Target Language-Aware Constrained Inference for Cross-lingual Dependency Parsing

Tao Meng
University of California, Los Angeles
tmeng@cs.ucla.edu

Nanyun Peng
University of Southern California
npeng@isi.edu

Kai-Wei Chang
University of California, Los Angeles
kw@kwchang.net

Abstract

Prior work on cross-lingual dependency parsing often focuses on capturing the commonalities between source and target languages and overlooks the potential of leveraging linguistic properties of the languages to facilitate the transfer. In this paper, we show that weak supervisions of linguistic knowledge for the target languages can improve a cross-lingual graph-based dependency parser substantially. Specifically, we explore several types of *corpus linguistic statistics* and compile them into *corpus-wise constraints* to guide the inference process during the test time. We adapt two techniques, Lagrangian relaxation and posterior regularization, to conduct inference with corpus-statistics constraints. Experiments show that the Lagrangian relaxation and posterior regularization inference improve the performances on 15 and 17 out of 19 target languages, respectively. The improvements are especially significant for target languages that have different word order features from the source language.

1 Introduction

Natural language processing (NLP) techniques have achieved remarkable performance in a variety of tasks when sufficient training data is available. However, obtaining high-quality annotations for low-resource language tasks is challenging, and this poses great challenges to process low-resource languages. To bridge the gap, cross-lingual transfer has been proposed to transfer models trained on high-resource languages (e.g., English) to low-resource languages (e.g., Tamil) to combat the resource scarcity problem. Recent studies have demonstrated successes of transferring models across languages without retraining for NLP tasks, such as named entity recognition (Xie et al., 2018), dependency parsing (Tiedemann, 2015; Agić et al., 2014), and question

answering (Joty et al., 2017), using a shared multi-lingual word embedding space (Smith et al., 2017) or delexicalization approaches (Zeman and Resnik, 2008; McDonald et al., 2013).

One key challenge for cross-lingual transfer is the *differences* among languages; for example, languages may have different word orders. When transferring a model learned from a source language to target languages, the performance may drop significantly due to the differences. To tackle this problem, various approaches have been proposed to better capture the *commonalities* between the source and the target languages (McDonald et al., 2011; Guo et al., 2016; Täckström et al., 2013; Agić, 2017; Ahmad et al., 2019); however, they overlook the potential to leverage linguistic knowledge about the target language to account for the *differences* between the source and the target languages to facilitate the transfer.

In this paper, we propose a complementary approach that studies how to leverage the linguistic knowledge about the target languages to help the transfer. Specifically, we use corpus linguistic statistics of the target languages as weak supervision signals to guide the test-time *inference* process when parsing with a graph-based parser. This approach is effective as the model only need to be trained *once* on the source language and applied to many target languages using different constraints *without retraining* the model.

We argue that certain corpus linguistic statistics such as the word order (e.g., how often an adjective appears before or after a noun) can be easily obtained from available resources such as World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). To incorporate the corpus linguistic statistics to a cross-lingual parser, we compile them into corpus-wise constraints and adopt two families of methods: 1) Lagrangian relaxation (LR) and 2) posterior regularization (PR) to

1117

solve the constrained inference problem. The algorithms take the original graph-based parsing inference as a sub-routine, and LR iteratively adjusts the pair-wise potentials until the constraints are (loosely) satisfied, while PR finds a feasible distribution and do inference based on that. The constrained inference framework is general and supports any knowledge that can be formulated as a first-order logic (Roth and Yih, 2004).

We evaluate the proposed approach under the single-source transfer setting using English as the source language and test on 19 target languages covering a broad range of language families with low-resource languages such as Tamil and Welsh. We demonstrate that by adding three simple corpus-wise constraints derived from WALS features, the performances improve in 15 and 17 out of 19 languages when using Lagrangian relaxation and posterior regularization techniques, respectively. The improvements are especially substantial when the target language features are distant from the source language. For example, our framework improves the UAS score of Urdu by 15.7%, and Tamil by 7.3%.[1]

2 Constrained Cross-Lingual Parsing

Our work focuses on the graph-based dependency parser (McDonald et al., 2005) in the zero-shot single-source transfer setting as in Ahmad et al. (2019). However, the proposed algorithms can be extended other transfer settings. Given a trained model, we derive corpus-statistics constraints and apply them to correct errors caused by word order differences between the source and the target language *during the inference time*. Figure 1 shows an example of how constraints can influence the inference results.

In this section, we first give a quick review of the graph-based parser and introduce the notations. We then discuss how to formulate corpus-wise constraints based on corpus linguistic statistics for guiding the graph-based parser.

2.1 Background: Graph-Based Parser

A graph-based parser learns a scoring function for every pair of words in a sentence and conducts inference to derive a directed spanning tree with the highest accumulated score. Formally, given the k-th sentence $\mathbf{w}_k = (w_{k1}, \ldots, w_{kL(k)})$ where $L(k)$

denotes the length of the k-th sentence, a graph-based parser learns a score matrix $S^{(k)}$, where $S_{ij}^{(k)}$ denotes the score to form an arc from word w_{ki} to word w_{kj}. Let y_k be an indicator function that $y_k(i, j) \in \{0, 1\}$ denotes the arc from w_{ki} to w_{kj}. The maximum directed spanning tree inference can be formulated as an integer linear programming (ILP) problem:

$$y_k^* = \arg \max_{y_k \in \mathcal{Y}_k} \sum_{i,j} S_{ij}^{(k)} y_k(i, j), \qquad (1)$$

where $\mathcal{Y}_k$ is the set of legal dependency trees of sentence k. In recent years, neural network approaches (Kiperwasser and Goldberg, 2016; Wang and Chang, 2016; Kuncoro et al., 2016; Dozat and Manning, 2017) have been applied to modeling the scoring matrix $S^{(k)}$ and have achieved great performance in dependency parsing.

