Typology-Aware Neural Dependency Parsing: Challenges and Directions

by

Adam Fisch

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

Author . . . .
Department of Electrical Engineering and Computer Science
January 17, 2020

Certified by . . . .
Regina Barzilay
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Typology-Aware Neural Dependency Parsing: Challenges and Directions

by

Adam Fisch

Submitted to the Department of Electrical Engineering and Computer Science
on January 17, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

This thesis explores the task of leveraging typology in the context of cross-lingual dependency parsing. While this linguistic information has shown great promise in pre-neural parsing, results for neural architectures have been mixed. The aim of the investigation put forth in this thesis is to better understand this state-of-the-art. Our main findings are as follows: 1) The benefit of typological information is derived from coarsely grouping languages into syntactically-homogeneous clusters rather than from learning to leverage variations along individual typological dimensions in a compositional manner; 2) Typology consistent with the actual corpus statistics yields better transfer performance; 3) Typological similarity is only a rough proxy of cross-lingual transferability with respect to parsing. Code for the work in this thesis is available at `https://github.com/ajfisch/TypologyParser`.

Thesis Supervisor: Regina Barzilay
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank Dingquan Wang, Jason Eisner, the MIT NLP group (special thanks to Jiaming Luo), and the anonymous reviewers for Fisch et al. (2019) for their valuable comments on the contents of this thesis. I am also grateful for my advisor Regina Barzilay's guidance and support over the course of this research, especially during its challenges. This work would not have been possible without the help of Jiang Guo, who collaborated with me on (Fisch et al., 2019), and fought together with me to the bottom of countless technical mysteries. Finally, I am also thankful for the love and encouragement of my family, especially my parents and brothers.

This research was supported in part by the National Science Foundation Graduate Research Fellowship under Grant #1122374. Any opinion, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the the NSF or the U.S. Government.

Bibliographic Note

Portions of this thesis are based on the peer-reviewed publication Fisch et al. (2019).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last decade, dependency parsers for resource-rich languages have steadily continued to improve. In parallel, significant research efforts have been dedicated towards advancing cross-lingual parsing. This direction seeks to capitalize on existing annotations in resource-rich languages by transferring them to the rest of the world's over 7,000 languages (Bender, 2011). The NLP community has devoted substantial resources towards this goal, such as the creation of universal annotation schemas, and the expansion of existing treebanks to diverse language families. Nevertheless, cross-lingual transfer gains remain modest when put in perspective: we show that the performance of cross-lingual transfer models can often be exceeded using only a handful of annotated sentences in the target language (Chapter 4). The considerable divergence of language structures proves challenging for current models.

One promising direction for handling these divergences is linguistic typology. Linguistic typology classifies languages according to their structural and functional features, providing a scaffold for comparing languages. By explicitly highlighting specific similarities and differences in languages' syntactic structures, typology holds great potential for facilitating cross-lingual transfer (O'Horan et al., 2016). Indeed, non-neural parsing approaches have already demonstrated empirical benefits of typology-aware models (Naseem et al., 2012; Ponti et al., 2018; Täckström et al., 2013; Zhang and Barzilay, 2015). While adding discrete typological attributes is straightforward for traditional feature-based approaches, for modern neural parsers finding an effective implementation choice is more of an open question. Not surprisingly, the reported results have been mixed. For instance, Ammar et al. (2016) found no benefit to using typology for parsing when using a neural-based model, while Wang and Eisner (2018) and Scholivet et al. (2019) did in several cases.

There are many possible hypotheses that can attempt to explain the state-of-the-art. Might neural models already implicitly learn typological information on their own? Is the hand-specified typology information sufficiently accurate—or provided in the right granularity—to always be useful? How do cross-lingual parsers use, or ignore, typology when making predictions? Without understanding answers to these questions, it is difficult to develop a principled way for robustly incorporating linguistic knowledge as an inductive bias for cross-lingual transfer.

In this paper, we explore these questions in the context of two predominantly-used typology-based neural architectures for delexicalized dependency parsing. We focus on delexicalized parsing in order to isolate the effects of syntax by removing lexical influences. The first method implements a variant of selective sharing (Naseem et al., 2012); the second adds typological information as an additional feature of the input sentence. Both models are built on top of the popular Biaffine Parser (Dozat and Manning, 2017). We study model performance across multiple forms of typological representation and resolution.

1.1 Key Findings

Our key findings are as follows:

- **Typology as Quantization** Cross-lingual parsers use typology to coarsely group languages into syntactically-homogeneous clusters, yet fail to significantly capture finer inter- and intra-cluster distinctions or typological feature compositions. Our results indicate that they primarily take advantage of the simple geometry of the typological space (e.g. language distances), rather than specific variations in individual typological dimensions (e.g. `SV` vs. `VS`).

