Target Language Adaptation of Discriminative Transfer Parsers

Oscar Täckström*
SICS | Uppsala University
Sweden
oscar@sics.se

Ryan McDonald
Google
New York
ryanmcd@google.com

Joakim Nivre*
Uppsala University
Sweden
joakim.nivre@lingfil.uu.se

Abstract

We study multi-source transfer parsing for resource-poor target languages; specifically methods for target language adaptation of delexicalized discriminative graph-based dependency parsers. We first show how recent insights on selective parameter sharing, based on typological and language-family features, can be applied to a discriminative parser by carefully decomposing its model features. We then show how the parser can be relexicalized and adapted using unlabeled target language data and a learning method that can incorporate diverse knowledge sources through ambiguous labelings. In the latter scenario, we exploit two sources of knowledge: arc marginals derived from the base parser in a self-training algorithm, and arc predictions from multiple transfer parsers in an ensemble-training algorithm. Our final model outperforms the state of the art in multi-source transfer parsing on 15 out of 16 evaluated languages.

1 Introduction

Many languages still lack access to core NLP tools, such as part-of-speech taggers and syntactic parsers. This is largely due to the reliance on *fully supervised* learning methods, which require large quantities of manually annotated training data. Recently, methods for *cross-lingual transfer* have appeared as a promising avenue for overcoming this hurdle for both part-of-speech tagging (Yarowsky et al., 2001; Das and Petrov, 2011) and syntactic dependency parsing (Hwa et al., 2005; Zeman and Resnik, 2008; Ganchev et al., 2009; McDonald et al., 2011; Naseem et al., 2012). While these methods do not yet compete with fully supervised approaches, they can drastically outperform both *unsupervised* methods (Klein and Manning, 2004) and *weakly supervised* methods (Naseem et al., 2010; Berg-Kirkpatrick and Klein, 2010).

A promising approach to cross-lingual transfer of syntactic dependency parsers is to use multiple source languages and to tie model parameters across related languages. This idea was first explored for weakly supervised learning (Cohen and Smith, 2009; Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010) and recently by Naseem et al. (2012) for multi-source cross-lingual transfer. In particular, Naseem et al. showed that by *selectively sharing* parameters based on typological features of each language, substantial improvements can be achieved, compared to using a single set of parameters for all languages. However, these methods all employ generative models with strong independence assumptions and weak feature representations, which upper bounds their accuracy far below that of feature-rich discriminative parsers (McDonald et al., 2005; Nivre, 2008).

In this paper, we improve upon the state of the art in cross-lingual transfer of dependency parsers from multiple source languages by adapting feature-rich discriminatively trained parsers to a specific target language. First, in §4 we show how selective sharing of model parameters based on typological traits can be incorporated into a delexicalized discriminative graph-based parsing model. This requires a careful decomposition of features into language-generic and language-specific sets in order to tie specific target language parameters to their relevant source language counterparts. The resulting parser outperforms the method of Naseem et al. (2012) on 12 out of 16 evaluated languages. Second, in §5 we introduce a train-

*Work primarily carried out while at Google, NY.

Proceedings of NAACL-HLT 2013, pages 1061–1071,
Atlanta, Georgia, 9–14 June 2013.

ing method that can incorporate diverse knowledge sources through ambiguously predicted labelings of unlabeled target language data. This permits effective relexicalization and target language adaptation of the transfer parser. Here, we experiment with two different knowledge sources: arc sets, which are filtered by marginal probabilities from the cross-lingual transfer parser, are used in an *ambiguity-aware self-training* algorithm (§5.2); these arc sets are then combined with the predictions of a different transfer parser in an *ambiguity-aware ensemble-training* algorithm (§5.3). The resulting parser provides significant improvements over a strong baseline parser and achieves a 13% relative error reduction on average with respect to the best model of Naseem et al. (2012), outperforming it on 15 out of the 16 evaluated languages.

2 Multi-Source Delexicalized Transfer

The methods proposed in this paper fall into the *delexicalized transfer* approach to multilingual syntactic parsing (Zeman and Resnik, 2008; McDonald et al., 2011; Cohen et al., 2011; Søgaard, 2011). In contrast to *annotation projection* approaches (Yarowsky et al., 2001; Hwa et al., 2005; Ganchev et al., 2009; Spreyer and Kuhn, 2009), delexicalized transfer methods do not rely on any bitext. Instead, a parser is trained on annotations in a source language, relying solely on features that are available in both the source and the target language, such as "universal" part-of-speech tags (Zeman and Resnik, 2008; Naseem et al., 2010; Petrov et al., 2012), cross-lingual word clusters (Täckström et al., 2012) or type-level features derived from bilingual dictionaries (Durrett et al., 2012).[1] This parser is then directly used to parse the target language. For languages with similar typology, this method can be quite accurate, especially when compared to purely unsupervised methods. For instance, a parser trained on English with only part-of-speech features can correctly parse the Greek sentence in Figure 1, even without knowledge of the lexical items since the sequence of part-of-speech tags determines the syntactic structure almost unambiguously.

