent POS h.

Combining the above two steps we have the following distribution for selecting a set S of size n:

$$P_{sel}(S|h) = P_{size}(n|h) P_{set}(S|h, n) \,. \quad (3)$$

[2] This number depends on the count of each distinct tag in S_R. For example if $S_R = \{N, N, N\}$ then $n(S_R) = 1$. If $S_R = \{N, D, V\}$ then $n(S_R) = 3!$.

[3] We acknowledge that assuming a uniform distribution over the permutations of the right and left dependents is linguistically counterintuitive. However, it simplifies the model by greatly reducing the number of parameters to learn.

ID	Feature Description	Values
81A	Order of Subject, Object and Verb	SVO, SOV, VSO, VOS, OVS, OSV
85A	Order of Adposition and Noun	Postpositions, Prepositions, Inpositions
86A	Order of Genitive and Noun	Genitive-Noun, Noun-Genitive
87A	Order of Adjective and Noun	Adjective-Noun, Noun-Adjective
88A	Order of Demonstrative and Noun	Demonstrative-Noun, Noun-Demonstrative
89A	Order of Numeral and Noun	Numeral-Noun, Noun-Numeral

Table 1: The set of typological features that we use in our model. For each feature, the first column gives the ID of the feature as used in WALS, the second column describes the feature and the last column enumerates the allowable values for the feature. Besides these values, each feature can also have a value of 'No dominant order'.

4.1.2 Ordering Component

The ordering component consists of distributions $P_{ord}(d|a,h,l)$ that determine whether tag a will be mapped to the left or right of the head tag h. We model it using the following log-linear model:

$$P_{ord}(d|a,h,l) = \frac{1}{Z_{ord}(a,h,l)} e^{\mathbf{w}_{ord}\cdot\mathbf{g}(d,a,h,\mathbf{v}_l)}$$
$$Z_{ord}(a,h,l) = \sum_{d\in\{R,L\}} e^{\mathbf{w}_{ord}\cdot\mathbf{g}(d,a,h,\mathbf{v}_l)}$$

Note that in the above equations the ordering component depends on the known typological features $\mathbf{v}_l$. In the setup when typological features are not known, $\mathbf{v}_l$ is replaced with the latent ordering feature set $\mathbf{b}_l$.

The feature vector $\mathbf{g}$ contains indicator features for combinations of a, h, d and individual features v_{li} (i.e., the i^{th} typological features for language l).

4.2 Typological Features

The typological features we use are a subset of order-related typological features from "The World Atlas of Language Structure" (Haspelmath et al., 2005). We include only those features whose values are available for all the languages in our dataset. Table 1 summarizes the set of features that we use. Note that we do not explicitly specify the correspondence between these features and the model parameters. Instead, we leave it for the model to learn this correspondence automatically.

4.3 Dependency Length Constraint

To incorporate the intuition that long distance dependencies are less likely, we impose a posterior constraint on dependency length. In particular, we use the Posterior Regularization (PR) framework of Graça et al. (2007). The PR framework incorporates constraints by adding a penalty term to the standard likelihood objective. This term penalizes the distance of the model posterior from a set $\mathcal{Q}$, where $\mathcal{Q}$ contains all the posterior distributions that satisfy the constraints. In our case the constraint is that the expected dependency length is less than or equal to a pre-specified threshold value b. If we denote the latent dependency trees by z and the observed sentences by x then

$$\mathcal{Q} = \{q(z|x) : E_q[f(x,z)] \leq b\} \quad (4)$$

where $f(x,z)$ computes the sum of the lengths of all dependencies in z with respect to the linear order of x. We measure the length of a dependency relation by counting the number of tokens between the head and its modifier. The PR objective penalizes the KL-divergence of the model posterior from the set $\mathcal{Q}$:

$$\mathcal{L}_\theta(x) - \mathrm{KL}\left(\mathcal{Q} \parallel p_\theta(z|x)\right)$$

where θ denotes the model parameters and the first term is the log-likelihood of the data. This objective can be optimized using a modified version of the EM algorithm (Graça et al., 2007).