- **Typology Quality** Typology that is consistent with the actual corpus statistics results in better transfer performance, most likely by capturing a better reflection of the typological variations within that sample. Typology granularity also matters. Finer-grained, high-dimensional representations prove harder to use robustly.

- **Typology vs. Parser Transferability** Typological similarity only partially explains cross-lingual transferability with respect to parsing. The geometry of the typological space does not fully mirror that of the "parsing" space, and therefore requires task-specific refinement.

1.2 Outline

The rest of this thesis is organized as follows:

- **Chapter 2** introduces different notions of typology and how it can be represented for different languages.

- **Chapter 3** presents our models for dependency parsing, both with and without any typology augmentation.

- **Chapter 4** details our experimental design and results to comprehensively evaluate and analyze the effects of the typology augmentation schemes we considered on the model parsing performance.

- **Chapter 5** presents a summary of related work.

- **Chapter 6** concludes the thesis and discusses directions for future work.

Chapter 2

Typology Representations

The performance of typology-aware neural dependency parsers can depend on the type of typological representation used. In this chapter we describe the several variants of language typology representation that we test and compare in our experiments.

2.1 Linguistic Typology

The standard representation of typology is sets of annotations by linguists for a variety of language-level properties, which we refer to in this thesis as "Linguistic Typology" ($\mathbf{T_L}$). These properties can be found in online databases such as The World Atlas of Language Structures (WALS)[1] (Dryer and Haspelmath, 2013). WALS contains over 190 features that describe aspects of phonology, morphology, and syntax—all curated by dozens of linguists for hundreds of languages. We consider the same subset of features related to word order as used by Naseem et al. (2012), represented as a k-hot vector $T \in \{0, 1\}^{\sum_f |V_f|}$, where V_f is the set of values feature f may take.

The typological tendencies for some languages are not always consistent, however, and can vary across different text samples. This is a deficiency for static databases such as WALS—sections 2.2 and 2.3 describe alternative representations that capture these variations. As a point of comparison, we also consider WALS features directly derived from the target corpus. Table 2.1 summarizes the rules we used to

[1] `https://wals.info/`

derive corpus-specific WALS features. The values are determined by the dominance of directionalities, e.g., if $\frac{\#\{\curvearrowright\}}{\#\{\curvearrowright\}+\#\{\curvearrowleft\}} > \delta$, then its typological feature is set to the right-direction value, and vice versa. In-between values are set to `Mixed`. In our experiments we use $\delta = 0.75$.

WALS ID	Condition	Values
82A	`relation` $\in$ {`nsubj`, `csubj`} $\wedge$ `h.p=VERB` $\wedge$ (`m.p=NOUN` $\vee$ `m.p=PRON`)	`VS`($\curvearrowright$), `SV`($\curvearrowleft$), `Mixed`
83A	`relation` $\in$ {`dobj`, `iobj`} $\wedge$ `h.p=VERB` $\wedge$ (`m.p=NOUN` $\vee$ `m.p=PRON`)	`VO`($\curvearrowright$), `OV`($\curvearrowleft$), `Mixed`
85A	(`h.p=NOUN` $\vee$ `h.p=PRON`) $\wedge$ `m.p=ADP`	`Prepositions`($\curvearrowleft$), `Postpositions`($\curvearrowright$)
86A	`h.p=NOUN` $\wedge$ `m.p=NOUN`	`Noun-Genitive`($\curvearrowright$), `Genitive-Noun`($\curvearrowleft$), `Mixed`
87A	`h.p=NOUN` $\wedge$ `m.p=ADJ`	`Adjective-Noun`($\curvearrowleft$), `Noun-Adjective`($\curvearrowright$), `Mixed`
88A	`relation` $\in$ {`det`} $\wedge$ `m.p=DET`	`Demonstrative-Noun`($\curvearrowleft$), `Noun-Demonstrative`($\curvearrowright$), `Mixed`

Table 2.1: Rules for determining the dependency arc set of each specific WALS feature type. The arc direction specificed in the parenthesis of each value indicates the global directional tendency of the corresponding typological feature.

2.2 Liu Directionalities

Liu (2010) proposed using a real-valued vector $T \in [0,1]^r$ of the average *directionalities* of each of a corpus' r dependency relations as a typological descriptor. We refer to them in this thesis as "Liu Directionalities" ($\mathbf{T}_L$). Clearly, this representation is related to its categorical linguistic counterpart. For example, a language with a dominantly left-directed `nsubj` treebank is likely 82A `SV`. These serve as a more fine-grained alternative to linguistic typology. Compared to WALS, there are rarely missing values, and the *degree* of dominance of each dependency ordering is directly encoded — potentially allowing for better modeling of local variance within a language. It is important to note, however, that true directionalities require a parsed corpus to be derived; thus, they are not a realistic option for cross-lingual parsing in practice (though Wang and Eisner (2017) indicate that they can be predicted from unparsed corpora with reasonable accuracy). Nevertheless, we include them for completeness.