From the probabilistic point of view, if we assume for different i, j, the edge probabilities $P(y_k(i, j) = 1 | \mathbf{w}_k)$ are mutually conditional independent, the probability of a whole parse tree can be written as

$$P(y_k | \mathbf{w}_k) = \prod_{i,j} P(y_k(i, j) = 1 | \mathbf{w}_k)^{y_k(i,j)}. \quad (2)$$

If we set $S_{ij}^{(k)} = \log P(y_k(i, j) = 1 | \mathbf{w}_k) + Z_j'$, where Z_j' is a constant term, then Eq. (1) can be regarded as the following maximum a posteriori (MAP) inference problem:

$$
\begin{aligned}
y_k^* &= \arg \max_{y_k \in \mathcal{Y}_k} P(y_k | \mathbf{w}_k) \\
&= \arg \max_{y_k \in \mathcal{Y}_k} \sum_{i,j} \log P(y_k(i, j) = 1 | \mathbf{w}_k) y_k(i, j).
\end{aligned}
$$
$$(3)$$

2.2 Corpus-Wise Constraints

Given the inference problems as in equations (1) and (3), additional constraints can be imposed to incorporate expert knowledge about the languages to help yield a better parser. Instance-level constraints have been explored in the literature of dependency parsing, both in the monolingual (Dryer, 2007) and cross-lingual transfer (Täckström et al., 2013) settings. However, most word order features for a language are non-deterministic and cannot be compiled into instance-wise constraints.

In this work, we introduce corpus-wise constraints to leverage the non-deterministic features for cross-lingual parser. We compile the following two types of corpus-wise constraints based on corpus linguistics statistics:

1118

- *Unary* constraints consider statistics regarding a particular POS tag (POS).
- *Binary* constraints consider statistics regarding a pair of POS tags (POS_1, POS_2).

Specifically, a unary constraint specifies the ratio r of the heads of a particular POS appears on the left of that POS.[2] Similarly, a binary constraint specifies the ratio r of POS_1 being on the left of POS_2 when there is an arc between POS_1 and POS_2.

The ratios r for the constraints are called corpus statistics, which can be estimated in one of the following ways: a) leveraging existing linguistics resources or consulting linguists; b) leveraging a higher-resource language that is similar to the target language (e.g., Finnish and Estonian) to collect the statistics. In this paper, we explore the first option and leverage the WALS features, which provide a reference for word order typology, to estimate the ratios.

Compile Constraints From WALS Features. For a particular language, once we collect the corpus-statistics of a pair of POS tags, we can formulate a binary constraint. There are different ways to estimate the corpus-statistics. For example, Östling (2015) utilizes a small amount of parallel data to estimate the dominant word orders. In this paper, we simply utilize a small subset of WALS features that show the dominant order of some POS pairs (e.g. adjective and noun) in a language. They can be directly compiled into binary constraints.

Similarly, we can estimate the ratio for unary constraints based on WALS features. For a particular POS tag, we choose all WALS features related to it to formulate a feature vector f. The mapping from the vector f to the unary constraint ratio r is learnable: for each language with annotated data, we can get a WALS feature vector f_{lang} and a ratio r_{lang} from the annotation. We only need a small amount of data to estimate r_{lang} well. Given a set of languages with feature vectors and estimated ratios, we can learn the mapping by a simple linear regression, and apply it to estimate the ratio of any target language to compile a unary constraint.

2.3 Formulate Constraints

In the following, we mathematically formulate the corpus-wise constraints. Note that these con-
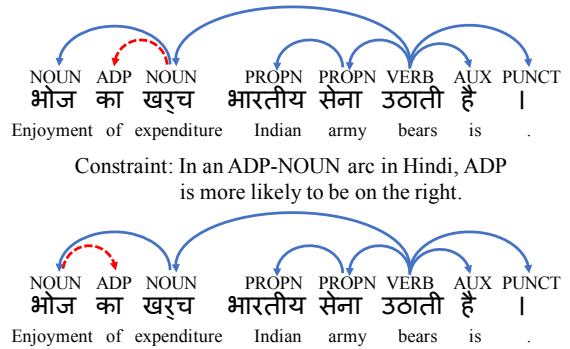
Figure 1: An running example of Hindi. On the top there is the inference result of a baseline model trained on English. In English, ADP is mostly on the left of the NOUN, so the potential of the correct ADP-NOUN arc is lower. With the help of corpus-statistics constraints, the potential is adjusted and the model gets correct inference result as shown in the bottom. The dashed lines highlight the difference.

straints are based on the statistics over the entire corpus. For a unary constraint $C_u^{(POS)}$, let P denotes a set of word with part-of-speech tag POS. We define $C_u^+ := \{(k, i, j) | w_{kj} \in P \wedge i < j\}$ as the set of arcs where the head of word in P is on its left and $C_u^- := \{(k, i, j) | w_{kj} \in P \wedge i > j\}$, conversely.