Learning with multiple languages has been shown to benefit unsupervised learning (Cohen and Smith, 2009; Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010). Annotations in multiple languages can be combined in delexicalized transfer as well, as long as the parser features are available across the involved languages. This idea was explored by McDonald et al. (2011), who showed that target language accuracy can be improved by simply concatenating delexicalized treebanks in multiple languages. In similar work, Cohen et al. (2011) proposed a mixture model in which the parameters of a generative target language parser is expressed as a linear interpolation of source language parameters, whereas Søgaard (2011) showed that target side language models can be used to selectively subsample training sentences to improve accuracy. Recently, inspired by the phylogenetic prior of Berg-Kirkpatrick and Klein (2010), Søgaard and Wulff (2012) proposed — among other ideas — a typologically informed weighting heuristic for linearly interpolating source language parameters. However, this weighting did not provide significant improvements over uniform weighting.

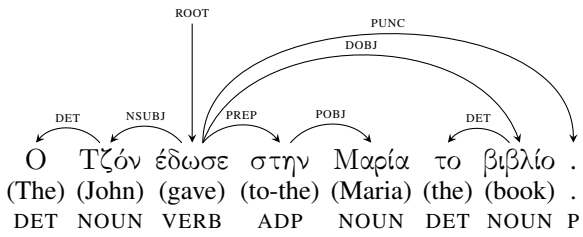

Figure 1: A Greek sentence which is correctly parsed by a delexicalized English parser, provided that part-of-speech tags are available in both the source and target language.

The aforementioned approaches work well for transfer between similar languages. However, their assumptions cease to hold for typologically divergent languages; a target language can rarely be described as a linear combination of data or model parameters from a set of source languages, as languages tend to share varied typological traits; this critical insight is discussed further in §4. To account for this issue, Naseem et al. (2012) recently introduced a novel generative model of dependency parsing, in which the generative process is factored into separate steps for the *selection* of dependents and their *ordering*. The parameters used in the selection step are all language independent, capturing only head-dependent attachment preferences. In the ordering step, however, parameters are *selectively shared* between subsets of

[1] Note that Täckström et al. (2012) and Durrett et al. (2012) do require bitext or a bilingual dictionary. The same holds for most cross-lingual representations, e.g., Klementiev et al. (2012).

Feature	Description
81A	Order of Subject, Object and Verb
85A	Order of Adposition and Noun
86A	Order of Genitive and Noun
87A	Order of Adjective and Noun
88A	Order of Demonstrative and Noun
89A	Order of Numeral and Noun

Table 1: Typological features from WALS (Dryer and Haspelmath, 2011), proposed for selective sharing by Naseem et al. (2012). Feature 89A has the same value for all studied languages, while 88A differs only for Basque. These features are therefore subsequently excluded.

source languages based on typological features of the languages extracted from WALS — the World Atlas of Language Structures (Dryer and Haspelmath, 2011) — as shown in Table 1. In the transfer scenario, where no supervision is available in the target language, this parser achieves the hitherto best published results across a number of languages; in particular for target languages with a word order divergent from the source languages.

However, the generative model of Naseem et al. is quite impoverished. In the fully supervised setting, it obtains substantially lower accuracies compared to a standard arc-factored graph-based parser (McDonald et al., 2005). Averaged across 16 languages,[2] the generative model trained with full supervision on the target language obtains an accuracy of 67.1%. A comparable lexicalized discriminative arc-factored model (McDonald et al., 2005) obtains 84.1%. Even when delexicalized, this model reaches 78.9%. This gap in supervised accuracy holds for all 16 languages. Thus, while selective sharing is a powerful device for transferring parsers across languages, the underlying generative model used by Naseem et al. (2012) restricts its potential performance.

3 Basic Models and Experimental Setup

Inspired by the superiority of discriminative graph-based parsing in the supervised scenario, we investigate whether the insights of Naseem et al. (2012) on selective parameter sharing can be incorporated into such models in the transfer scenario. We first review the basic graph-based parser framework and the experimental setup that we will use throughout. We then delve into details on how to incorporate selective sharing in this model in §4. In §5, we show how learning with ambiguous labelings in this parser can be used for further target language adaptation, both through self-training and through ensemble-training.

[2]Based on results in Naseem et al. (2012), excluding English.

3.1 Discriminative Graph-Based Parser

Let x denote an input sentence and let $y \in \mathcal{Y}(x)$ denote a dependency tree, where $\mathcal{Y}(x)$ is the set of well-formed dependency trees spanning x. Henceforth, we restrict $\mathcal{Y}(x)$ to projective dependency trees, but all our methods are equally applicable in the non-projective case. Provided a vector of model parameters θ, the probability of a dependency tree $y \in \mathcal{Y}(x)$, conditioned on a sentence x, has the following form:

$$p_\theta(y \mid x) = \frac{\exp\left\{\theta^\top \Phi(x, y)\right\}}{\sum_{y' \in \mathcal{Y}(x)} \exp\left\{\theta^\top \Phi(x, y')\right\}} \,.$$

Without loss of generality, we restrict ourselves to first-order models, where the feature function $\Phi(x, y)$ factors over individual arcs (h, m) in y, such that

$$\Phi(x, y) \;= \sum_{(h,m) \in y} \phi(x, h, m) \,,$$

where $h \in [0, |x|]$ and $m \in [1, |x|]$ are the indices of the head word and the dependent word of the arc; $h = 0$ represents a dummy ROOT token. The model parameters are estimated by maximizing the log-likelihood of the training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$,

$$\mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log\; p_\theta(y_i \mid x_i) \,.$$

We use the standard gradient-based L-BFGS algorithm (Liu and Nocedal, 1989) to maximize the log-likelihood. Eisner's algorithm (Eisner, 1996) is used for inference of the Viterbi parse and arc-marginals.