5 Parameter Learning

Our model is parameterized by the parameters θ_{sel}, θ_{size} and $\mathbf{w}_{ord}$. We learn these by maximizing the likelihood of the training data. As is standard, we add ℓ_2 regularization on the parameters and tune it on source languages. The likelihood is marginalized over all latent variables. These are:

- For sentences in the target language: all possible derivations that result in the observed POS tag sequences. The derivations include the choice of unordered sets size n, the unordered sets themselves S, their left/right al-

locations and the orderings within the left and right branches.

- For all languages: all possible values of the latent features $\mathbf{b}_l$.[4]

Since we are learning with latent variables, we use the EM algorithm to monotonically improve the likelihood. At each E step, the posterior over latent variables is calculated using the current model. At the M step this posterior is used to maximize the likelihood over the fully observed data. To compensate for the differences in the amount of training data, the counts from each language are normalized before computing the likelihood.

The M step involves finding maximum likelihood parameters for log-linear models in Equations 3 and 4. This is done via standard gradient based search; in particular, we use the method of BFGS.

We now briefly discuss how to calculate the posterior probabilities. For estimating the $\mathbf{w}_{ord}$ parameters we require marginals of the type $P(b_{li}|\mathcal{D}_l; \mathbf{w}^t)$ where $\mathcal{D}_l$ are the sentences in language l, b_{li} is the i_{th} latent feature for the language l and $\mathbf{w}^t$ are the parameter values at iteration t. Consider doing this for a source language l. Since the parses are known, we only need to marginalize over the other latent features. This can be done in a straightforward manner by using our probabilistic model. The complexity is exponential in the number of latent features, since we need to marginalize over all features other than b_{li}. This is feasible in our case, since we use a relatively small number of such features.

When performing unsupervised learning for the target language, we need to marginalize over possible derivations. Specifically, for the M step, we need probabilities of the form $P(\text{a modifies h}|\mathcal{D}_l; \mathbf{w}^t)$. These can be calculated using a variant of the inside outside algorithm. The exact version of this algorithm would be exponential in the number of dependents due to the $\frac{1}{n(S_r)}$ term in the permutation factor. Although it is possible to run this exact algorithm in our case, where the number of dependents is limited to 4, we use an approximation that works well in practice: instead of $\frac{1}{n(S_r)}$ we use $\frac{1}{|S_r|!}$. In this case the runtime is no longer exponential in the number of children, so inference is much faster.

[4]This corresponds to the case when typological features are not known.

Finally, given the trained parameters we generate parses in the target language by calculating the maximum a posteriori derivation. This is done using a variant of the CKY algorithm.

6 Experimental Setup

Datasets and Evaluation We test the effectiveness of our approach on 17 languages: Arabic, Basque, Bulgarian, Catalan, Chinese, Czech, Dutch, English, German, Greek, Hungarian, Italian, Japanese, Portuguese, Spanish, Swedish and Turkish. We used datasets distributed for the 2006 and 2007 CoNLL Shared Tasks (Buchholz and Marsi, 2006; Nivre et al., 2007). Each dataset provides manually annotated dependency trees and POS tags. To enable crosslingual sharing, we map the gold part-of-speech tags in each corpus to a common coarse tagset (Zeman and Resnik, 2008; Søgaard, 2011; McDonald et al., 2011; Naseem et al., 2010). The coarse tagset consists of 11 tags: noun, verb, adjective, adverb, pronoun, determiner, adposition, numeral, conjunction, particle, punctuation mark, and X (a catch-all tag). Among several available fine-to-coarse mapping schemes, we employ the one of Naseem et al. (2010) that yields consistently better performance for our method and the baselines than the mapping proposed by Petrov et al. (2011).

As the evaluation metric, we use directed dependency accuracy. Following standard evaluation practices, we do not evaluate on punctuation. For both the baselines and our model we evaluate on all sentences of length 50 or less ignoring punctuation.

Training Regime Our model typically converges quickly and does not require more than 50 iterations of EM. When the model involves latent typological variables, the initialization of these variables can impact the final performance. As a selection criterion for initialization, we consider the performance of the final model averaged over the supervised source languages. We perform ten random restarts and select the best according to this criterion. Likewise, the threshold value b for the PR constraint on the dependency length is tuned on the source languages, using average test set accuracy as the selection criterion.