For a binary constraint $C_b^{(POS_1, POS_2)}$, we denotes P_1, P_2 as a set of word with part-of-speech tag POS_1, POS_2, respectively. We then define C_b^+ as the set of arcs with two ends $w_{ki} \in P_1$ and $w_{kj} \in P_2$, and w_{ki} is on the left of w_{kj}. We define C_b^- Similarly. Formally,

$$
\begin{aligned}
C_b^+ :=&\{(k,i,j)|w_{ki} \in P_1 \wedge w_{kj} \in P_2 \wedge i < j\} \\
&\cup \{(k,j,i)|w_{ki} \in P_1 \wedge w_{kj} \in P_2 \wedge i < j\}, \\
C_b^- :=&\{(k,i,j)|w_{ki} \in P_1 \wedge w_{kj} \in P_2 \wedge i > j\} \\
&\cup \{(k,j,i)|w_{ki} \in P_1 \wedge w_{kj} \in P_2 \wedge i > j\}.
\end{aligned}
$$

For notational simplicity, we use C to denote all constraints including both the unary and binary ones. The ratio function $R(C, Y)$ for a constraint C given the parse trees Y can be defined as:

$$
R(C,Y) = \frac{\sum_k \sum_{(i,j):(k,i,j) \in C^+} y_k(i,j)}{\sum_k \sum_{(i,j):(k,i,j) \in C^+ \cup C^-} y_k(i,j)}.
$$

We want to enforce the ratio $R(C, Y)$ estimated from Y to be consistent with a value r (see Sec. 2.2), which formulates a constraint

$$
r - \theta \leq R(C, Y) \leq r + \theta
$$

where θ is a tolerance margin. Note that the instance-level hard constraint is a special case of
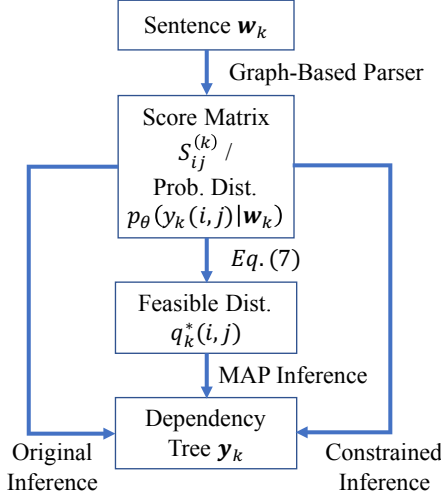
Figure 2: The pipelines of the baseline method (left), Lagrangian relaxation (right) and posterior regularization (middle). Lagrangian relaxation converts constrained inference to an unconstrained optimization problem using Lagrange's method. Posterior regularization method is working on the distribution space. For a given distribution, PR finds the closest feasible distribution and conduct MAP inference.

the corpus-statistics constraint when $r = 0$ or $r = 1$.

Given a set of corpus-statistics constraints $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ with corresponding corpus statistics $\mathbf{r} = \{r_1, r_2, \ldots, r_n\}$, the objective of the constrained inference is:

$$\max_{Y \in \mathcal{Y}} \quad \sum_k \sum_{i,j \leq L(k)} S_{ij}^{(k)} y_k(i,j), \tag{4}$$
$$\text{s.t.} \quad r_i - \theta_i \leq R(C_i, Y) \leq r_i + \theta_i, i \in [N],$$

where $\mathcal{Y}$ denotes the set of all possible dependency trees. As all the constraints can be written as a linear inequality with respect to $y_k(i,j)$. Eq. (4) is an ILP.

3 Inference with Corpus-Statistics Constraints

The ILP problem in Eq. (4) is in general an NP-hard problem; especially, it involves variables associated with the entire corpus. Without constraints, Eq. (4) can be decoupled into K sub-problems, and the inference with respect to each sentence can be solved independently as in Eq. (1). In this way, an efficient inference algorithm such as maximum directed spanning tree algorithm (Chu and Liu, 1965) can be used.

However, with the corpus-wise constraints, directly solving Eq. (4) is infeasible. Therefore, we explore two algorithms for inference with corpus-statistics constraints: Lagrangian relaxation and

posterior regularization (Ganchev et al., 2010). The Lagrangian relaxation algorithm introduces Lagrangian multipliers to relax the constraint optimization problem to an unconstrained optimization problem, and estimates the Lagrangian multipliers with gradient-based methods. The posterior regularization algorithm uses the constraints of the target language to define a feasible set of parse tree distributions, and find a feasible distribution that is closest to the parse tree distribution trained on the source language by minimizing the KL-divergence. The constrained inference problem can then be converted into an MAP inference problem on the best feasible distribution. Figure 2 illustrates the procedure of the original inference, Lagrangian relaxation, and posterior regularization.

3.1 Lagrangian Relaxation

Lagrangian relaxation has been applied in various NLP applications (Rush and Collins, 2012, 2011). In Eq. (4), each constraint C_i involves two inequality constraints: $R(C_i, Y) - r_i + \theta \geq 0$, and $r_i + \theta - R(C_i, Y) \geq 0$. Instead of treating these two constraints separately, we consider a heuristic to optimize with equality constraints $R(C_i, Y) = r_i, i \in [N]$ and terminate earlier when constraints in Eq. (4) are satisfied. Despite this approach does not guarantee the solution is optimal if all the constraints are satisfied as the original Lagrangian relaxation algorithm does, in practice, the inference converges faster (as the number of Lagrangian multipliers is half) and the parsing performance maintains.

