Working Hard or Hardly Working: Challenges of Integrating Typology into Neural Dependency Parsers

Adam Fisch*  Jiang Guo*  Regina Barzilay
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{fisch, jiang.guo, regina}@csail.mit.edu

Abstract

This paper 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 our investigation 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.¹

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 tree-banks to diverse language families. Nevertheless, cross-lingual transfer gains remain modest when put in perspective: the performance of transfer models can often be exceeded using only a handful of annotated sentences in the target language (Section 5). 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. 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; 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.² The first method implements a variant of selective sharing (Naseem et al., 2012);

¹The first two authors contributed equally.
²Code: github.com/ajfisch/TypologyParser

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.

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 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.

2 Typology Representations

**Linguistic Typology, **$T_L$**: The standard representation of typology is sets of annotations by linguists for a variety of language-level properties. These properties can be found in online databases such as The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). 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. 

**Liu Directionalities, **$T_D$**: 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. 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. Nevertheless, we include them for completeness.

3 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 $p_i$, and encodes the sequence with a bi-directional LSTM to produce tag-level contextual representations $h_i$. Each $h_i$ is then mapped into head- and child-specific representations for arc and relation prediction, $h_{arc-dep}^i$, $h_{arc}^i$, $h_{rel-dep}^i$, and $h_{rel}^i$, using four separate multi-layer perceptrons.

For decoding, arc scores are computed as:

$$s_{ij}^{arc} = (h_{i}^{arc-dep})^T (U_{arc}) h_{j}^{arc-dep} + b_{arc}$$ (1)

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

$$s_{(i,j),r}^{rel} = (h_{i}^{rel-dep})^T (U_{r}) h_{j}^{rel-dep} + b_{r}$$ (2)

Both $s_{ij}^{arc}$ and $s_{(i,j),r}^{rel}$ 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).
Table 1: A comparison of all methods on held-out test languages. UAS results are reported over the train splits of the held-out languages, following (Wang and Eisner, 2018). \( B^* \) and \( +T_S^* \) are the baseline and surface statistics model results, respectively, of (Wang and Eisner, 2018). \(^\dagger\) Fine-tune is the result of adapting our baseline model using only 10 sentences from the target language. All of our reported numbers are the average of three runs with different random seeds. Results with differences that are statistically insignificant compared to the baseline are marked with \(^\dagger\) (arc-level paired permutation test with \( p \geq 0.05 \)).

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

4 Typology Augmented Parsing

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 isn’t an explicit way of using selective sharing. Here we choose to directly incorporate selective sharing as a bias term for arc-scoring:

\[
s_{ij}^{\text{arc-aug}} = s_{ij}^{\text{arc}} + \nu^T f_{ij}
\]

where \( \nu \) is a learned weight vector and \( f_{ij} \) is a feature vector engineered using Täckström et al.’s head-modifier feature templates (Appendix B).

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

\[
\Phi = W_2 \cdot \tanh \left( W_1 \cdot T^{(l)} + b \right) \\
h = \text{BiLSTM}\left( \{ p_1 + \Phi, \ldots, p_n + \Phi \} \right)
\]

This approach assumes the parser is able to learn to use information in \( T^{(l)} \in \{ T_L^{(l)}, T_D^{(l)}, T_S^{(l)} \} \) to induce some distinctive change in encoding \( h \).

5 Experiments

Data: We conduct our analysis on the Universal Dependencies v1.2 dataset (Nivre et al., 2015) 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 1. We perform hyper-parameter tuning via 5-fold cross-validation on the training languages.

Results: Table 1 presents our cross-lingual transfer results. 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 (\( T_D \)) driving the most measurable improvements — achieving statistically significant gains on 13/15 languages. The Linguistic Typology (\( T_L \)) gives statistically significant gains on 10/15 languages. Nevertheless, the results are still modest. Fine-tuning on only 10 sentences yields a \( 2.3 \times \) larger average UAS increase, a noteworthy point of reference.

\( ^4\)Wang and Eisner (2018)’s final \( T_S^* \) also contains additional neural features that we omitted, as we found it to underperform using only hand-engineered features.

\( ^5\)Two treebanks are excluded from evaluation, following the setting of Wang and Eisner (2018).
6 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 1 for an illustration). 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 $T_L$ on 7/15 languages. A gap does still partially remain; future work may investigate this further.

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 1 and Fig. 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 naive 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 2 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.

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 permuting the behaviors of real languages. Following their work, we add 8,820 GD treebanks synthesized from the 20 UD training languages, giving 8,840 training treebanks in total. Table 3 presents the results of training on this setting. While GD helps the weaker $T_S$ 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.7

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 $\rightarrow$ Verb), yet follows a large number of Verb $\rightarrow$ Subject patterns in UD v1.2. Fig. 2 fur-

---

7Sourcing 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.

---

Table 1: Performance of typology prediction using hidden states of the parser’s encoder, compared to a majority baseline which predicts the most frequent category.

<table>
<thead>
<tr>
<th>WALS ID</th>
<th>82A</th>
<th>83A</th>
<th>85A</th>
<th>86A</th>
<th>87A</th>
<th>88A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logreg</td>
<td>87</td>
<td>85</td>
<td>97</td>
<td>92</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>Majority</td>
<td>61</td>
<td>56</td>
<td>87</td>
<td>75</td>
<td>51</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 2: Performance of typology prediction using hidden states of the parser’s encoder, compared to a majority baseline which predicts the most frequent category.

<table>
<thead>
<tr>
<th>+GD</th>
<th>$B^*$</th>
<th>$+T_S^*$</th>
<th>Baseline</th>
<th>$+T_L^*$</th>
<th>$+T_D$</th>
<th>$+T_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>–</td>
<td>67.11</td>
<td>68.45</td>
<td>69.23</td>
<td>68.36</td>
<td>67.12</td>
</tr>
</tbody>
</table>

Table 3: Average UAS results when training with Galactic Dependencies. The Linguistic Typology ($T_L^*$) here is computed directly from the corpora using the rules in Appendix E. All of our reported numbers are the average of three runs.

---

---
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, we observe that while there is some correlation between the two, they are far from perfectly aligned. 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.

Table 4: Precision@k for identifying the best parsing transfer language, for the k typological neighbors.

<table>
<thead>
<tr>
<th></th>
<th>P@1</th>
<th>P@3</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_L</td>
<td>13</td>
<td>33</td>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>T_D</td>
<td>27</td>
<td>67</td>
<td>67</td>
<td>93</td>
</tr>
<tr>
<td>T_S</td>
<td>13</td>
<td>27</td>
<td>27</td>
<td>73</td>
</tr>
</tbody>
</table>

Figure 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 Appendix E.

Other recent progress in cross-lingual parsing 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. Ahmad et al. (2019) showed structural-sensitivity is important for modern parsers; insensitive parsers suffer. Data transfer is an alternative solution to alleviate the typological divergences, such as annotation projection (Tiedemann, 2014) 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.

8 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.

Acknowledgments

We thank Dingquan Wang, Jason Eisner, the MIT NLP group (special thanks to Jiaming Luo), and the reviewers for their valuable comments. This research is supported in part by the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity, via contract #FA8650-17-C-9116, and the National Science Foundation Graduate Research Fellowship under Grant #1122374. Any opinion, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the the NSF, ODNI, IARPA, or the U.S. Government. We are also grateful for the support of MIT Quest for Intelligence program.
References


Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. TACL, 4:491–505.
