Neural Unsupervised Parsing Beyond English

Katharina Kann, Anhad Mohananey, Kyunghyun Cho and **Samuel R. Bowman**
New York University, USA
{kann, anhad, kyunghyun.cho, bowman}@nyu.edu

Abstract

Recently, neural network models which automatically infer syntactic structure from raw text have started to achieve promising results. However, earlier work on unsupervised parsing shows large performance differences between non-neural models trained on corpora in different languages, even for comparable amounts of data. With that in mind, we train instances of the PRPN architecture (Shen et al., 2018a)—one of these unsupervised neural network parsers—for Arabic, Chinese, English, and German. We find that (i) the model strongly outperforms trivial baselines and, thus, acquires at least some parsing ability for all languages; (ii) good hyperparameter values seem to be universal; (iii) how the model benefits from larger training set sizes depends on the corpus, with the model achieving the largest performance gains when increasing the number of sentences from 2,500 to 12,500 for English. In addition, we show that, by sharing parameters between the related languages German and English, we can improve the model's unsupervised parsing F1 score by up to 4% in the low-resource setting.

1 Introduction

Unsupervised parsing, the task of inducing hierarchical syntactic structure from a large amount of unlabeled text, has been widely studied in natural language processing (NLP) (Carroll and Charniak, 1992; Pereira and Schabes, 1992; Klein and Manning, 2002, 2004). Work on this task bears on open research questions involving human language learning and grammar design by demonstrating what can be learned without substantial prior knowledge. Further, it can also be practically relevant for low-resource languages or language styles.

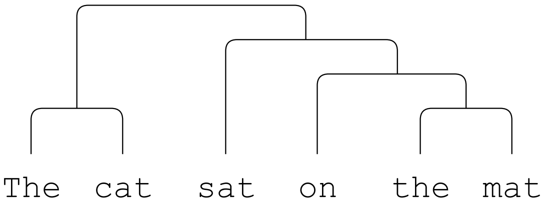

Figure 1: The constituency parse tree of the sentence *The cat sat on the mat*. In this work, we experiment with models that discover such syntactic structures in an unsupervised manner.

Recently, multiple types of neural network models have been added to the line of research on unsupervised parsing. Latent tree learning models learn to parse via optimization of a downstream task objective (Yogatama et al., 2017; Maillard et al., 2017; Choi et al., 2018). In contrast, generative unsupervised parsing models learn to model syntactic structure while being trained to language model (Shen et al., 2018a,c). While the latter model family has been able to generate parse trees which show a high accordance with expert annotations, its members, with the prominent Parsing Reading Predict-Network (PRPN) being no exception, have mostly been evaluated on English.[1] Thus, it is not obvious whether and when obtained results would hold true for other languages, especially if they are unrelated to English or dispose of significantly smaller training corpora. Some non-neural models for grammar induction—i.e., models, which perform unsupervised parsing—show language-dependent performance variation (Snyder et al., 2009), which motivates our investigation of recent neural models.

In this work, we first aim to answer the following research questions, focusing on the parsing-reading-predict network (PRPN; Shen et al., 2018a) and experimenting with Arabic, Chinese,

[1] Some work, e.g. Kim et al. (2019a), present PRPN results on Chinese, but without further analysis of language-dependent differences.

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 209–218
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

and German datasets: (i) Do neural grammar induction models succeed on languages which are unrelated to English? (ii) Does the required amount of training data vary significantly by language? (iii) Do optimal hyperparameter values differ between languages? Answering those questions provides insight into what to expect when applying a neural grammar induction model to a new, potentially low-resource, language. For instance, we find that the model outperforms trivial baselines for all languages and good hyperparameter values seem to be largely language-independent.

Second, motivated by transfer learning approaches for similar tasks like supervised dependency parsing (de Lhoneux et al., 2018), we propose a multilingual model trained on the related languages English and German, such that information from each language is leveraged for the other one. We find that, for small training corpora, this improves performance and reduces the required number of parameters.

Contributions. To summarize, we make the following contributions: (i) We perform the first thorough study of the PRPN's parsing ability across multiple languages. To facilitate this analysis and interpret the meaning of the experimental results, we compare to baselines and upper bounds for each language. (ii) For each language, we study variations in hyperparameter trends. We further investigate how the PRPN's performance depends on the training set size, and how this differs between languages. (iii) We present a multilingual variant of the model, which has been obtained via parameter sharing across two related languages. We show that this improves the model's performance. Our experiments with sharing parameters between English and German result in up to 4% gain in parsing F1 over training on each language separately.