3.2 Data Sets and Experimental Setup

To facilitate comparison with the state of the art, we use the same treebanks and experimental setup as Naseem et al. (2012). Notably, we use the mapping proposed by Naseem et al. (2010) to map from fine-grained treebank specific part-of-speech tags to coarse-grained "universal" tags, rather than the more recent mapping proposed by Petrov et al. (2012). For

Delexicalized MSTParser

$\texttt{d} \otimes \texttt{l}$; $[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p}$; $[\texttt{d} \otimes \texttt{l}] \otimes \texttt{m.p}$
$[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} \otimes \texttt{m.p}$
$[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} \otimes \texttt{h}_{+1}\texttt{.p} \otimes \texttt{m}_{-1}\texttt{.p} \otimes \texttt{m.p}$
$[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h}_{-1}\texttt{.p} \otimes \texttt{h.p} \otimes \texttt{m}_{-1}\texttt{.p} \otimes \texttt{m.p}$
$[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} \otimes \texttt{h}_{+1}\texttt{.p} \otimes \texttt{m.p} \otimes \texttt{m}_{+1}\texttt{.p}$
$[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h}_{-1}\texttt{.p} \otimes \texttt{h.p} \otimes \texttt{m.p} \otimes \texttt{m}_{+1}\texttt{.p}$
$\texttt{h.p} \otimes \texttt{between.p} \otimes \texttt{m.p}$

Bare

$\texttt{l}$
$[\texttt{l}] \otimes \texttt{h.p}$
$[\texttt{l}] \otimes \texttt{m.p}$
$[\texttt{l}] \otimes \texttt{h.p} \otimes \texttt{m.p}$

Selectively shared

$\texttt{d} \otimes \texttt{w.81A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{VERB} \wedge \texttt{m.p} = \texttt{NOUN}]$
$\texttt{d} \otimes \texttt{w.81A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{VERB} \wedge \texttt{m.p} = \texttt{PRON}]$
$\texttt{d} \otimes \texttt{w.85A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{ADP} \wedge \texttt{m.p} = \texttt{NOUN}]$
$\texttt{d} \otimes \texttt{w.85A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{ADP} \wedge \texttt{m.p} = \texttt{PRON}]$
$\texttt{d} \otimes \texttt{w.86A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{NOUN} \wedge \texttt{m.p} = \texttt{NOUN}]$
$\texttt{d} \otimes \texttt{w.87A} \otimes \mathbb{1}[\texttt{h.p} = \texttt{NOUN} \wedge \texttt{m.p} = \texttt{ADJ}]$

Figure 2: Arc-factored feature templates for graph-based parsing. Direction: $\texttt{d} \in \{\texttt{LEFT}, \texttt{RIGHT}\}$; dependency length: $\texttt{l} \in \{1, 2, 3, 4, 5+\}$; part of speech of head / dependent / words between head and dependent: $\texttt{h.p}$ / $\texttt{m.p}$ / $\texttt{between.p} \in \{\texttt{NOUN}, \texttt{VERB}, \texttt{ADJ}, \texttt{ADV}, \texttt{PRON}, \texttt{DET}, \texttt{ADP}, \texttt{NUM}, \texttt{CONJ}, \texttt{PRT}, \texttt{PUNC}, \texttt{X}\}$; token to the left / right of $\texttt{z}$: $\texttt{z}_{-1}$ / $\texttt{z}_{+1}$; WALS features: $\texttt{w.X}$ for $\texttt{X} = \texttt{81A}, \texttt{85A}, \texttt{86A}, \texttt{87A}$ (see Table 1). $[\cdot]$ denotes an optional template, e.g., $[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} \otimes \texttt{m.p}$ expands to templates $\texttt{d} \otimes \texttt{l} \otimes \texttt{h.p} \otimes \texttt{m.p}$ and $\texttt{h.p} \otimes \texttt{m.p}$, so that the template also falls back on its undirectional variant.

each target language evaluated, the treebanks of the remaining languages are used as *labeled* training data, while the target language treebank is used for testing only (in §5 a different portion of the target language treebank is additionally used as *unlabeled* training data). We refer the reader to Naseem et al. (2012) for detailed information on the different treebanks. Due to divergent treebank annotation guidelines, which makes fine-grained evaluation difficult, all results are evaluated in terms of unlabeled attachment score (UAS). In line with Naseem et al. (2012), we use gold part-of-speech tags and evaluate only on sentences of length 50 or less excluding punctuation.