Among all the 37 relation types defined in Universal Dependencies, we select a subset of dependency relations which appear in at least 20 languages, as listed in Table 2.2. For relation types that are missing in a specific language, we simply put its value (directionality) as 0.5 without making any assumption to its tendency.

cc	conj	case	nsubj	nmod	dobj
advcl	amod	advmod	neg	nummod	xcomp
ccomp	cop	acl	aux	punct	det
iobj	dep	csubj	parataxis	mwe	name
nsubjpass	compound	auxpass	csubjpass	mark	appos
vocative	discourse				

Table 2.2: Subset of the Universal Dependency relations used for deriving the Liu Directionalities typology.

2.3 Surface Statistics

It is possible to derive a proxy measure of typology from part-of-speech tag sequences alone—which we refer to in this thesis as "Surface Statistics" ($\mathbf{T_L}$). Wang and Eisner (2017) found surface statistics to be highly predictive of language typology. For example, a language with many initial `NOUN VERB` subsequences is also fairly likely to be `82A SV`. Wang and Eisner (2018) replaced typological features entirely with surface statistics in their augmented dependency parser. Surface statistics have the advantage of being readily available and are not restricted to narrow linguistic definitions, but are less informed by the true underlying structure. We compute the set of hand-engineered features used in Wang and Eisner (2018), yielding a real-valued vector $T \in [0, 1]^{2380}$.

Chapter 3

Dependency Parsing Models

The performance of typology-aware neural dependency parsers can also depend on the *how* the typological representation is integrated into the model architecture. In this chapter we describe the different architecture choices used in our experiments.

3.1 Baseline Parsing Architecture

We use the graph-based Deep Biaffine Attention neural parser of Dozat and Manning (2017) as our baseline model. Given a delexicalized sentence s consisting of n part-of-speech tags, the Biaffine Parser embeds each tag $\boldsymbol{p}_i$, and encodes the sequence with a bi-directional LSTM to produce tag-level contextual representations $\boldsymbol{h}_i$. Each $\boldsymbol{h}_i$ is then mapped into head- and child-specific representations for arc and relation prediction, $\boldsymbol{h}_i^{\text{arc-dep}}$, $\boldsymbol{h}_i^{\text{arc-head}}$, $\boldsymbol{h}_i^{\text{rel-dep}}$, and $\boldsymbol{h}_i^{\text{rel-head}}$, using four separate multi-layer perceptrons. For decoding, arc scores are computed as:

$$s_{ij}^{\text{arc}} = \left(\boldsymbol{h}_i^{\text{arc-head}}\right)^T \left(U^{\text{arc}} \boldsymbol{h}_j^{\text{arc-dep}} + \boldsymbol{b}^{\text{arc}}\right) \tag{3.1}$$

while the score for dependency label r for edge (i, j) is computed in a similar fashion:

$$\begin{aligned} s^{\text{rel}}_{(i,j),r} = &\left(\boldsymbol{h}_i^{\text{rel-head}}\right)^T U_r^{\text{rel}} \boldsymbol{h}_j^{\text{rel-dep}} + \\ &\left(\boldsymbol{u}_r^{\text{rel-head}}\right)^T \boldsymbol{h}_i^{\text{rel-head}} + \\ &\left(\boldsymbol{u}_r^{\text{rel-dep}}\right)^T \boldsymbol{h}_j^{\text{rel-dep}} + b_r \end{aligned} \tag{3.2}$$

Both s^{arc}_{ij} and $s^{\text{rel}}_{(i,j),r}$ are trained greedily using cross-entropy loss with the correct head or label. At test time the final tree is composed using the Chu-Liu-Edmonds (CLE) maximum spanning tree algorithm (Chu and Liu, 1965; Edmonds, 1967).

3.2 Typology-Augmented Parsers

Selective Sharing Naseem et al. (2012) introduced the idea of *selective sharing* in a generative parser, where the features provided to a parser were controlled by its typology. The idea was extended to discriminative models by Täckström et al. (2013). For neural parsers which do not rely on manually-defined feature templates, however, there is not an explicit way of using selective sharing. Here we choose to directly incorporate selective sharing as a bias term for arc-scoring:

$$s^{\text{arc-aug}}_{ij} = s^{\text{arc}}_{ij} + \boldsymbol{v}^\top \boldsymbol{f}_{ij} \tag{3.3}$$

where $\boldsymbol{v}$ is a learned weight vector and $\boldsymbol{f}_{ij}$ is a feature vector engineered using Täckström et al.'s head-modifier feature templates (Table 3.1).