2 Unsupervised Constituency Parsing

Human language is governed by a set of syntactic rules, which all grammatical or acceptable sentences follow. Unsupervised parsing or *grammar induction* is the task of detecting such structure automatically, i.e., without any human annotation. The underlying research question is: how much of this latent structure of language can be discovered from raw text alone and how much requires an inherent bias towards acquiring a valid grammar or additional (potentially non-linguistic) information? From a practical NLP perspective, grammar induction enables us to obtain syntactic information without labeled data, i.e., even in low-resource settings and for resource-poor languages. This information can then be of help for downstream tasks like machine translation (Aharoni and Goldberg, 2017).

In this work, we explore unsupervised *constituency* parsing. The PRPN, which we experiment with, aims at detecting so-called constituents in sentences. Thus, it splits a sentence's tokens into groups, usually based on their meaning. For example, in Figure 1, *The cat* and *on the mat* are constituents, inter alia. Constituency parsing is recursive: sentences are constructed from units which themselves consist of even smaller constituents which, in turn, consist of smaller groups of words, and so on.

This recursive structure of language is represented explicitly in recursive neural networks (Socher et al., 2011), which are also known as Tree-LSTMs. Successful induction of constituents enables us to then process input sentences using such a computational model instead of a sequential one like, e.g., a standard long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997). Since this can be relevant especially for low-resource languages with limited or zero training examples, we investigate in this work how the PRPN model for grammar induction behaves across languages and different (unlabeled) training set sizes.

3 Model and Setup

In this section, we first introduce the PRPN, which is the object of our investigations. We then present all datasets, baselines, and upper bounds we use in the set of experiments described in the next sections.

3.1 PRPN

The PRPN[2] consists of three principal components: (i) a parsing network, (ii) a reading network, and (iii) a predict network. We will summarize them in this section. A detailed explanation can be found in Shen et al. (2018a).

Parsing network. Given an input sentence $x = (x_0, x_1, \cdots, x_n)$ of length $n + 1$, the parsing net-

[2] We use the code from the original paper with minimal modifications: `https://github.com/yikangshen/PRPN.git`

work uses two convolutional layers to predict syntactic distances d_i between x_{i-1} and x_i for all pairs of adjacent words. This syntactic distance represents syntactic relationships between tokens in a sentence. Mathematically, the syntactic distance d_i between x_{i-1} and x_i is predicted by the PRPN as

$$h_i = \text{ReLU}(W_c \begin{bmatrix} x_{i-L} \\ x_{i-L+1} \\ \cdots \\ x_i \end{bmatrix} + b_c) \qquad (1)$$

$$d_i = \text{ReLU}(W_d h_i + b_d), \qquad (2)$$

where L is a a look-back parameter, which defines how many previous token are taken into account to compute the syntactic distance. W_c and b_c are kernel parameters, W_d and b_d represent another convolutional layer with a unit kernel size. Since the PRPN belongs to the family of unsupervised models with regard to parsing, there is no direct supervision on absolute values of the syntactic distance. Instead, these distances are trained via a language modeling downstream task.

For generating a constituency parse tree from the distances computed by the PRPN model, a recursive algorithm is used. Following this algorithm, we first, for a given set of distances $d_0, \dots, d_n$, find the maximum distance d_i. Then, we split the input sequence $x = (x_0, x_1, \cdots, x_n)$ corresponding to the distances into a left $(0, \dots, i-1)$ and a right $(i, \dots, n)$ part or subtree. For each of those, we then again find the maximum value to split the sequence into two parts. This process repeats till the leaf nodes are reached, thus generating the entire tree.

Reading network. The reading network processes a sentence based on gate values g_i^t, which are, at each time step t, computed from the syntactic distances as follows:

$$\alpha_j^t = \frac{\text{hardtanh}((d_t - d_j)\tau) + 1}{2} \qquad (3)$$

and

$$g_i^t = \prod_{j=i+1}^{t-1} \alpha_j^t. \qquad (4)$$

where hardtanh(x) is defined as

$$\text{hardtanh}(x) = \max(-1, \min(1, x)), \qquad (5)$$

and τ is a temperature parameter controlling the sensitivity to differences between distances.