3.3 Baseline Models

We compare our models to two multi-source baseline models. The first baseline, *NBG*, is the generative model with selective parameter sharing from Naseem et al. (2012).[3] This model is trained without target language data, but we investigate the use of such data in §5.4. The second baseline, *Delex*, is a delexicalized projective version of the well-known graph-based MSTParser (McDonald et al., 2005). The feature templates used by this model are shown to the left in Figure 2. Note that there is no selective sharing in this model.

The second and third columns of Table 2 show the unlabeled attachment scores of the baseline models for each target language. We see that *Delex* performs well on target languages that are related to the majority of the source languages. However, for languages that diverge from the Indo-European majority family, the selective sharing model, *NBG*, achieves substantially higher accuracies.

4 Feature-Based Selective Sharing

The results for the baseline models are not surprising considering the feature templates used by *Delex*. There are two fundamental issues with these features when used for direct transfer. First, all but one template include the arc direction. Second, some features are sensitive to local word order; e.g., $[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} \otimes \texttt{h}_{+1}\texttt{.p} \otimes \texttt{m}_{-1}\texttt{.p} \otimes \texttt{m.p}$, which models direction as well as word order in the local contexts of the head and the dependent. Such features do not transfer well across typologically different languages.

In order to verify that these issues are the cause of the poor performance of the *Delex* model, we remove all directional features and all features that model local word order from *Delex*. The feature templates of the resulting *Bare* model are shown in the center of Figure 2. These features only model selectional preferences and dependency length, analogously to the selection component of *NBG*. The performance of *Bare* is shown in the fourth column of Table 2. The removal of most of the features results in a performance drop on average. However, for languages outside of the Indo-European family, *Bare* is often more accurate, especially for Basque, Hungarian and Japanese, which supports our hypothesis.

4.1 Sharing Based on Typological Features

After removing all directional features, we now carefully reintroduce them. Inspired by Naseem et al.

[3] Model "D-,T_o" in Table 2 from Naseem et al. (2012).

		Graph-Based Models				
Lang.	NBG	Delex	Bare	Share	Similar	Family
ar	**57.2**	43.3	43.1	52.7	52.7	52.7
bg	**67.6**	64.5	56.1	65.4	62.4	65.4
ca	71.9	72.0	58.1	66.1	**80.2**	77.6
cs	43.9	40.5	43.1	42.5	**45.3**	43.5
de	54.0	57.0	49.3	55.2	58.1	**59.2**
el	61.9	63.2	57.7	62.9	59.9	**63.2**
es	62.3	66.9	52.6	59.3	**69.0**	67.1
eu	39.7	29.5	43.3	**46.8**	**46.8**	**46.8**
hu	56.9	56.2	60.5	**64.5**	**64.5**	**64.5**
it	68.0	70.8	55.7	63.5	**74.6**	72.5
ja	62.3	38.9	50.6	57.1	64.6	**65.9**
nl	56.2	**57.9**	51.6	55.0	51.8	56.8
pt	76.2	77.5	63.0	72.7	**78.4**	**78.4**
sv	52.0	61.4	55.9	58.8	48.8	**63.5**
tr	59.1	37.4	36.0	41.7	**59.5**	59.4
zh	**59.9**	45.1	47.9	54.8	54.8	54.8
avg	59.3	55.1	51.5	57.4	60.7	**62.0**

Table 2: Unlabeled attachment scores of the multi-source transfer models. Boldface numbers indicate the best result per language. Underlined numbers indicate languages whose group is not represented in the training data (these default to *Share* under *Similarity* and *Family*). *NBG* is the "D-,T_o" model in Table 2 from Naseem et al. (2012).

(2012), we make use of the typological features from WALS (Dryer and Haspelmath, 2011), listed in Table 1, to selectively share directional parameters between languages. As a natural first attempt at sharing parameters, one might consider forming the cross-product of all features of *Delex* with all WALS properties, similarly to a common domain adaptation technique (Daumé III, 2007; Finkel and Manning, 2009). However, this approach has two issues. First, it results in a huge number of features, making the model prone to overfitting. Second, and more critically, it ties together languages via features for which they are not typologically similar. Consider English and French, which are both prepositional and thus have the same value for WALS property 85A. These languages will end up sharing a parameter for the feature $[\texttt{d} \otimes \texttt{l}] \otimes \texttt{h.p} = \texttt{NOUN} \otimes \texttt{m.p} = \texttt{ADJ} \otimes \texttt{w.85A}$; yet they have the exact opposite direction of attachment preference when it comes to nouns and adjectives. This problem applies to any method for parameter mixing that treats all the parameters as equal.

Like Naseem et al. (2012), we instead share parameters more selectively. Our strategy is to use the relevant part-of-speech tags of the head and dependent to select which parameters to share, based on very basic linguistic knowledge. The resulting features are shown to the right in Figure 2. For example, there is a shared directional feature that models the order of Subject, Object and Verb by conjoining WALS feature 81A with the arc direction and an indicator feature that fires only if the head is a verb and the dependent is a noun. These features would not be very useful by themselves, so we combine them with the *Bare* features. The accuracy of the resulting *Share* model is shown in column five of Table 2. Although this model still performs worse than *NBG*, it is an improvement over the *Delex* baseline and actually outperforms the former on 5 out of the 16 languages.