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

Table 3.1: Arc-factored feature templates for selective sharing. Arc direction: $\texttt{d} \in \{\textsc{Left}, \textsc{Right}\}$; Part-of-speech tag of head / modifier: `h.p` / `m.p`. WALS features: `w.X` for `X=81A` (order of `Subject, Verb` and `Object`), `85A` (order of `Adposition` and `Noun`), `86A` (order of `Genitive` and `Noun`), `87A` (order of `Adjective` and `Noun`).

Input Features We follow Wang and Eisner (2018) and encode the typology for language l with an MLP, and concatenate it with each input:

$$\Phi = W_2 \cdot \tanh\left(W_1 \cdot \mathbf{T}^{(l)} + \boldsymbol{b}\right) \tag{3.4}$$

$$\boldsymbol{h} = \texttt{BiLSTM}\left(\{\boldsymbol{p}_1 \oplus \Phi, \ldots, \boldsymbol{p}_n \oplus \Phi\}\right) \tag{3.5}$$

This approach assumes the parser is able to learn to use information in $\mathbf{T}^{(l)} \in \{\mathbf{T}_L^{(l)}, \mathbf{T}_D^{(l)}, \mathbf{T}_S^{(l)}\}$ to induce some distinctive change in encoding $\boldsymbol{h}$.

3.3 Fine-tuning

We also compare to simple model fine-tuning on a few labelled sentences from the target language. We fine-tune the baseline architecture from Section 3.1 using only 10 examples for supervision.

Chapter 4

Experiments

4.1 Data

We conduct our analysis on the Universal Dependencies v1.2 dataset (Nivre et al., 2015)[1] and follow the same train-test partitioning of languages as Wang and Eisner (2018). We train on 20 treebanks and evaluate cross-lingual performance on the other 15; test languages are shown in Table 4.1. Two treebanks that overlap with the training languages are excluded from evaluation, following the setting of Wang and Eisner (2018). We perform hyper-parameter tuning via five-fold cross-validation on the training languages. Test results are reported over the *train* splits of the held-out languages.

4.2 Training

To train our baseline parser and its typology-augmented variants, we use Adam (Kingma and Ba, 2015) with a learning rate of 10^{-3} for 200K updates (2M when using GD). We use a batch size of 500 tokens. Early stopping is also employed, based on the validation set in the training languages. For *fine-tune*, we perform 100 SGD updates with no early-stopping. Following Dozat and Manning (2017), we use a 3-layered bidirectional LSTM (`BiLSTM`) (Hochreiter and Schmidhuber, 1997) with a hidden size

[1] We evaluate on this older release of UD for fair comparison to Wang and Eisner (2018).

of 400. The hidden sizes of the MLPs for predicting arcs and dependency relations are 500 and 100, respectively.

Our baseline model shares all parameters across languages. During training, we truncate each training treebank to a maximum of 500K tokens for efficiency. Batch updates are composed of examples derived from a single language, and are sampled uniformly, such that the number of per-language updates are proportional to the size of each language's treebank. Following Wang and Eisner (2018), when training on GD, we sample a batch from a real language with probability 0.2, and a batch of GD data otherwise.

All reported numbers are the average of three runs with different random seeds. All models are implemented in PyTorch (Paszke et al., 2019).

4.3 Results

Tables 4.1 and 4.2 present our cross-lingual transfer results for unlabelled attachment scores (UAS) and labelled attachment scores (LAS), respectively. The entries for B* and $+\mathbf{T}_S^*$ are the baseline and surface statistics model results, respectively, of Wang and Eisner (2018).[2]

On UAS scores, our baseline model improves over the benchmark in (Wang and Eisner, 2018) by more than 6%. As expected, using typology yields mixed results. Selective sharing provides little to no benefit over the baseline. Incorporating the typology vector as an input feature is more effective, with the Liu Directionalities ($\mathbf{T}_D$) driving the most measurable improvements — achieving statistically significant gains on 13/15 languages. The Linguistic Typology ($\mathbf{T}_L$) gives statistically significant gains on 10/15 languages. Nevertheless, the results are still modest. Fine-tuning on only 10 sentences yields a **2.3× larger** average UAS increase, a noteworthy point of reference. LAS scores show similar trends.

[2] Wang and Eisner (2018)'s final $\mathbf{T}_S^*$ also contains additional neural features that we omitted, as we found it to under-perform using only hand-engineered features.