The reading network then computes the memory state m_t at time step t from the input x_t, previous memory states $(m_{t-N_m}, \cdots, m_{t-1})$, and gate values $(g_0^t, \cdots, g_{t-1}^t)$. The memory consists of two sequences of vectors: a hidden tape $H_{t-1} = \{h_{t-N_m}, \cdots, h_{t-1}\}$, and a memory tape $C_{t-1} = \{c_{t-N_m}, \cdots, c_{t-1}\}$. Thus, the hidden state at time step t is defined as $m_t = (h_t, c_t)$, wherein h_t and c_t constitute the hidden and memory tape respectively. N_m represents the length of the memory span. At each step, the reading network performs an update using a structured attention mechanism, which is a variant of vanilla attention, but considers dependency relationship from the tree structure.

$$k_t = W_h h_{t-1} + W_x x_t \qquad (6)$$

$$q_i^t = \text{softmax} \frac{h_i k_t^T}{\sqrt{(\delta_k)}} \qquad (7)$$

$$s_i^T = \frac{g_i^t q_i^t}{\sum_i g_i^t} \qquad (8)$$

$$\begin{bmatrix} \tilde{h}_t \\ \tilde{c}_t \end{bmatrix} = \sum_{i=1}^{t-1} s_i^t \begin{bmatrix} h_i \\ c_i \end{bmatrix} \qquad (9)$$

Here, δ_k is the dimension of the hidden states. New values for h_t and c_t are then computed from x_t, $\tilde{h}_t$ and $\tilde{c}_t$ via the LSTM recurrent update.

Predict network. The last component of the PRPN predicts the probability of the next token x_{t+1} based on the memory states $m_0, \cdots, m_t$ from the reading network as well as the gates $g_0^{t+1}, \cdots, g_t^{t+1}$ from the parsing network. Since the true x_{t+1} is unknown at time step $t+1$, the predict network computes a temporary value for d_{t+1}:

$$d'_{t+1} = \text{ReLU}(W'_d h_t + b'_d). \qquad (10)$$

Obtaining α^{t+1} and g_i^{t+1} in the same way as before, the probability distribution of the next token x_{t+1} is then computed via a feed-forward network.

3.2 Datasets

We use existing constituency treebanks for Arabic, Chinese, English, and German. Table 1 shows statistics for all languages. For efficiency, we ignore sentences longer than 100 words during both training and evaluation in all our experiments.

	Arabic	Chinese	English	German
vocab size	21,902	23,714	15,617	10,367
train	18,087	57,251	43,738	18,598
dev	2,422	6,736	1,699	1,000
test	2,556	7,075	2,416	1,000

Table 1: Dataset statistics.

English. We perform English constituency parsing experiments for comparison with the original work by Shen et al. (2018a). We use the Wall Street Journal Section of the Penn Treebank (Marcus et al., 1999). We use parts 00-21 for training, 22 for validation and 23 for testing.

Arabic. We use the Arabic Treebank v3.0 (Maamouri et al., 2004). We randomly split the files into training, development and test: 200 files are used for each of test and development, and the remaining 1434 constitute our training set.

Chinese. We use the Chinese Penn Treebank v8.0 (Xue et al., 2005). Again, we randomly separate files into splits: the development and test sets consist of 300 files each, while the training set consists of the remaining 2407.

German. We use the NEGRA corpus (Skut et al., 1997), which consists of approximately $350,000$ words of German newspaper text (20,602 sentences). We divide the dataset into training, development, and test splits as suggested by Dubey and Keller (2003).

3.3 Parsing Baselines and Upper Bounds

We compare to the following baselines and upper bounds to better evaluate the performance of our models:

Best binary tree upper bound (BB). Since our datasets contain n-ary trees, but the PRPN only produces binary trees, obtaining a perfect F1 score is impossible. This upper bound represents the best score which can be obtained with binary trees.