In the following, we derive the constrained inference algorithm for corpus-statistics constraints. First, we rewrite the equality constraint $R(C, Y) = r$ by substituting $R(C, Y)$ with Eq. (2.3):

$$(1-r) \sum_k \sum_{(i,j):(k,i,j) \in C_i^+} y_k(i,j)$$
$$-r \sum_k \sum_{(i,j):(k,i,j) \in C_i^-} y_k(i,j) = 0, \tag{5}$$

We use F(C) to denote the left-hand-side of Eq. (5), which is linear w.r.t. y_k. Then, the Lagrangian relaxation of the constrained inference problem can be written as:

$$L(Y, \lambda; \mathcal{C}) = \sum_k \sum_{i,j} S_{ij}^{(k)} y_k(i,j) + \sum_{i \in [N]} \lambda_i F(C_i),$$

Algorithm 1 Lagrangian Relaxation for Constraint Inference
___

Input: Constraints $\mathcal{C} = \{C_i\}_{i=1}^N$, corresponding ratio $\mathbf{r} = \{r_i\}_{i=1}^N$, tolerance margin $\theta = \{\theta_i\}_{i=1}^N$, learning rate decay η, initial learning rate α_0
Output: parse trees $\hat{Y}$
___
1: $\alpha \leftarrow \alpha_0$
2: $\lambda_i \leftarrow 0, \ i \in [N]$
3: **repeat**
4:    $\hat{Y} \leftarrow \arg\max_{Y \in \mathcal{Y}} L(Y, \lambda; \mathcal{C})$
5:    $\hat{r}_i \leftarrow R(C_i, \hat{Y}), \ i \in [N]$
6:    **if** $\forall i, abs(r_i - \hat{r}_i) \leq \theta$ **then**
7:       **return** $\hat{Y}$
8:    $\lambda_i \leftarrow \alpha(\hat{r}_i - r_i), \ i \in [N]$
9:    $\alpha \leftarrow \eta\alpha$
10: **until** MAX_ITER times
11: **return** $\hat{Y}$
___

where λ_i is called Lagrangian multiplier. It is well-known that we can solve the dual form of the constrained inference problem:

$$\max_{Y \in \mathcal{Y}} \min_{\lambda \geq 0} L(Y, \lambda; \mathcal{C}).$$

To solve the dual form, we initialize λ_i to be 0. At iteration t, we firstly conduct an constraint-augmented inference with a fixed $\lambda^{(t)}$:

$$\hat{Y}^{(t)} = \max_{Y \in \mathcal{Y}} L(Y, \lambda^{(t)}; \mathcal{C}). \quad (6)$$

As $F(C)$ is a linear function w.r.t y_k, we combine it with $S_{ij}^{(k)} y_k(i,j)$. In this way, the inference problem Eq. (6) can be treated as a special case of Eq. (1) with a different scoring matrix $S^{(k)}$. In this way, we can treat the inference on every sentence independently and leverage existing inference techniques.

After solving the constraint-augmented inference, we compute the ratio of every constraint $\hat{r}_i^{(t)} = R(C_i, \hat{Y}^{(t)})$, and use gradient ascent algorithm to update the Lagrangian multipliers

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \alpha^{(t)}(r_i - \hat{r}_i).$$

Here $\alpha^{(t)}$ denote the step size at iteration t. The algorithm is shown in Algorithm 1.

3.2 Posterior Regularization

From a probabilistic point of view, the parser model learns parameters θ to realize Eq. (2). During the inference, the model predict a probability distribution $p_\theta(\mathbf{Y}|\mathbf{W})$ over possible parse trees given a sentence. The posterior regularization algorithm first defines a feasible set of the probability distributions w.r.t the given constraints, and looks for the closest feasible distribution $q^*(\mathbf{Y})$ to the model distribution $p_\theta(\mathbf{Y}|\mathbf{W})$. The best parse tree is given by $\arg\max_{\mathbf{Y}} q^*(\mathbf{Y})$. Specifically, we define the feasible set as:

$$Q = \{q(\mathbf{Y}) | r_i - \theta_i \leq R(C_i, q(\mathbf{Y})) \leq r_i + \theta_i, i \in [N]\},$$

where $R(C, q) = \frac{\sum_k \sum_{(i,j):(k,i,j)\in C^+} q_k(i,j)}{\sum_k \sum_{(i,j):(k,i,j)\in C^+ \cup C^-} q_k(i,j)}$, with $q_k(i,j) = E_{Y \sim q(Y)}[y_k(i,j)]$.