Language	B^*	$+\mathbf{T}_S^*$	Our Baseline	Selective Sharing	$+\mathbf{T}_L$	$+\mathbf{T}_D$	$+\mathbf{T}_S$	Fine-tune
Basque	49.89	54.34	56.18	56.54	$56.35^\dagger$	56.77	56.50	60.71
Croatian	65.03	67.78	74.86	75.23	74.07	77.39	75.20	78.39
Greek	65.91	68.37	70.09	70.49	68.05	71.66	70.47	73.35
Hebrew	62.58	66.27	68.85	68.61	72.02	72.75	69.21	73.88
Hungarian	58.50	64.13	63.81	64.78	70.28	66.40	64.21	72.50
Indonesian	55.22	64.63	63.68	64.96	69.73	67.73	66.25	73.34
Irish	58.58	61.51	61.72	$61.49^\dagger$	65.88	66.49	62.21	66.76
Japanese	54.97	60.41	57.28	57.80	63.83	64.28	57.04	72.72
Slavonic	68.79	71.13	75.18	$75.17^\dagger$	74.65	74.17	$75.16^\dagger$	73.11
Persian	40.38	34.20	53.87	53.61	45.14	56.72	53.03	59.92
Polish	72.15	76.85	76.01	$75.93^\dagger$	79.51	71.09	76.29	77.78
Romanian	66.55	69.69	73.00	73.40	75.20	76.34	73.82	75.15
Slovenian	72.21	76.06	81.21	80.99	81.39	81.36	80.92	82.43
Swedish	72.26	75.32	79.39	79.64	80.28	80.10	79.22	81.29
Tamil	51.59	57.53	57.81	58.85	59.70	60.37	58.39	62.94
Average	60.97	64.55	67.53	67.83	69.07	69.57	67.86	72.28

Table 4.1: A comparison of UAS scores of all methods on held-out test languages. Results with differences that are statistically *insignificant* compared to the baseline are marked with † (arc-level paired permutation test with $p \geq 0.05$).

Language	B^*	$+\mathbf{T}_S^*$	Our Baseline	Selective Sharing	$+\mathbf{T}_L$	$+\mathbf{T}_D$	$+\mathbf{T}_S$	Fine-tune
Basque	27.07	31.46	34.64	34.79	36.49	36.83	34.90	43.04
Croatian	48.68	52.29	61.34	$61.41^\dagger$	59.86	63.72	61.60	65.07
Greek	50.10	56.73	56.51	$56.53^\dagger$	55.16	60.18	$56.59^\dagger$	64.66
Hebrew	49.71	53.29	41.15	41.05	43.58	43.63	41.50	43.14
Hungarian	42.85	47.73	32.65	33.43	34.14	32.01	33.07	44.26
Indonesian	39.46	47.63	47.17	48.21	51.82	50.78	49.22	62.23
Irish	39.06	40.75	39.63	$39.60^\dagger$	43.02	42.14	40.24	48.58
Japanese	37.57	40.6	43.32	43.69	47.67	48.10	42.85	60.59
Slavonic	40.03	43.95	57.35	$57.40^\dagger$	56.89	56.69	57.19	53.88
Persian	30.06	24.6	35.72	35.59	32.85	39.78	34.93	49.72
Polish	50.08	54.85	61.67	61.57	64.69	57.20	61.71	65.68
Romanian	50.90	53.42	55.77	56.21	$55.99^\dagger$	59.28	56.48	59.12
Slovenian	57.09	61.48	70.86	70.01	70.44	70.03	70.29	73.81
Swedish	55.35	58.42	67.24	67.40	66.92	68.03	67.04	68.65
Tamil	28.39	37.81	33.81	34.57	34.96	36.61	34.70	47.46
AVG	43.09	47.00	49.26	49.43	50.30	51.00	49.49	56.66

Table 4.2: A comparison of LAS scores of all methods on held-out test languages. Results with differences that are statistically *insignificant* compared to the baseline are marked with † (arc-level paired permutation test with $p \geq 0.05$).

4.4 Analysis

Typology as Quantization Adding simple, discrete language identifiers to the input has been shown to be useful in multi-task multi-lingual settings (Ammar et al., 2016; Johnson et al., 2017). We hypothesize that the model utilizes typological information for a similar purpose by clustering languages by their parsing behavior. Testing this to the extreme, we encode languages using one-hot representations of their cluster membership. The clusters are computed by applying K-Means to WALS feature vectors (see Figure 4-1 for an illustration). We use Euclidean distance as our metric, another extreme simplification. There is no guarantee that all dimensions should be given equal weight, as indicated in Table 4.5. In this sparse form, compositional aspects of cross-lingual sharing are erased. Performance using this impoverished representation, however, only suffers slightly compared to the original — dropping by just 0.56% UAS overall and achieving statistically significant parity or better with $\mathbf{T}_L$ on 7/15 languages. A gap does still partially remain; future work may investigate this further—for example, this might be explained by soft versus hard clustering.