Shen et al. (2018b) upper bound (SUB). Our second upper bound is a supervised parser, differing from the one presented by Shen et al. (2018b) only in that we do not predict or use any tags. It is trained on the gold annotations of the training set. We choose this particular approach since it is, like PRPN, based on the concept of syntactic distances: the parse tree is recovered from predicted distances between words in a sentence. Thus, we see it as the supervised approach which is most comparable to the PRPN. Since our model does not predict labels which are used to recover n-ary trees in the original work, we compute the F1 score for this approach only with respect to binary gold trees. This is acceptable for our purposes, since we are interested in the supervised parser upper bound only to get an idea of the difficulty of the datasets in our different languages. For the supervised SUB baseline, we use hidden state and embedding dimensions of 100 and 300, respectively, and keep the default settings from Shen et al. (2018b) for all other hyperparameters.

Left/right-branching trees baseline (LBR/RBR). Our next baseline consists of purely left- or right-branching trees. LBR refers to the F1 score strictly left-branching binary trees obtain compared to the gold annotations, and RBR denotes the score of strictly right-branching trees.

Balanced trees baseline (BTB). Finally, this baseline is similar to LBR/RBR, but considering balanced binary trees, which are created by recursively splitting each span into halves. For odd lengths, the middle word becomes a part of the right subtree.

4 Monolingual Experiments

4.1 Language-Dependence of Hyperparameters

It is of practical importance to know whether a set of hyperparameters found for one language transfers to another one without any changes, especially for low-resource language without annotated (development) data. Therefore, we ask the following questions: (i) Do hyperparameters depend on the language when using our datasets? (ii) How to choose good hyperparameters for a new language? We aim at answering these questions with respect to the PRPN.

Setup. We perform an extensive random hyperparameter search, training 45 models for each language, i.e., 180 models in total. The hyperparameters we vary are *embedding size*, *hidden state size*, and *learning rate*. Random combinations of values are selected uniformly from the following parameter ranges: embedding size in $[100, 400]$, hidden state size in $[200, 400]$, and learning rate in $[0.0005, 0.0015]$.

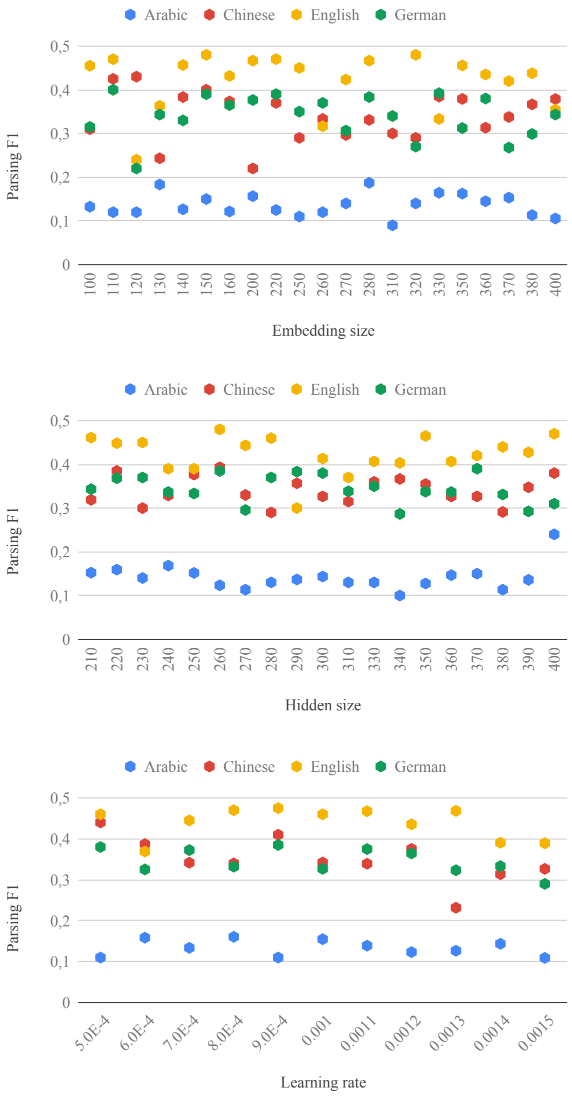

Figure 2: Parsing F1 as a function of different hyperparameters.