To measure the distance between two distribution, we use the KL-divergence, and find the best feasible distribution $q^*(\mathbf{Y})$:

$$q^*(\mathbf{Y}) = \arg\min_{q \in Q} KL(q(\mathbf{Y}) \| p_\theta(\mathbf{Y}|\mathbf{W})). \quad (7)$$

If the feasible set has the expectation form:

$$\{q | \mathbb{E}_{\mathbf{Y} \sim q}[\phi(\mathbf{Y})] \leq \mathbf{b}\}, \quad (8)$$

Eq. (7) has a simple close form solution (Ganchev et al., 2010):

$$q^*(\mathbf{Y}) = \frac{p_\theta(\mathbf{Y}|\mathbf{W}) \exp(-\lambda^* \cdot \phi(\mathbf{Y}))}{Z(\lambda^*)}, \quad (9)$$

where λ^* is the solution of

$$\lambda^* = \arg\max_{\lambda \geq 0} -\mathbf{b} \cdot \lambda - \log Z(\lambda),$$
$$Z(\lambda) = \sum_{\mathbf{Y}'} p_\theta(\mathbf{Y}'|\mathbf{W}) \exp(-\lambda^* \cdot \phi(\mathbf{Y}')). \quad (10)$$

In the rest of this section, we first show that the feasible set Q we considered above can be reformulated in the form of Eq. (8), and then we discuss how to solve Eq. (10). To show that the inequality $R(C, q(\mathbf{Y})) \leq r$, in Q can be formulated in the form of Eq. (8), we set

$$\phi_C(\mathbf{Y}) = \sum_{k,i,j} y_k(i,j)\phi_C(k,i,j),$$

and

$$\phi_C(k,i,j) = \begin{cases} 1-r & (k,i,j) \in C^+ \\ -r & (k,i,j) \in C^- \\ 0 & (k,i,j) \notin C^+ \cup C^-. \end{cases} \quad (11)$$

Similarly, we can derive $R(C, q(\mathbf{Y})) \geq r$, into the same form and rewrite Q as

$$Q = \{q(\mathbf{Y}) \mid \mathbb{E}_{\mathbf{Y} \sim q} \phi(\mathbf{Y}) \leq \mathbf{0}\},$$

Algorithm 2 Posterior Regularization

Input: Constraints $\mathcal{C} = \{C_i\}_{i=1}^N$, corresponding ratio $\mathbf{r} = \{r_i\}_{i=1}^N$, tolerance margin $\theta = \{\theta_i\}_{i=1}^N$
Output: parse trees $\hat{Y}$

1: $p_\theta(y_k(i,j) = 1) \leftarrow$ normalize$\left(\exp\left(S_{ij}^{(k)}\right)\right)$
2: $\phi \leftarrow$ defined by Eq. (11)
3: $\lambda_i \leftarrow 0,\ i \in [N]$
4: **repeat**
5:    estimate $\mathbf{g} \leftarrow \partial \log Z(\lambda)/\partial \lambda$ in a batch
6:    update λ based on $\mathbf{g}$
7: **until** MAX_ITER times
8: **for** each (k, i, j) **do**
9:    $q_k(i,j) \leftarrow$ defined by Eq. (12)
10: $\hat{Y} \leftarrow$ MAP inference based on q
11: **return** $\hat{Y}$

where $\phi = (\phi_{C_1}, \phi_{C_2}, ..., \phi_{C_N})$ is a collection of the constraints. The detailed derivations can be found in Appendix A.

We solve Eq. (10) by sub-gradient descent[3]. Noting that there can be exponential number of terms in $Z(\lambda)$, we firstly need to factorize $Z(\lambda)$ from corpus level to instance level and arc level, and compute the gradient. The technical details are in Appendix A. With the optimal λ^*, we can compute the feasible distribution $q^*(\mathbf{Y})$ given p_θ and $\mathbf{W}$. Noting that the solution Eq. (9) can also be factorized to arc-level:

$$q_k^*(i,j) \propto p_\theta(y_k(i,j)|\mathbf{w_k}) \exp(-\lambda^* \cdot \phi(k,i,j)), \quad (12)$$

here $q_k^*(i,j)$ denote the arc-level distribution $q^*(y_k(i,j) = 1)$ satisfying Eq. (3). We then do MAP inference based on q, which is actually a minimal spanning tree problem same as before. Algorithm 2 summarizes the process.

4 Experiments

In this section, we evaluate the proposed algorithms by transferring an English dependency parser to 19 target languages covering 13 language families of real low-resource languages. We first introduce the experimental setup including data selection and constraint details and then discuss the results as well as in-depth analysis.

4.1 Setup

Model and Data We train the best performing Att-Graph parser proposed in Ahmad et al. (2019)

on English and transfer it to 19 target languages in UD Tree Bank v2.2 (Nivre et al., 2018).[4] The model takes words and predicted POS tags[5] as input, and achieve transfer by leveraging pre-trained multi-lingual FastText (Bojanowski et al., 2017) embeddings that project the word embeddings from different languages into the same space using an offline transformation method (Smith et al., 2017; Conneau et al., 2018). The SelfAtt-Graph model uses a Transformer (Vaswani et al., 2017) with relative position embedding as the encoder and a deep biaffine scorer (Dozat and Manning, 2017) as the decoder. We follow the setting in Ahmad et al. (2019) to train and tune only on the source language (English) and directly transfer to all the target languages. We modify their decoder to incorporate constraints with the proposed constrained inference algorithms during the transfer phase without retraining the model. All the hyper-parameters are specified in Appendix Table 4 together with hyper-parameters for the inference algorithms in Appendix Table 5.