4.2 Sharing Based on Language Groups

While *Share* models selectional preferences and arc directions for a subset of dependency relations, it does not capture the rich local word order information captured by *Delex*. We now consider two ways of selectively including such information based on language similarity. While more complex sharing could be explored (Berg-Kirkpatrick and Klein, 2010), we use a flat structure and consider two simple groupings of the source and target languages.

First, the *Similar* model consists of the features used by *Share* together with the features from *Delex* in Figure 2. The latter are conjoined with an indicator feature that fires only when the source and target languages share values for all the WALS features in Table 1. This is accomplished by adding the template

$$\texttt{f} \otimes [\texttt{w.81A} \otimes \texttt{w.85A} \otimes \texttt{w.86A} \otimes \texttt{w.87A} \otimes \texttt{w.88A}]$$

for each template `f` in *Delex*. This groups: 1) Catalan, Italian, Portuguese and Spanish; 2) Bulgarian, Czech and English; 3) Dutch, German and Greek; and 4) Japanese and Turkish. The remaining languages do not share all WALS properties with at least one source language and thus revert to *Share*, since they cannot exploit these grouped features.

Second, instead of grouping languages according to WALS, the *Family* model is based on a simple subdivision into Indo-European languages (Bulgarian, Catalan, Czech, Greek, English, Spanish, Italian,

Dutch, Portuguese, Swedish) and Altaic languages (Japanese, Turkish). This is accomplished with indicator features analogous to those used in *Similar*. The remaining languages are again treated as isolates and revert to *Similar*.

The results for these models are given in the last two columns of Table 2. We see that by adding these rich features back into the fold, but having them fire only for languages in the same group, we can significantly increase the performance — from 57.4% to 62.0% on average when considering *Family*. If we consider our original *Delex* baseline, we see an absolute improvement of 6.9% on average and a relative error reduction of 15%. Particular gains are seen for non-Indo-European languages; e.g., Japanese increases from 38.9% to 65.9%. Furthermore, *Family* achieves a 7% relative error reduction over the *NBG* baseline and outperforms it on 12 of the 16 languages. This shows that a discriminative graph-based parser can achieve higher accuracies compared to generative models when the features are carefully constructed.

5 Target Language Adaptation

While some higher-level linguistic properties of the target language have been incorporated through selective sharing, so far no features specific to the target language have been employed. Cohen et al. (2011) and Naseem et al. (2012) have shown that using expectation-maximization (EM) to this end can in some cases bring substantial accuracy gains. For discriminative models, self-training has been shown to be quite effective for adapting monolingual parsers to new domains (McClosky et al., 2006), as well as for *relexicalizing* delexicalized parsers using unlabeled target language data (Zeman and Resnik, 2008). Similarly Täckström (2012) used self-training to adapt a multi-source direct transfer named-entity recognizer (Täckström et al., 2012) to different target languages, "relexicalizing" the model with word cluster features. However, as discussed in §5.2, standard self-training is not optimal for target language adaptation.

5.1 Ambiguity-Aware Training

In this section, we propose a related training method: *ambiguity-aware training*. In this setting a discriminative probabilistic model is induced from automatically inferred *ambiguous labelings* over unlabeled target language data, in place of gold-standard dependency trees. The ambiguous labelings can combine multiple sources of evidence to guide the estimation or simply encode the underlying uncertainty from the base parser. This uncertainty is marginalized out during training. The structure of the output space, e.g., projectivity and single-headedness constraints, along with regularities in the feature space, can together guide the estimation, similar to what occurs with the expectation-maximization algorithm.

Core to this method is the idea of an *ambiguous labeling* $\tilde{\boldsymbol{y}}(x) \subseteq \mathcal{Y}(x)$, which encodes a set of possible dependency trees for an input sentence x. In subsequent sections we describe how to define such labelings. Critically, $\tilde{\boldsymbol{y}}(x)$ should be large enough to capture the correct labeling, but on the other hand small enough to provide concrete guidance for model estimation. Ideally, $\tilde{\boldsymbol{y}}(x)$ will capture heterogenous knowledge that can aid the parser in target language adaptation. In a first-order arc-factored model, we define $\tilde{\boldsymbol{y}}(x)$ in terms of a collection of *ambiguous arc sets* $\mathcal{A}(x) = \{\mathcal{A}(x,m)\}_{m=1}^{|x|}$, where $\mathcal{A}(x,m)$ denotes the set of ambiguously specified heads for the mth token in x. Then, $\tilde{\boldsymbol{y}}(x)$ is defined as the set of all projective dependency trees spanning x that can be assembled from the arcs in $\mathcal{A}(x)$.