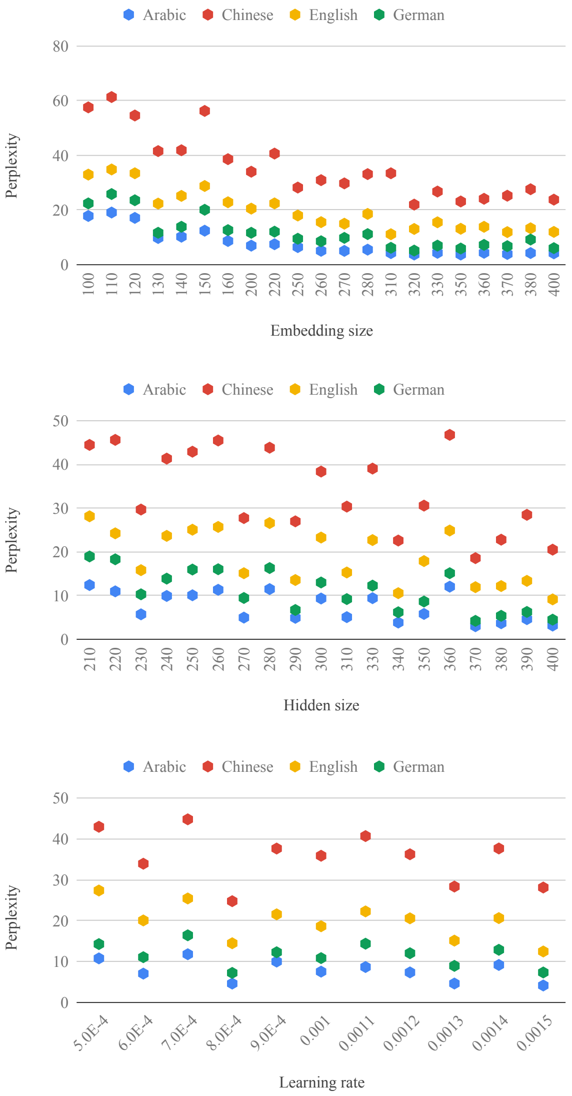

Figure 3: Language modeling perplexity as a function of different hyperparameters.

Results. The resulting parsing performances as a function of one hyperparameter at a time are shown in the three plots in Figure 2. The corresponding plots for language modeling perplexity can be found in Figure 3. We observe the following:

- First of all, we find no clear trend regarding which values yield the best parsing performance for either language. This shows that, as far as parsing is concerned, the PRPN is robust to hyperparameter changes. Furthermore, this also indicates that likely any values from within our ranges would be acceptable for a new language.

- Next, we look at the language modeling perplexity of all 180 models. While changes to the hyperparameters seem to not affect parsing performance, perplexity slightly decreases for larger embeddings sizes, as can be seen in Figure 2. This suggests that tuning hyperparameters using a language modeling objective for languages without annotated parsing data might not be helpful.

Because we do not find any substantial language-specific differences regarding the hyperparameter preferences of the model, we keep the default parameters of Shen et al. (2018a) for all following experiments: we use an embedding size of 200, a hidden state size of 400, and a learning rate of 0.001.

	Arabic	Chinese	English	German
PRPN	0.17 (.09)	0.28 (.07)	0.43 (.01)	0.41 (.03)
LBR	0.035	0.018	0.002	0.029
RBR	0.028	0.068	0.056	0.155
BTB	0.126	0.139	0.117	0.127
SUB	0.774	0.822	0.881	0.799
BB	0.914	0.924	0.908	0.878

Table 2: F1 scores for grammar induction. PRPN results are averaged over 5 training runs, with standard deviation in parentheses.

4.2 Unsupervised Parsing

After settling on hyperparameters, we evaluate the PRPN's performance for unsupervised parsing in Arabic, Chinese, English, and German. We train for a maximum of 150 epochs, but stop training anytime after 100 epochs if the training loss does not decrease for 5 consecutive epochs.

Results. All results are shown in Table 2. While unsupervised parsing of English using the PRPN has previously been studied widely (Shen et al., 2018a; Htut et al., 2018), we include the scores on English for comparison. We make the following observations:

- LBR, RBR, and BTB perform poorly for all languages. Since this has been shown to not be the case for English sentences only up to length 10 after filtering of punctuation, we conclude that, for longer sentences and with punctuation, these simple baselines are weak. This might be easily explained by the fact that the diversity of parses increases for longer sentences. The PRPN's F1 score is far higher that that of all trivial baselines for all languages, showing the effectiveness of the model overall.

- The F1 scores obtained by the PRPN differ a lot across languages, ranging from 0.17 for Arabic to 0.43 for English. This is in contrast to the BB and BTB scores, which differ only by 0.046 and 0.022 between extremes, respectively.