Constraints We consider two types of constraints: 1) instance-level projective constraints for avoiding creating crossing arcs in the dependency trees, 2) corpus-statistics constraints constructed by the process described in Section 2.2. We consider the following three corpus-statistics constraints: $C1 = (NOUN)$, $C2 = (NOUN, ADP)$, $C3 = (NOUN, ADJ)$; intuitively, $C1$ concerns about the ratio of nouns being on the right of their heads; $C2$ concerns about the ratio of nouns being on the left of adpositions among all noun-adposition arcs; $C3$ concerns about the ratio of nouns being on the left of adjectives among all noun-adjective arcs.

For binary constraints, $C2$ and $C3$ can be directly compiled from WALS feature $85A$ and $87A$ respectively. We encode "dominant order" specified in WALS as the ratio being always greater than 0.75 (i.e., $r = 0.875$ and $\theta = 0.125$). If there is no dominant order or the feature is missing, we set $r = 0.5$ and $\theta = 0.25$. Some WALS features like $82A, 83A$ are also about word order, but we need to specify the arc types to utilize them. For simplicity, we only consider forming constraints from the POS tags in this paper. To estimate the

---

[3]In implementation, we use stochastic gradient descent with Adam optimizer (Kingma and Ba, 2015)

[4]We make the selection to prioritize the coverage of language families and low resource languages. The language family information can be found in Table 1.

[5]We use predicted POS tags provided in UD v2.2.

Family	Lang.	Features	Baseline	Lagrangian Relaxation			Posterior Regularization		
				Oracle	WALS	△WALS	Oracle	WALS	△WALS
IE.Germanic	en	1,1,1	90.5	90.3	90.4	-0.1	90.4	90.6	+0.1
IE.Indic	ur	-1,-1,1	18.3	35.2	34.0	+15.7	35.0	33.7	+15.4
IE.Indic	hi	-1,-1,1	34.3	52.4	53.4	+19.1	51.3	49.1	+14.8
Dravidian	ta	-1,-1,1	36.1	42.8	43.4	+7.3	43.1	43.0	+6.9
Turkic	tr	-1,-1,1	31.2	35.2	37.1	+5.9	35.1	36.3	+5.1
Afro-Asiatic	ar	1, 1, -1	38.5	47.3	45.3	+6.8	45.8	43.7	+5.2
Afro-Asiatic	he	1, 1, -1	55.7	58.8	57.6	+1.9	58.3	57.6	+1.9
Austronesian	id	1, 1, -1	49.3	53.1	52.3	+3.0	52.3	51.9	+2.6
Korean	ko	-1,-1,1	34.0	37.1	37.2	+3.2	36.3	36.4	+2.4
IE.Celtic	cy	1, 1, -1	47.3	54.2	51.7	+4.4	53.8	50.0	+2.7
IE.Romance	ca	1, 1, -1	73.9	74.9	73.8	-0.1	74.9	74.7	+0.8
IE.Romance	fr	1, 1, -1	77.8	79.1	78.7	+0.9	79.0	79.0	+1.2
Uralic	et	1, -1,1	65.3	65.5	65.8	+0.5	65.7	66.0	+0.7
Uralic	fi	1, -1,1	66.7	67.1	67.0	+0.3	66.9	67.1	+0.4
IE.Slavic	hr	1, 1, 1	62.2	63.7	63.2	+1.0	63.6	63.4	+1.2
IE.Slavic	bg	1, 1, 1	79.6	79.7	79.2	+0.0	79.7	79.7	+0.1
IE.Baltic	lv	1, 1, 1	70.3	70.7	69.5	-0.8	70.5	69.9	-0.4
IE.Latin	la	?, ?, ?	47.4	48.0	45.6	-1.8	48.1	47.3	-0.1
IE.Germanic	da	1, 1, 1	76.6	76.6	76.5	-0.1	76.6	76.6	+0.0
IE.Germanic	nl	0, 1, 1	67.5	67.6	67.5	+0.0	67.9	67.9	+0.4
Average Performance			54.3	58.4	57.8	+3.5	58.1	57.5	+3.1

Table 1: Cross-lingual transfer performances for dependency parsing on 19 languages from 13 different families, with the performance on the source language (English) as a reference. Performances are reported per UAS (we observe similar trends for LAS and details can be found in the appendix Table 7). We compare the baseline model (Ahmad et al., 2019) with our two algorithms (Lagrangian relaxation and posterior regularization) considering the oracle constraints, and the corpus-statistics constraints compiled from WALS. Columns △WALS denote the improvements bring by leveraging WALS feature as constraints. We also create a Features column to show three WALS features [83A,85A,87A] for each language. The values {1, -1, 0, ?} stand for the same as English, opposite to English, no dominant order, and feature missing, respectively.

ratio for unary constraint $C1$, we use the WALS features $82A, 83A, 85A, 86A, 87A, 88A, 89A$ that are related to $NOUN$ to form feature vectors, and do regression on languages in the test set except the target language to predict the constraint ratio. The process guarantees the target language remain unseen during the ratio estimation process. The ratios on the regression training languages are estimated by sampling 100 sentences in the training set per language.

We also consider an oracle setting where we collect a "ground-truth" ratio of each constraint for the target language to estimate an upper bound of our inference algorithms. In the oracle setting, we estimate the ratio on the whole training corpus of the target language and set the margin to $\theta = 0.01$.

4.2 Parsing Performances

We first compare the performances of the cross-lingual dependency parser with or without con-

straints. Table 1 illustrates the results for the 19 target languages we selected,[6] along with the performance on the source language (English). The performance on English is not as high as the dependency parsers specialized for English, because to achieve transfer, we have to freeze the pre-trained multi-lingual word embeddings. Yet this parser achieved the best single-source transfer performances according to Ahmad et al. (2019).