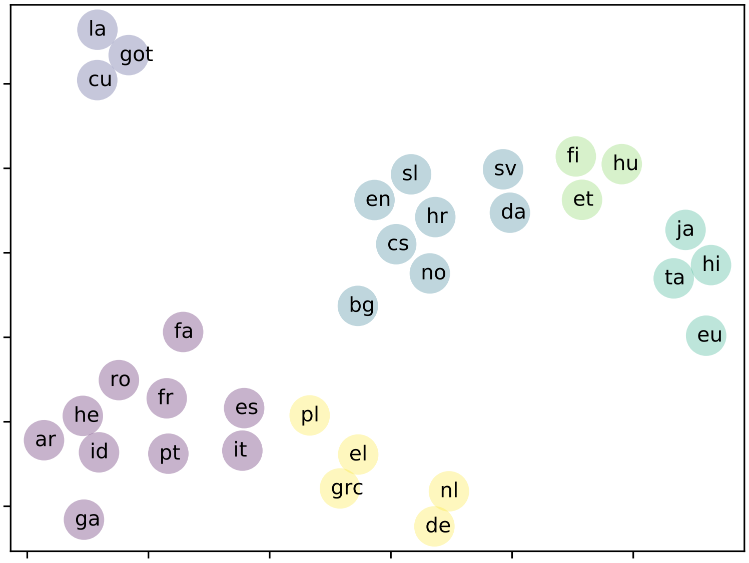

Figure 4-1: t-SNE projection of WALS vectors with clustering. Persian (fa) is an example of a poorly performing language that is also far from its cluster center.

WALS ID	82A	83A	85A	86A	87A	88A
Logreg	87	85	97	92	94	92
Majority	61	56	87	75	51	82

Table 4.3: Performance of typology prediction using hidden states of the parser's encoder, compared to a majority baseline which predicts the most frequent category.

+GD	B*	+$\mathbf{T}_S^*$	Baseline	+$\mathbf{T}_L^\ddagger$	+$\mathbf{T}_D$	+$\mathbf{T}_S$
Average	–	67.11	68.45	69.23	68.36	67.12

Table 4.4: Average UAS results when training with Galactic Dependencies. The Linguistic Typology ($\mathbf{T}_L^\ddagger$) is computed directly from the corpora using the rules Table 2.1.

This phenomenon is also reflected in the performance when the original WALS features are used. Test languages that do belong to compact clusters have higher performance on average than that of those who are isolates (e.g., Persian, Basque). Indeed from Table 4.1 and Fig. 4-1 we observe that the worst performing languages are isolated from their cluster centers. Even though their typology vectors can be viewed as compositions of training languages, the model appears to have limited generalization ability. This suggests that the model does not effectively use individual typological features.

This can likely be attributed to the training routine, which poses two inherent difficulties: 1) the parser has few examples (entire languages) to generalize from, making it hard from a learning perspective and 2) a naïve encoder can already implicitly capture important typological features within its hidden state, using only the surface forms of the input. This renders the additional typology features redundant. Table 4.3 presents the results of probing the final max-pooled output of the `BiLSTM` encoder for typological features on a *sentence level.* We find they are nearly linearly separable — logistic regression achieves greater than 90% accuracy on average on held out sentences from the 15 training languages.

Wang and Eisner (2018) attempt to address the learning problem by using the synthetic Galactic Dependencies (GD) dataset (Wang and Eisner, 2016) as a form of data augmentation. GD constructs "new" treebanks with novel typological qualities by systematically combining the behaviors of real languages. Following their work,

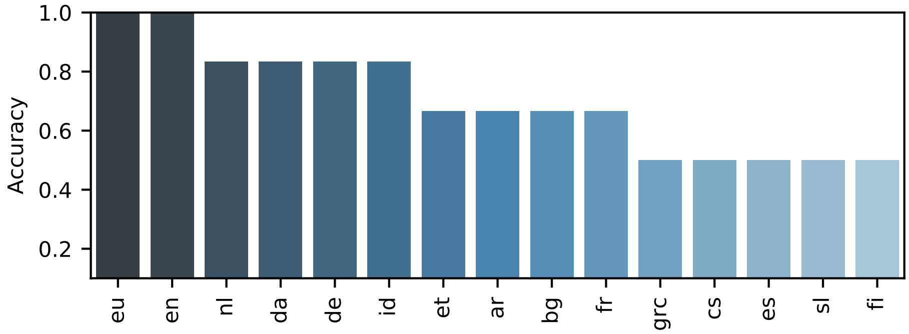

Figure 4-2: Averaged matching accuracy of the linguistically-defined WALS features on 15 randomly sampled languages compared to their corpus-specific values derived from UD v1.2. Rules for deriving the features from corpus are described in Table 2.1.