Baselines We compare against the state-of-the-art multilingual dependency parsers that do not use parallel corpora for training. All the systems were eval-

uated using the same fine-to-coarse tagset mapping. The first baseline, *Transfer*, uses direct transfer of a discriminative parser trained on all the source languages (McDonald et al., 2011). This simple baseline achieves surprisingly good results, within less than 3% difference from a parser trained using parallel data. In the second baseline (*Mixture*), parameters of the target language are estimated as a weighted mixture of the parameters learned from annotated source languages (Cohen et al., 2011). The underlying parsing model is the dependency model with valance (DMV) (Klein and Manning, 2004). Originally, the baseline methods were evaluated on different sets of languages using a different tag mapping. Therefore, we obtained new results for these methods in our setup. For the *Transfer* baseline, for each target language we trained the model on all other languages in our dataset. For the *Mixture* baseline, we trained the model on the same four languages used in the original paper — English, German, Czech and Italian. When measuring the performance on these languages, we selected another set of four languages with a similar level of diversity.[5]

7 Results

Table 2 summarizes the performance for different configurations of our model and the baselines.

Comparison against Baselines On average, the selective sharing model outperforms both baselines, yielding 8.9% gain over the weighted mixture model (Cohen et al., 2011) and 5.9% gain over the direct transfer method (McDonald et al., 2011). Our model outperforms the weighted mixture model on 15 of the 17 languages and the transfer method on 12 of the 17 languages. Most of the gains are obtained on non-Indo-European languages, that have little similarity with the source languages. For this set, the average gain over the transfer baseline is 14.4%. With some languages, such as Japanese, achieving gains of as much as 30%.

On Indo-European languages, the model performance is almost equivalent to that of the best performing baseline. To explain this result we consider the performance of the supervised version of our model which constitutes an upper bound on the performance. The average accuracy of our supervised model on these languages is 66.8%, compared to the 76.3% of the unlexicalized MST parser. Since Indo-European languages are overrepresented in our dataset, a target language from this family is likely to exhibit more similarity to the training data. When such similarity is substantial, the transfer baseline will benefit from the power of a context-rich discriminative parser.

A similar trait can be seen by comparing the performance of our model to an oracle version of our model which selects the optimal source language for a given target language (column 7). Overall, our method performs similarly to this oracle variant. However, the gain for non Indo-European languages is 1.9% vs -1.3% for Indo-European languages.

Analysis of Model Properties We first test our hypothesis about the universal nature of the dependent selection. We compare the performance of our model (column 6) against a variant (column 8) where this component is trained from annotations on the target language. The performance of the two is very close – 1.8%, supporting the above hypothesis.

To assess the contribution of other layers of selective sharing, we first explore the role of typological features in learning the ordering component. When the model does not have access to observed typological features, and does not use latent ones (column 4), the accuracy drops by 2.6%[6]. For some languages (e.g., Turkish) the decrease is very pronounced. Latent typological features (column 5) do not yield the same gain as observed ones, but they do improve the performance of the typology-free model by 1.4%.

Next, we show the importance of using raw target language data in training the model. When the model has to make all the ordering decisions based on meta-linguistic features without account for unique properties of the target languages, the performance decreases by 0.9% (see column 3).

To assess the relative difficulty of learning the ordering and selection components, we consider model variants where each of these components is

[5] We also experimented with a version of the Cohen et al. (2011) model trained on all the source languages. This setup resulted in decreased performance. For this reason, we chose to train the model on the four languages.

[6] In this setup, the ordering component is trained in an unsupervised fashion on the target language.