As is shown in Table 1, the improvements by our constrained inference algorithms are dramatic in a few languages that have very distinct word order features from the source language. For example, the parsing performance of Hindi (hi) improves about 15% in UAS with WALS features via both Lagrangian relaxation and posterior regularization inference. The improvements are less

---

[6]We also run on all languages in Ahmad et al. (2019) for completeness and observe similar trends. The results can be found in Appendix Table 7.

obvious for languages that are in the same family as English such as Danish(da) and Dutch(nl). This is expected as the corpus linguistic statistics of these languages are similar to English thus the constraints are mostly satisfied with the baseline parser. Comparing Lagrangian relaxation and posterior regularization, we find posterior regularization being more robust and less sensitive to the errors in the corpus-statistics estimation, while Lagrangian relaxation gives a higher improvement on average. Overall, the two proposed constrained inference algorithms improved the transfer performance by 3.5% and 3.1% per UAS on average on 19 target languages.

For languages like Finnish (fi) and Estonian (et), the WALS setting works even better than the oracle. We suspect the reason being the large margin we set in the WALS setting. When the estimated corpus-statistics is different from the real ratio in the test set, the large margin relaxes the constraints, thus could result in better performances.

Discussion. Despite the major experiments and analysis are conducted using English as the only source language, our approach is general and does not have restriction on the choice of the source language(s). To verify this claim, we run experiments with Hebrew as the source language. Under the oracle setting, Lagrangian relaxation and posterior regularization improve the baseline by 4.4% and 4.1%, respectively.

We observed that if we compile WALS features into hard constraints (i.e., set $r = 0$ or 1), the constraint inference framework only improves performance on half of the languages. For example, in Estonian (et), the performance *drops* about 3%. This is because WALS only provides the dominant order. Therefore, treating WALS as hard constraints introduces error to the inference.

Finally, we assume if we can access to native speakers, the corpus-statistics can be estimated by a few partial annotations of parse trees. In our simulation, using less than 300 arcs, we can achieve the same performance as using the oracle.

4.3 Contributions of Individual Constraints

We analyze the contribution of each constraint demonstrated in Table 2. Here we use the oracle setting to reduce the noise introduced by corpus-statistics estimation errors. The results are based on Lagrangian relaxation inference. As shown in Table 2, Despite some languages have

Model	UAS	coverage	Δ
baseline	54.3	N/A	N/A
+Proj.	54.6	N/A	+0.3
+Proj.+C1	57.0	0.24	+2.4
+Proj.+C2	55.7	0.08	+1.1
+Proj.+C3	55.0	0.07	+0.4
oracle	58.4	N/A	+4.1

Table 2: Ablation study: average UAS of baseline model with different sets of constraints. Proj. represent projective constraints. C1-C3 and oracle are introduced in 4.1. The improvements for projective constraint and oracle are compared to baseline. For the other three constraint sets the improvement is compared to model with projective constraint.

Const.	statistics	improvement
+Proj.	N/A	+0.1
C1	0.30/0.36/0.94	+6.9
C2	0.00/0.06/1.00	+11.3
C3	0.14/0.27/0.12	+0.5
All	N/A	+18.1

Table 3: Contribution of individual constraints and their statistics in Hindi. The second column lists the ratios estimated from oracle in English/ baseline in Hindi/ oracle in Hindi, respectively. The improvement is measured in UAS. The improvement of constraints is computed same as Table 2

non-projective dependencies, we observed performance improvements on almost all the languages when the projective constraint is enforced. All the constraints we formulated have positive contributions to the performance improvements. $C1 = (NOUN)$ brings the largest gain probably because its widest coverage.

Table 1 shows that the performance of Hindi improves from 34% to over 51% per UAS for both inference algorithms. To better understand where the improvements come from, we conduct an analysis to breakdown the contribution of each individual constraint for Hindi. Table 3 shows the results. We can see that since the corpus linguistic statistics between Hindi and English are distinct, the baseline model only achieves low performance. With the constrained inference, especially the postposition constraint (C2), the proposed inference algorithm bring significant improvement.

To verify the effectiveness of the constraints, we analyze the relation between the performance improvements and corpus statistics ratio gaps between the source and the target languages. To quantify the ratio gap, we weight constraints by

their coverage rate and compute the weighted average of the ratio difference between source and target languages. Results show that the performance improvement is highly related to the ratio gap. The Pearson Correlation Coefficient is **0.938**. The figure showing the correlation between performance gap (as per UAS) and the corpus statistics ratio gap is in the Appendix Figure 3.

5 Related Work

Cross-Lingual Transfer for Parsing Many approaches have been developed to transfer a dependency parser. However, they mainly focus on better capture information from the source language(s). McDonald et al. (2011); Guo et al. (2016); Täckström et al. (2013); Chen et al. (2019) consider transferring a parser trained on multiple source languages. Agić (2017); Lin et al. (2019) selects good source languages by comparing part-of-speech tags sequences. Søgaard (2011); Täckström et al. (2013) chooses suitable data points from the source language. Pires et al. (2019) uses multilingual BERT to leverage language features from multiple languages. Ahmad et al. (2019) design an order-free model to take out the order features from the source language. Xiao and Guo (2014); Guo et al. (2015) learn an alignment from source words to target words. Ponti et al. (2018) learn an anisomorphism from the source parsing tree to target. Rasooli and Collins (2019) reorder the source data before training. In contrast, we focus on incorporating linguistic properties in the target languages.