- The large difference between the PRPN's performance on one hand and BB as well as SUB, our supervised parsing upper bound, on the other, demonstrates that there is still room for improvement. Furthermore, the differences between languages for SUB is relatively small. In particular, it is smaller than 5% for all languages. The difference in unsupervised constituency parsing performance across languages could thus be attributed to the PRPN's inherent preference toward English-like languages.

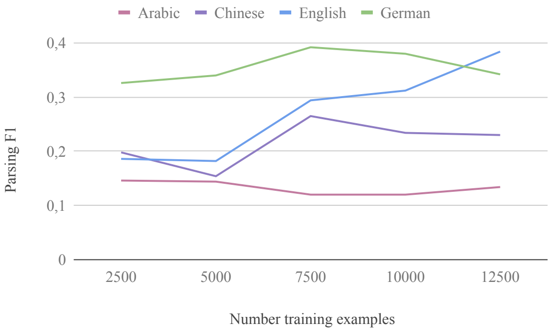

Figure 4: Learning curves for all languages.

4.3 Low-Resource Settings

We now aim to understand how the PRPN's performance for each language depends on the amount of data available. While large unlabeled corpora can be obtained easily for languages which are popular in NLP research, this is not the case for the majority of the world's languages. Thus, this question has practical relevance.

Setup. To simulate different low-resource settings for this experiment, we use the first $n \in \{2500, 5000, 7500, 10000, 12500\}$ sentences from each training set. Since the PRPN has shown to be largely robust to hyperparameter changes and we stop training based on the training loss, development sets are not used. Test sets are kept unmodified. As before, all results are averaged over 5 training runs with different random seeds.

Results. The learning curves in Figure 4 show the F1 scores of all models as a function of the amount of sentences available for training. We observe the following:

- While performance for English strongly increases with more training instances, this is not the case for the remaining languages. The Chinese F1 score increases only slightly for additional examples. The performance for Arabic and German is roughly constant. The fact that the PRPN is better at leveraging additional English data is another indicator that the PRPN might be better suited for English than for the other three languages in our experiments.

- Comparing Figure 4 to Table 2, we see a gap between the PRPN's performances for 12,500 examples and for the entire training sets for all languages. Thus, we conclude that disposing of more than 12,500 examples is generally beneficial.

5 Multilingual Experiment

5.1 Motivation

A multilingual model for unsupervised constituency parsing has desirable benefits. First, since only one model is needed for a set of languages, less memory is required to store all parameters. This facilitates, for instance, the application on mobile devices. Second, neural models which have been trained simultaneously on multiple languages have been shown to leverage knowledge from related languages to improve performance on other languages in the case of limited training data. Such cross-lingual transfer has been successfully employed for a variety of tasks, e.g., for supervised dependency parsing (de Lhoneux et al., 2018), for machine translation (Johnson et al., 2017a), or for paradigm completion (McCarthy et al., 2019).

5.2 Setup

Since transfer learning is mostly needed and also particularly effective in the low-data regime, we combine the training sets of 2,500 examples for English and German—the two of our languages which are related—to form a multilingual training set. We then train PRPN models on this combined training set, sharing all parameters.

The model's hyperparameters remain the same as before, and we again train for a maximum of 150 epochs, or until the training loss has not decreased for 5 consecutive epochs. As for our previous experiments, the reported results are averaged over 5 training runs with different random seeds.

5.3 Results

Results for single-language models as well as the multilingual versions are shown in Table 3. The parsing performance of the multilingual model is 4.0% and 2.4% higher than that of the single-language models for English and German, respectively. Thus, the PRPN model indeed benefits from transfer learning, i.e., it can share information across related languages for unsupervised constituency parsing.

	single language	multilingual
English	0.186 (.08)	0.226 (.03)
German	0.326 (.02)	0.350 (.06)

Table 3: Monolingual and multilingual PRPN test results for 2500 training sentences. Results are averaged over 5 training runs, with standard deviations in parentheses.

6 Related Work

Unsupervised parsing. Previous work on non-neural models for unsupervised parsing includes Clark (2001) and Klein and Manning (2002) for constituency parsing and Carroll and Charniak (1992); Klein and Manning (2004); Cohn et al. (2010); Spitkovsky et al. (2011); and Jiang et al. (2016) for dependency parsing. For Chinese, German, and English, previous work also observed differences in F1 scores; an overview can be found in Bod (2006). We have not included these non-neural baselines as part of our results due to lack of availability of trustworthy implementations.