	Baselines		Selective Sharing Model							
	Mixture	Transfer	(D-,T_o)	(D+)	(D+,T_l)	(D+,T_o)	Best Pair	Sup. Sel.	Sup. Ord.	MLE
Catalan	64.9	69.5	71.9	66.1	66.7	71.8	74.8	70.2	73.2	72.1
Italian	61.9	68.3	68.0	65.5	64.2	65.6	68.3	65.1	70.7	72.3
Portuguese	72.9	75.8	76.2	72.3	76.0	73.5	76.4	77.4	77.6	79.6
Spanish	57.2	65.9	62.3	58.5	59.4	62.1	63.4	61.5	62.6	65.3
Dutch	50.1	53.9	56.2	56.1	55.8	55.9	57.8	56.3	58.6	58.0
English	45.9	47.0	47.6	48.5	48.1	48.6	44.4	46.3	60.0	62.7
German	54.5	56.4	54.0	53.5	54.3	53.7	54.8	52.4	56.2	58.0
Swedish	56.4	63.6	52.0	61.4	60.6	61.5	63.5	67.9	67.1	73.0
Bulgarian	67.7	64.0	67.6	63.5	63.9	66.8	66.1	66.2	69.5	71.0
Czech	39.6	40.3	43.9	44.7	45.4	44.6	47.5	53.2	51.2	58.9
Arabic	44.8	40.7	57.2	58.8	60.3	58.9	57.6	62.9	61.9	64.2
Basque	32.8	32.4	39.7	40.1	39.8	47.6	42.0	46.2	47.9	51.6
Chinese	46.7	49.3	59.9	52.2	52.0	51.2	65.4	62.3	65.5	73.5
Greek	56.8	60.4	61.9	67.5	67.3	67.4	60.6	67.2	69.0	70.5
Hungarian	46.8	54.3	56.9	58.4	58.8	58.5	57.0	57.4	62.0	61.6
Japanese	33.5	34.7	62.3	56.8	61.4	64.0	54.8	63.4	69.7	75.6
Turkish	28.3	34.3	59.1	43.6	57.8	59.2	56.9	66.6	59.5	67.6
Average	50.6	53.6	58.6	56.9	58.3	**59.5**	59.5	61.3	63.7	66.8

Table 2: Directed dependency accuracy of different variants of our selective sharing model and the baselines. The first section of the table (column 1 and 2) shows the accuracy of the weighted mixture baseline (Cohen et al., 2011) (Mixture) and the multi-source transfer baseline (McDonald et al., 2011) (Transfer). The middle section shows the performance of our model in different settings. D$\pm$ indicates the presence/absence of raw target language data during training. T_o indicates the use of observed typological features for all languages and T_l indicates the use of latent typological features for all languages. The last section shows results of our model with different levels of oracle supervision: a. (Best Pair) Model parameters are borrowed from the best source language based on the accuracy on the target language b. (Sup. Sel.) Selection component is trained using MLE estimates from target language c. (Sup. Ord.) Ordering component is trained using MLE estimates from the target language d. (MLE) All model parameters are trained on the target language in a supervised fashion. The horizontal partitions separate language families. The first three families are sub-divisions of the Indo-European language family.

trained using annotations in the target language. As shown in columns 8 and 9, these two variants outperform the original model, achieving 61.3% for supervised selection and 63.7% for supervised ordering. Comparing these numbers to the accuracy of the original model (column 6) demonstrates the difficulty inherent in learning the ordering information. This finding is expected given that ordering involves selective sharing from multiple languages.

Overall, the performance gap between the selective sharing model and its monolingual supervised counterpart is 7.3%. In contrast, the unsupervised monolingual variant of our model achieves a meager 26%.[7] This demonstrates that our model can effectively learn relevant aspects of syntactic structure from a diverse set of languages.

[7]This performance is comparable to other generative models such as DMV (Klein and Manning, 2004).

8 Conclusions

We present a novel algorithm for multilingual dependency parsing that uses annotations from a diverse set of source languages to parse a new unannotated language. Overall, our model consistently outperforms the multi-source transfer based dependency parser of McDonald et al. (2011). Our experiments demonstrate that the model is particularly effective in processing languages that exhibit significant differences from the training languages.

Acknowledgments

The authors acknowledge the support of the NSF (IIS-0835445), the MURI program (W911NF-10-1-0533), the DARPA BOLT program, and the ISF (1789/11). We thank Tommi Jaakkola, Ryan McDonald and the members of the MIT NLP group for their comments.

References

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *ACL*, pages 1288–1297.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, pages 149–164.

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

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*, pages 50–61.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford: Blackwell.

Jason Eisner and Noah A. Smith. 2010. Favor short dependencies: Parsing with soft and hard constraints on dependency length. In *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing*, pages 121–150.

João Graça, Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *Advances in NIPS*, pages 569–576.

Joseph H Greenberg. 1963. Some universals of language with special reference to the order of meaningful elements. In Joseph H Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press.

Z.S. Harris. 1968. *Mathematical structures of language*. Wiley.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

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

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, pages 478–485.

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

Ryan T. McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *EMNLP*, pages 1234–1244.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. In *ArXiv*, April.

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

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of ACL/AFNLP*, pages 73–81.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL (Short Papers)*, pages 682–686.

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

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

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, January.