Constrained Inference for Parsing Several previous studies show that adding constraints in inference time improves the performance of models. Grave and Elhadad (2015) consider incorporating constraints to promote popular types of arcs in an unsupervised setting. Naseem et al. (2010); Li et al. (2019) train a parser with constraints compiled from the frequency of particular arcs. Compared with the previous work, we focus on cross-lingual transfer with word order constraints.

Finally, prior studies have noticed that the word order information is significant for parsing and use it as features (Ammar et al., 2016; Naseem et al., 2012; Rasooli and Collins, 2017; Zhang and Barzilay, 2015; Dryer, 2007). Täckström et al. (2013) further propose to decompose these features from models for adapting target languages. Wang and Eisner (2018a) use the statistics of sur-face part-of-speech (POS) tags of target languages to learn the word order. Wang and Eisner (2018b) use POS tags of target languages together with a similar language, and design a stochastic permutation process to synthetic the word order. However, none of them consider using the word order features as constraints.

Incorporating Constraints In NLP Tasks Constraints are widely incorporated in variety of NLP tasks. To name a few, Roth and Yih (2004) propose to formulate constrained inferences in NLP as integer linear programming problems. To solve the intractable structure, Rush and Collins (2012) decompose the structure and incorporate constraints on some composite tasks. To improve the performance of a model, Chang and Collins (2011); Peng et al. (2015) incorporate constraints on exact decoding tasks and inference tasks on graphical models, and Chang et al. (2013); Dalvi (2015); Martins (2015) incorporate corpus-level constraints on semi-supervised multilabel classification and coreference resolution. Zhao et al. (2017) incorporate corpus-level constraints to avoid amplifying gender bias on visual semantic role labeling and multilabel classification. In contrast to previous work, we incorporate corpus-level constraints to facilitate dependency parser in the cross-lingual transfer setting.

6 Conclusion

We propose to leverage corpus-linguistic statistics to guide the inference of cross-lingual dependency parsing. We compile these statistics into corpus-statistic constraints and design two inference algorithms on top of a graph-based parser based on Lagrangian relaxation and posterior regularization. Experiments on 19 languages show that our approach improves the performance of the cross-lingual parser substantially. In the future, we plan to study the design and incorporation of fine-grained constraints considering multipule languages for cross-lingual transfer. We also plan to adapt this constrained inference framework to other cross-lingual structured prediction problems, such as semantic role labeling.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa Workshop on Universal Dependencies, UDW@NoDaLiDa 2017, Gothenburg, Sweden, May 22, 2017*, pages 1–10. Association for Computational Linguistics.

Željko Agić, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kai-Wei Chang, S. Sundararajan, and S. Sathiya Keerthi. 2013. Tractable semi-supervised learning of complex structured prediction models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, volume 8190 of *Lecture Notes in Computer Science*, pages 176–191. Springer.

Yin-Wen Chang and Michael Collins. 2011. Exact decoding of phrase-based translation models through lagrangian relaxation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 26–37. ACL.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multisource cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3098–3112. Association for Computational Linguistics.

Y. J. Chu and T. H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Internation Conference on Learning Representations*.

Bhavana Bharat Dalvi. 2015. *Constrained Semi-supervised Learning in the Presence of Unanticipated Classes*. Ph.D. thesis, Google Research.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. *Internation Conference on Learning Representations*.

Matthew S Dryer. 2007. Word order. *Language typology and syntactic description*, 1:61–131.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049.

Edouard Grave and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1375–1384.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2734–2740.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, and Noah A Smith. 2016. Distilling an ensemble of greedy dependency parsers into one mst parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1744–1753.

Bowen Li, Jianpeng Cheng, Yang Liu, and Frank Keller. 2019. Dependency grammar induction with a neural variational transition-based parser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6658–6665.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3125–3135. Association for Computational Linguistics.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1427–1437. The Association for Computer Linguistics.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL-2005*, pages 91–98, Ann Arbor, Michigan, USA.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 205–211.

Nanyun Peng, Ryan Cotterell, and Jason Eisner. 2015. Dual decomposition inference for graphical models over strings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 917–927. The Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1531–1542. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Mohammad Sadegh Rasooli and Michael Collins. 2019. Low-resource syntactic transfer with unsupervised source reordering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3845–3856. Association for Computational Linguistics.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE.

Alexander M Rush and Michael Collins. 2011. Exact decoding of syntactic translation models through Lagrangian relaxation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 72–82.

Alexander M Rush and MJ Collins. 2012. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *Journal of Artificial Intelligence Research*, 45:305–362.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *Internation Conference on Learning Representations*.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 682–686. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071. Association for Computational Linguistics.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted pos labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Dingquan Wang and Jason Eisner. 2018a. Surface statistics of an unknown language indicate how to parse it. *Transactions of the Association for Computational Linguistics (TACL)*.

Dingquan Wang and Jason Eisner. 2018b. Synthetic data made to order: The case of parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1337.

Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2306–2315.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379. Association for Computational Linguistics.

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*.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2979–2989. Association for Computational Linguistics.