Following the success of the neural PRPN model in 2018, various other neural unsupervised parsing approaches have been developed and shown promising results. Shen et al. (2018c) enhance the vanilla LSTM network with master forget and input gates to learn the tree structure through soft gating. Drozdov et al. (2019) use a recursive autoencoder-based architecture. Kim et al. (2019b) employ unsupervised recurrent neural network grammars, and Kim et al. (2019a) employ compound probabilistic context free grammars. Shi et al. (2019) show how image captions can be successfully leveraged to identify constituents in sentences. None of these papers performs an explicit analysis of differences between languages.

Jin et al. (2019) extend the PCFG approach to show results on Chinese, English and German. There are certain question that remain unanswered about multilingual grammar induction, especially related to cross-lingual transfer and difference in hyper parameters. In this work, we focus on adapting the PRPN to a multilingual setting, since it is the first neural model which has been shown to obtain robust unsupervised parsing results. Although we have primarily focused on PRPN due to its overall success, it would be interesting to observe whether similar trends in relative performance among languages hold for other models mentioned above. We leave this for future work.

A closely related line of research, which is often referred to as *latent tree learning*, aims to create a parse structure which is well-suited for a particular NLP application. Common choices are sentence classification tasks like natural language inference (Yogatama et al., 2017; Maillard et al., 2017; Choi et al., 2018), machine translation (Bisk and Tran, 2018), or toy datasets where the correct parse can trivially be found by humans (Jacob et al., 2018; Nangia and Bowman, 2018). Latent tree learning models have been shown to outperform sequential models and TreeRNNs on multiple datasets (Maillard et al., 2017; Choi et al., 2018). However, the parses predicted by latent tree models have been shown to mostly be nonsensical (Williams et al., 2018).

Supervised parsing. This research is further related to the line of work on supervised parsing. Two main parsing paradigms exist: dependency parsing, which is concerned with the relationships between words in a sentence, and constituency parsing (or *phrase-structure parsing*), which is what we are interested in here (cf. Figure 1). Neural network models have pushed the state of the art for supervised constituency parsing in the last years. Possible approaches include methods to either build parse trees sequentially by estimating transition probabilities (Zhu et al., 2013; Cross and Huang, 2016), employ a chart-based approach, which performs exact structured inference via dynamic programming (Durrett and Klein, 2015; Stern et al., 2017), or cast the problem as a sequence labeling task (Gómez-Rodríguez and Vilares, 2018). Another, rather new option is to predict syntactic distances between words, which can then be converted into trees (Shen et al., 2018b). This is the same core concept that the PRPN is based on. Thus, we consider Shen et al. (2018b)'s approach one of our upper bounds on the unsupervised parsing performance of the PRPN.

Cross-lingual transfer. Cross-lingual transfer (Wu, 1997; Yarowsky et al., 2001), i.e., using knowledge gained from one (usually high-resource) language for solving a task in another (usually low-resource) language, is very common when working on resource-poor languages in NLP. There are two very intuitive ways of realizing such a transfer (Liu et al., 2019): One way is to translate the test data into a high-resource language and to solve the task using a system for that second language. Another way is to translate large amount of training data into a low-resource language and train a system in that language. Other methods have been developed as well, many based on parameter sharing—as we do in this work—, e.g., for cross-lingual natural language inference (Conneau et al., 2018), morphological generation (Kann et al., 2017; McCarthy et al., 2019), dialogue systems (Schuster et al., 2019), or machine translation (Johnson et al., 2017b; Aharoni et al., 2019). While we are not aware of any previous work exploring cross-lingual transfer for unsupervised parsing as done in this paper, approaches have been developed which leverage high-resource language data for *supervised* parsing in low-resource languages (Søgaard, 2011; Naseem et al., 2012).