we add $8,820$ GD treebanks synthesized from the 20 UD training languages, giving $8,840$ training treebanks in total. Table 4.4 presents the results of training on this setting. While GD helps the weaker $\mathbf{T}_S^*$ substantially, the same gains are not realized for models built on top of our stronger baseline—in fact, the baseline only narrows the gap even further by increasing by 0.92% UAS overall.[3]

Typology Quality The notion of typology is predicated on the idea that some language features are consistent across different language samples, yet in practice this is not always the case. For instance, *Arabic* is listed in WALS as `SV` (82A, `Subject`⌒`Verb`), yet follows a large number of `Verb`⌒`Subject` patterns in UD v1.2. Fig. 4-2 further demonstrates that for some languages these divergences are significant (see Appendix F for concrete examples). Given this finding, we are interested in measuring the impact this noise has on typology utilization. Empirically, $\mathbf{T}_D$, which is consistent with the corpus, performs best. Furthermore, updating our typology features for $\mathbf{T}_L$ to match the dominant ordering of the corpus yields a slight improvement of 0.21% UAS overall, with statistically significant gains on 7/15 languages.

In addition to the quality of the representation, we can also analyze the impact of its resolution. In theory, a richer, high-dimensional representation of typology may capture subtle variations. In practice, however, we observe an opposite effect,

[3] Sourcing a greater number of real languages may still be helpful. The synthetic GD setting is not entirely natural, and might be sensitive to hyper-parameters.

where the Linguistic Typology ($\mathbf{T}_L$) and the Liu Directionalities ($\mathbf{T}_D$) outperform the surface statistics ($\mathbf{T}_S$), with $|\mathbf{T}_L| \approx |\mathbf{T}_D| \ll |\mathbf{T}_S|$. This is likely due to the limited number of languages used for training (though training on GD exhibits the same trend). This suggests that future work may consider using targeted dimensionality reduction mechanisms, optimized for performance.

Typology vs. Parser Transferability The implicit assumption of all the typology based methods is that the typological similarity of two languages is a good indicator of their parsing transferability. As a measure of parser transferability, for each language we select the oracle source language which results in the best transfer performance. We then compute precision@k for the nearest k neighbors in the typological space, i.e. whether the best source appears in the k nearest neighbors. As shown in Table 4.5, we observe that while there is some correlation between the two, they are far from perfectly aligned. $\mathbf{T}_D$ has the best alignment, which is consistent with its corresponding best parsing performance. Overall, this divergence motivates the development of approaches that better match the two distributions.

	P@1	P@3	P@5	P@10
$\mathbf{T}_L$	13	33	60	80
$\mathbf{T}_D$	27	67	67	93
$\mathbf{T}_S$	13	27	27	73

Table 4.5: Precision at k for identifying the best *parsing* transfer language, for the k typological neighbors.

Chapter 5

Related Work

This chapter provides a brief overview of other related works in the field with respect to cross-lingual parsing. Cross-lingual parsing is a long-standing task in natural language processing (McDonald et al., 2011; Søgaard and Wulff, 2012; Zeman and Resnik, 2008). Various approaches have tried to tackle the problem from different angles. Recent progress in the field has focused on lexical alignment (Guo et al., 2015, 2016; Schuster et al., 2019). Data augmentation (Wang and Eisner, 2017) is another promising direction, but at the cost of greater training demands. Both directions do not directly address structure. With respect to structure, Ahmad et al. (2019) showed structural-sensitivity is important for modern parsers; insensitive parsers suffer. Post-hoc constraints can be applied at test-time to attempt to match corpus level statistics with known typology (Meng et al., 2019). Performance, however, can vary based on how pronounced typological divergences are with respect to the given data sample (e.g., Figure 4-2), and can still lag behind simple fine-tuning methods (e.g., Section 3.3). Data transfer is an alternative solution to alleviate the typological divergences, such as annotation projection (Hwa et al., 2005; McDonald et al., 2011; Tiedemann, 2014; Yarowsky et al., 2001) and source treebank reordering (Rasooli and Collins, 2019). These approaches are typically limited by parallel data and imperfect alignments. Our work aims to understand cross-lingual parsing in the context of model transfer, with typology serving as language descriptors, with the goal of eventually addressing the issue of structure.