Methods for learning with ambiguous labelings have previously been proposed in the context of multi-class classification (Jin and Ghahramani, 2002), sequence-labeling (Dredze et al., 2009), log-linear LFG parsing (Riezler et al., 2002), as well as for discriminative reranking of generative constituency parsers (Charniak and Johnson, 2005). In contrast to Dredze et al., who allow for weights to be assigned to partial labels, we assume that the ambiguous arcs are weighted uniformly. For target language adaptation, these weights would typically be derived from unreliable sources and we do not want to train the model to simply mimic their beliefs. Furthermore, with this assumption, learning is simply achieved by maximizing the *marginal* log-likelihood of the ambiguous training set $\widetilde{\mathcal{D}} = \{(x_i, \tilde{\boldsymbol{y}}(x_i)\}_{i=1}^{n}$,

$$\mathcal{L}(\theta; \widetilde{\mathcal{D}}) = \sum_{i=1}^{n} \log \left\{ \sum_{y \in \tilde{\boldsymbol{y}}(x_i)} p_\theta(y \mid x_i) \right\} - \lambda \left\| \theta \right\|_2^2 .$$

In maximizing the marginal log-likelihood, the model is free to distribute probability mass among the trees

in the ambiguous labeling to its liking, as long as the marginal log-likelihood improves. The same objective function is used by Riezler et al. (2002) and Charniak and Johnson (2005). A key difference is that in these works, the ambiguity is constrained through a supervised signal, while we use ambiguity as a way to achieve self-training, using the base-parser itself, or some other potentially noisy knowledge source as the sole constraints. Note that we have introduced an ℓ_2-regularizer, weighted by λ. This is important as we are now training *lexicalized* target language models which can easily overfit. In all experiments, we optimize parameters with L-BFGS. Note also that the marginal likelihood is non-concave, so that we are only guaranteed to find a local maximum.

5.2 Ambiguity-Aware Self-Training

In standard self-training — hereafter referred to as *Viterbi self-training* — a base parser is used to label each unlabeled sentence with its most probable parse tree to create a self-labeled data set, which is subsequently used to train a supervised parser. There are two reasons why this simple approach may work. First, if the base parser's errors are not too systematic and if the self-training model is not too expressive, self-training can reduce the variance on the new domain. Second, self-training allows for features in the new domain with low support — or no support in the case of lexicalized features — in the base parser to be "filled in" by exploiting correlations in the feature representation. However, a potential pitfall of this approach is that the self-trained parser is encouraged to blindly mimic the base parser, which leads to error reinforcement. This may be particularly problematic when relexicalizing a transfer parser, since the lexical features provide the parser with increased power and thereby an increased risk of overfitting to the noise. To overcome this potential problem, we propose an *ambiguity-aware self-training* (AAST) method that is able to take the noise of the base parser into account.

We use the arc-marginals of the base parser to construct the ambiguous labeling $\tilde{\boldsymbol{y}}(x)$ for a sentence x. For each token $m \in [1, |x|]$, we first sort the set of arcs in which m is the dependent, $\{(h, m)\}_{h=0}^{|x|}$, by the marginal probabilities of the arcs:

$$p_\theta(h, m \mid x) = \sum_{\{y \in \mathcal{Y}(x) \mid (h,m) \in y\}} p_\theta(y \mid x)$$

We next construct the ambiguous arc set $\mathcal{A}(x, m)$ by adding arcs (h, m) in order of decreasing probability, until their cumulative probability exceeds σ, i.e. until

$$\sum_{(h,m) \in \mathcal{A}(x,m)} p_\theta(h, m \mid x) \;\geq\; \sigma\,.$$

Lower values of σ result in more aggressive pruning, with $\sigma = 0$ corresponding to including no arc and $\sigma = 1$ corresponding to including all arcs. We always add the highest scoring tree $\hat{y}$ to $\tilde{\boldsymbol{y}}(x)$ to ensure that it contains at least one complete projective tree.

Figure 3 outlines an example of how (and why) AAST works. In the Greek example, the genitive phrase Η παραμονή σκαφών (*the stay of vessels*) is incorrectly analyzed as a flat noun phrase. This is not surprising given that the base parser simply observes this phrase as DET NOUN NOUN. However, looking at the arc marginals we can see that the correct analysis is available during AAST, although the actual marginal probabilities are quite misleading. Furthermore, the genitive noun σκαφών also appears in other less ambiguous contexts, where the base parser correctly predicts it to modify a noun and not a verb. This allows the training process to add weight to the corresponding lexical feature pairing σκαφών with a noun head and away from the feature pairing it with a verb. The resulting parser correctly predicts the genitive construction.

5.3 Ambiguity-Aware Ensemble-Training

While ambiguous labelings can be used as a means to improve self-training, any information that can be expressed as hard arc-factored constraints can be incorporated, including linguistic expert knowledge and annotation projected via bitext. Here we explore another natural source of information: the predictions of other transfer parsers. It is well known that combining several diverse predictions in an ensemble often leads to improved predictions. However, in most ensemble methods there is typically no learning involved once the base learners have been trained (Sagae and Lavie, 2006). An exception is the method of Sagae and Tsujii (2007), who combine the outputs of many parsers on unlabeled data to train a parser for a new domain. However, in that work the learner is not exposed to the underlying ambiguity of the base parsers; it is only given the Viterbi parse of the combination system as the gold standard. In contrast,