7 Conclusion

We investigated the behavior of the PRPN, a neural unsupervised constituency parsing model, for the languages Arabic, Chinese, English, and German. While, overall, our experiments showed that the model strongly outperformed trivial baselines for all the languages, we made the following additional observations: (i) With regards to its parsing performance, the model is robust to hyperparameter changes for all four languages. (ii) Parsing F1 and language modeling perplexity were not correlated. (iii) The PRPN's unsupervised parsing performance differed a lot between languages, while trivial baselines and upper bounds obtained similar scores. (iv) The model was able to leverage additional training data for English better in low-resource settings, and for no language did the PRPN reach its maximum observed F1 score with 12,500 training instances. Finally, we proposed to train a multilingual PRPN model for the related languages English and German. Besides requiring less parameters, this led to an up to 4% higher F1 score in the low-data regime.

Acknowledgments

This work has benefited from the support of Samsung Research under the project *Improving Deep Learning using Latent Structure* and from the donation of a Titan V GPU by NVIDIA Corporation.

References

Roee Aharoni and Yoav Goldberg. 2017. Towards string-to-tree neural machine translation. In *ACL*.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL*.

Yonatan Bisk and Ke Tran. 2018. Inducing grammars with and for neural machine translation. In *WNMT*.

Rens Bod. 2006. An all-subtrees approach to unsupervised parsing. In *ACL*.

Glenn Carroll and Eugene Charniak. 1992. *Two experiments on learning probabilistic dependency grammars from corpora*. AAAI Technical Report.

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Learning to compose task-specific tree structures. In *AAAI*.

Alexander Clark. 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *CoNLL*.

Trevor Cohn, Phil Blunsom, and Sharon Goldwater. 2010. Inducing tree-substitution grammars. *JMLR*, 11:3053–3096.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *EMNLP*.

James Cross and Liang Huang. 2016. Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *EMNLP*.

Andrew Drozdov, Pat Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised latent tree induction with deep inside-outside recursive autoencoders. *arXiv:1904.02142*.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for german using sister-head dependencies. In *ACL*.

Greg Durrett and Dan Klein. 2015. Neural CRF parsing. In *ACL–IJCNLP*.

Carlos Gómez-Rodríguez and David Vilares. 2018. Constituent parsing as sequence labeling. In *EMNLP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phu Mon Htut, Kyunghyun Cho, and Samuel R Bowman. 2018. Grammar induction with neural language models: An unusual replication. In *EMNLP*.

Athul Paul Jacob, Zhouhan Lin, Alessandro Sordoni, and Yoshua Bengio. 2018. Learning hierarchical structures on-the-fly with a recurrent-recursive model for sequences. In *Repl4NLP*.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *EMNLP*.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019. Unsupervised learning of pcfgs with normalizing flow. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2442–2452.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017a. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017b. Googles multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *ACL*.

Yoon Kim, Chris Dyer, and Alexander M Rush. 2019a. Compound probabilistic context-free grammars for grammar induction. *arXiv:1906.10225*.

Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019b. Unsupervised recurrent neural network grammars. *arXiv:1904.03746*.

Dan Klein and Christopher D Manning. 2002. A generative constituent-context model for improved grammar induction. In *ACL*.

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

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *EMNLP*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *ACL*.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*.

Jean Maillard, Stephen Clark, and Dani Yogatama. 2017. Jointly learning sentence embeddings and syntax with unsupervised tree-LSTMs. In *ICLR*.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42. *CD-ROM. Philadelphia, Penn.: Linguistic Data Consortium*.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Miikka Silfverberg, Sebastian J Mielke, Jeffrey Heinz, Ryan Cotterell, et al. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *SIGMORPHON*.

Nikita Nangia and Samuel Bowman. 2018. ListOps: A diagnostic dataset for latent tree learning. In *NAACL-SRW*.

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

Fernando Pereira and Yves Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *ACL*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL*.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018a. Neural language modeling by jointly learning syntax and lexicon. In *ICLR*.

Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, and Yoshua Bengio. 2018b. Straight to the tree: Constituency parsing with neural syntactic distance. In *ACL*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018c. Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv:1810.09536*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. *arXiv:1906.02890*.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *ANLP*.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *ACL–IJCNLP*.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *ICML*.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *NAACL–HLT*.

Valentin I Spitkovsky, Hiyan Alshawi, Angel X Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *EMNLP*.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *ACL*, volume 1, pages 818–827.

Adina Williams, Andrew Drozdov, and Samuel R Bowman. 2018. Do latent tree learning models identify meaningful structure in sentences? *TACL*, 6:253–267.

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

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

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

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. 2017. Learning to compose words into sentences with reinforcement learning. In *ICLR*.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *ACL*.