Chapter 6

Conclusion

Realizing the potential for typology may require rethinking current approaches. We can further drive performance by refining typology-based similarities into a metric more representative of actual transfer quality. Ultimately, we would like to design models that can directly leverage typological compositionality for distant languages.

Future Work

The work presented in this thesis answers some important questions about typology usage in modern neural dependency parsers, but still leaves some unanswered—and introduces additional ones:

- **Typology Representation:** Both the information content and quality of typology affects parsing performance. Current common typological representations (e.g., Chapter 2) do not appear to be well-suited for the task at hand. Instead, future work may seek to *learn* typological representations that capture universal linguistic properties that are indeed useful for cross-lingual transfer.

- **Learning Problem:** Learning to use typology from only a (relatively) few languages in a generalizeable way is a fundamentally hard machine learning problem (Section 4.4). More regularized training or directions such as meta-learning may yield more success.

- **Analysis Tools:** Interpreting the performance of neural models is difficult. Though we present several important analyses in this work, our scope is nevertheless still limited. Developing robust methods to measure and quantify if models are using typology or other inductive biases in the ways we expect is an important area of research.

Bibliography

Ahmad, W., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *NAACL*, pages 2440–2452.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *TACL*, 4:431–444.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*.

Chu, Y.-J. and Liu, T.-H. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Dozat, T. and Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. In *ICLR*.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B:233–240.

Fisch, A., Guo, J., and Barzilay, R. (2019). Working hard or hardly working: Challenges of integrating typology into neural dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2015). Cross-lingual dependency parsing based on distributed representations. In *ACL-IJCNLP*, volume 1, pages 1234–1244.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2016). A representation learning framework for multi-source transfer parsing. In *AAAI*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8).

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

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

Meng, T., Peng, N., and Chang, K.-W. (2019). Target language-aware constrained inference for cross-lingual dependency parsing. In *EMNLP-IJCNLP*.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *ACL*, pages 629–637. Association for Computational Linguistics.

Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bosco, C., Bowman, S., Celano, G. G. A., Connor, M., de Marneffe, M.-C., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Erjavec, T., Farkas, R., Foster, J., Galbraith, D., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Gonzales, B., Guillaume, B., Hajič, J., Haug, D., Ion, R., Irimia, E., Johannsen, A., Kanayama, H., Kanerva, J., Krek, S., Laippala, V., Lenci, A., Ljubešić, N., Lynn, T., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., Mori, S., Nurmi, H., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Prokopidis, P., Pyysalo, S., Ramasamy, L., Rosa, R., Saleh, S., Schuster, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Simov, K., Smith, A., Štěpánek, J., Suhr, A., Szántó, Z., Tanaka, T., Tsarfaty, R., Uematsu, S., Uria, L., Varga, V., Vincze, V., Žabokrtský, Z., Zeman, D., and Zhu, H. (2015). Universal dependencies 1.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

O'Horan, H., Berzak, Y., Vulic, I., Reichart, R., and Korhonen, A. (2016). Survey on the use of typological information in natural language processing. In *COLING*, pages 1297–1308.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Ponti, E. M., Reichart, R., Korhonen, A., and Vulić, I. (2018). Isomorphic transfer of syntactic structures in cross-lingual nlp. In *ACL*, pages 1531–1542.

Rasooli, M. S. and Collins, M. (2019). Low-resource syntactic transfer with unsupervised source reordering. In *NAACL*, pages 3845–3856.

Scholivet, M., Dary, F., Nasr, A., Favre, B., and Ramisch, C. (2019). Typological features for multilingual delexicalised dependency parsing. In *NAACL*, pages 3919–3930.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *NAACL*, pages 1599–1613.

Søgaard, A. and Wulff, J. (2012). An empirical etudy of non-lexical extensions to delexicalized transfer. In *COLING*.

Täckström, O., McDonald, R., and Nivre, J. (2013). Target language adaptation of discriminative transfer parsers. In *NAACL*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *COLING*, pages 1854–1864.

Wang, D. and Eisner, J. (2016). The galactic dependencies treebanks: Getting more data by synthesizing new languages. *TACL*, 4:491–505.

Wang, D. and Eisner, J. (2017). Fine-grained prediction of syntactic typology: Discovering latent structure with supervised learning. *TACL*, 5:147–161.

Wang, D. and Eisner, J. (2018). Surface statistics of an unknown language indicate how to parse it. *TACL*, 6:667–685.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Zhang, Y. and Barzilay, R. (2015). Hierarchical low-rank tensors for multilingual transfer parsing. In *EMNLP*.