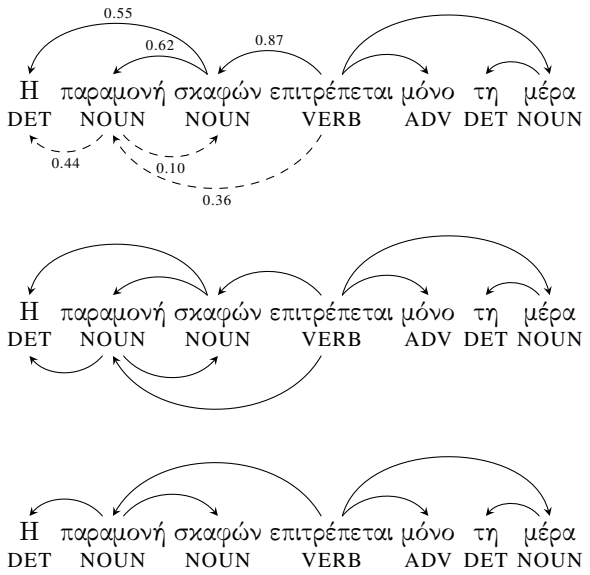

Figure 3: An example of ambiguity-aware self-training (AAST) on a sentence from the Greek self-training data. The sentence roughly translates to *The stay of vessels is permitted only for the day*. **Top:** Arcs from the base model's Viterbi parse are shown above the sentence. When only the part-of-speech tags are observed, the parser tends to treat everything to the left of the verb as a head-final noun phrase. The dashed arcs below the sentence are the arcs for the true genitive construction *stay of vessels*. These arcs and the corresponding incorrect arcs in the Viterbi parse are marked with their marginal probabilities. **Middle:** The ambiguous labeling $\tilde{\boldsymbol{y}}(x)$, which is used as supervision in AAST. Additional non-Viterbi arcs are present in $\tilde{\boldsymbol{y}}(x)$; for clarity, these are not shown. When learning with AAST, probability mass will be pushed towards any tree consistent with $\tilde{\boldsymbol{y}}(x)$. Marginal probabilities are ignored at this stage, so that all arcs in $\tilde{\boldsymbol{y}}(x)$ are treated as equals. **Bottom:** The Viterbi parse of the AAST model, which has selected the correct arcs from $\tilde{\boldsymbol{y}}(x)$.

we propose an *ambiguity-aware ensemble-training* (AAET) method that treats the union of the ensemble predictions for a sentence x as an ambiguous labeling $\tilde{\boldsymbol{y}}(x)$. An additional advantage of this approach is that the ensemble is compiled into a single model and therefore does not require multiple models to be stored and used at runtime.

It is straightforward to construct $\tilde{\boldsymbol{y}}(x)$ from multiple parsers. Let $\mathcal{A}_k(x, m)$ be the set of arcs for the mth token in x according to the kth parser in the ensemble. When arc-marginals are used to construct the ambiguity set, $|\mathcal{A}_k(x, m)| \geq 1$, but when the Viterbi-parse is used, $\mathcal{A}_k(x, m)$ is a singleton. We next form $\mathcal{A}(x, m) = \bigcup_k \mathcal{A}_k(x, m)$ as the ensemble arc ambiguity set from which $\tilde{\boldsymbol{y}}(x)$ is assembled. In this study, we combine the arc sets of two base parsers: first, the arc-marginal ambiguity set of the base parser (§5.2); and second, the Viterbi arc set from the *NBG* parser of Naseem et al. (2012) in Table 2.[4] Thus, the latter will have singleton arc ambiguity sets, but when combined with the arc-marginal ambiguity sets of our base parser, the result will encode uncertainty derived from both parsers.

5.4 Adaptation Experiments

We now study the different approaches to target language adaptation empirically. As in Naseem et al. (2012), we use the CoNLL training sets, stripped of all dependency information, as the unlabeled target language data in our experiments. We use the *Family* model as the base parser, which is used to label the unlabeled target data with the Viterbi parses as well as with the ambiguous labelings. The final model is then trained on this data using standard lexicalized features (McDonald et al., 2005). Since labeled training data is unavailable in the target language, we cannot tune any hyper-parameters and simply set $\lambda = 1$ and $\sigma = 0.95$ throughout. Although the latter may suggest that $\tilde{\boldsymbol{y}}(x)$ contains a high degree of ambiguity, in reality, the marginal distributions of the base model have low entropy and after filtering with $\sigma = 0.95$, the average number of potential heads per dependent ranges from 1.4 to 3.2, depending on the target language.

The ambiguity-aware training methods, that is ambiguity-aware self-training *(AAST)* and ambiguity-aware ensemble-training *(AAET)*, are compared to three baseline systems. First, *NBG+EM* is the generative model of Naseem et al. (2012) trained with expectation-maximization on additional unlabeled target language text. Second, *Family* is the best discriminative model from the previous section. Third, *Viterbi* is the basic Viterbi self-training model. The results of each of these models are shown in Table 3.

There are a number of things that can be observed. First, *Viterbi* self-training helps slightly on average, but the gains are not consistent and there are even drops in accuracy for some languages. Second, *AAST* outperforms the *Viterbi* variant on all languages and

[4] We do not have access to the marginals of *NBG*.

			Target Adaptation		
Lang.	NBG+EM	Family	Viterbi	AAST	AAET
ar	**59.3**	52.7	52.6	53.5	58.7
bg	67.0	65.4	66.4	67.9	**73.0**
ca	71.7	77.6	78.0	**79.9**	76.1
cs	44.3	43.5	43.6	44.4	**48.3**
de	54.1	59.2	59.7	**62.5**	61.5
el	67.9	63.2	64.5	65.5	**69.6**
es	62.0	67.1	68.2	**68.5**	66.9
eu	47.8	46.8	47.5	48.6	**49.4**
hu	58.6	64.5	64.6	65.6	**67.5**
it	65.6	72.5	71.6	72.4	**73.4**
ja	64.1	65.9	65.7	68.8	**72.0**
nl	56.6	56.8	57.9	58.1	**60.2**
pt	75.8	78.4	79.9	**80.7**	79.9
sv	61.7	63.5	63.4	**65.5**	**65.5**
tr	59.4	59.4	59.5	64.1	**64.2**
zh	51.0	54.8	54.8	57.9	**60.7**
avg	60.4	62.0	62.4	64.0	**65.4**

Table 3: Target language adaptation using unlabeled target data. *AAST*: ambiguity-aware self-training. *AAET*: ambiguity-aware ensemble-training. Boldface numbers indicate the best result per language. Underlined numbers indicate the best result, excluding *AAET*. *NBG+EM* is the "D+,T_o" model from Naseem et al. (2012).

nearly always improves on the base parser, although it sees a slight drop for Italian. *AAST* improves the accuracy over the base model by 2% absolute on average and by as much as 5% absolute for Turkish. Comparing this model to the *NBG+EM* baseline, we observe an improvement by 3.6% absolute, outperforming it on 14 of the 16 languages. Furthermore, ambiguity-aware self-training appears to help more than expectation-maximization for generative (unlexicalized) models. Naseem et al. observed an increase from 59.3% to 60.4% on average by adding unlabeled target language data and the gains were not consistent across languages. *AAST*, on the other hand, achieves consistent gains, rising from 62.0% to 64.0% on average. Third, as shown in the rightmost column of Table 3, ambiguity-aware ensemble-training is indeed a successful strategy; *AAET* outperforms the previous best self-trained model on 13 and *NB&G+EM* on 15 out of 16 languages. The relative error reduction with respect to the base *Family* model is 9% on average, while the average reduction with respect to *NBG+EM* is 13%.

Before concluding, two additional points are worth making. First, further gains may potentially be achievable with feature-rich discriminative models. While the best generative transfer model of Naseem et al. (2012) approaches its upper-bounding supervised accuracy (60.4% vs. 67.1%), our relaxed self-training model is still far below its supervised counterpart (64.0% vs. 84.1%). One promising statistic along these lines is that the oracle accuracy for the ambiguous labelings of *AAST* is 75.7%, averaged across languages, which suggests that other training algorithms, priors or constraints could improve the accuracy substantially. Second, relexicalization is a key component of self-training. If we use delexicalized features during self-training, we only observe a small average improvement from 62.0% to 62.1%.

6 Conclusions

We contributed to the understanding of multi-source syntactic transfer in several complementary ways. First, we showed how selective parameter sharing, based on typological features and language family membership, can be incorporated in a discriminative graph-based model of dependency parsing. We then showed how ambiguous labelings can be used to integrate heterogenous knowledge sources in parser training. Two instantiations of this framework were explored. First, an ambiguity-aware self-training method that can be used to effectively relexicalize and adapt a delexicalized transfer parser using unlabeled target language data. Second, an ambiguity-aware ensemble-training method, in which predictions from different parsers can be incorporated and further adapted. On average, our best model provides a relative error reduction of 13% over the state-of-the-art model of Naseem et al. (2012), outperforming it on 15 out of 16 evaluated languages.

Acknowledgments We thank Alexander Rush for help with the hypergraph framework used for inference. Tahira Naseem kindly provided us with her data sets and the predictions of her systems. This work benefited from many discussions with Yoav Goldberg and members of the Google parsing team. We finally thank the three anonymous reviewers for their valuable feedback. The work of the first author was partly funded by the Swedish National Graduate School of Language Technology (GSLT).

References

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of ACL*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL*.

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

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

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Mark Dredze, Partha Pratim Talukdar, and Koby Crammer. 2009. Sequence learning from data with multiple labels. In *Proceedings of the ECML/PKDD Workshop on Learning from Multi-Label Data*.

Matthew S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. http://wals.info/.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of EMNLP-CoNLL*.

Jason Eisner. 1996. Three new probabilistic models for dependency parsing: an exploration. In *Proceedings of COLING*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical Bayesian domain adaptation. In *Proceedings of HLT-NAACL*.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of ACL-IJCNLP*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Rong Jin and Zoubin Ghahramani. 2002. Learning with multiple labels. In *Proceedings of NIPS*.

Dan Klein and Chris D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of ACL*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of ACL*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of EMNLP*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of ACL*.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of ACL*.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *Proceedings of NAACL*.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of EMNLP-CoNLL*.

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

Anders Søgaard and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. In *Proceedings of COLING*.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of ACL*.

Kathrin Spreyer and Jonas Kuhn. 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In *Proceedings of CONLL*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL-HLT*.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure (WILS 2012)*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP Workshop: NLP for Less Privileged Languages*.