Compound Probabilistic Context-Free Grammars
for Grammar Induction

Yoon Kim
Harvard University
Cambridge, MA, USA
yoonkim@seas.harvard.edu

Chris Dyer
DeepMind
London, UK
cdyer@google.com

Alexander M. Rush
Harvard University
Cambridge, MA, USA
srush@seas.harvard.edu

Abstract

We study a formalization of the grammar induction problem that models sentences as being generated by a compound probabilistic context-free grammar. In contrast to traditional formulations which learn a single stochastic grammar, our context-free rule probabilities are modulated by a per-sentence continuous latent variable, which induces marginal dependencies beyond the traditional context-free assumptions. Inference in this grammar is performed by collapsed variational inference, in which an amortized variational posterior is placed on the continuous variable, and the latent trees are marginalized with dynamic programming. Experiments on English and Chinese show the effectiveness of our approach compared to recent state-of-the-art methods for grammar induction from words with neural language models.

1 Introduction

Grammar induction is the task of inducing hierarchical syntactic structure from data. Statistical approaches to grammar induction require specifying a probabilistic grammar (e.g. formalism, number and shape of rules), and fitting its parameters through optimization. Early work found that it was difficult to induce probabilistic context-free grammars (PCFG) from natural language data through direct methods, such as optimizing the log likelihood with the EM algorithm (Lari and Young, 1990; Carroll and Charniak, 1992). While the reasons for the failure are manifold and not completely understood, two major potential causes are the ill-behaved optimization landscape and the overly strict independence assumptions of PCFGs. More successful approaches to grammar induction have thus resorted to carefully-crafted auxiliary objectives (Klein and Manning, 2002), priors or non-parametric models (Kurihara and Sato, 2006; Johnson et al., 2007; Liang et al., 2007; Wang and Blunsom, 2013), and manually-engineered features (Huang et al., 2012; Golland et al., 2012) to encourage the desired structures to emerge.

We revisit these aforementioned issues in light of advances in model parameterization and inference. First, contrary to common wisdom, we find that parameterizing a PCFG’s rule probabilities with neural networks over distributed representations makes it possible to induce linguistically meaningful grammars by simply optimizing log likelihood. While the optimization problem remains non-convex, recent work suggests that there are optimization benefits afforded by over-parameterized models (Arora et al., 2018; Xu et al., 2018; Du et al., 2019), and we indeed find that this neural PCFG is significantly easier to optimize than the traditional PCFG. Second, this factored parameterization makes it straightforward to incorporate side information into rule probabilities through a sentence-level continuous latent vector, which effectively allows different contexts in a derivation to coordinate. In this compound PCFG—continuous mixture of PCFGs—the context-free assumptions hold conditioned on the latent vector but not unconditionally, thereby obtaining longer-range dependencies within a tree-based generative process.

To utilize this approach, we need to efficiently optimize the log marginal likelihood of observed sentences. While compound PCFGs break efficient inference, if the latent vector is known the distribution over trees reduces to a standard PCFG. This property allows us to perform grammar induction using a collapsed approach where the latent trees are marginalized out exactly with dynamic programming. To handle the latent vector, we employ standard amortized inference using reparameterized samples from a variational
posterior approximated from an inference network (Kingma and Welling, 2014; Rezende et al., 2014).

On standard benchmarks for English and Chinese, the proposed approach is found to perform favorably against recent neural network-based approaches to grammar induction (Shen et al., 2018, 2019; Drozdov et al., 2019; Kim et al., 2019).

2 Probabilistic Context-Free Grammars

We consider context-free grammars (CFG) consisting of a 5-tuple \( G = (S, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R}) \) where \( S \) is the distinguished start symbol, \( \mathcal{N} \) is a finite set of nonterminals, \( \mathcal{P} \) is a finite set of preterminals, \( \Sigma \) is a finite set of terminal symbols, and \( \mathcal{R} \) is a finite set of rules of the form,

\[
S \rightarrow A, \quad A \in \mathcal{N} \\
A \rightarrow BC, \quad A \in \mathcal{N}, \ B, C \in \mathcal{N} \cup \mathcal{P} \\
T \rightarrow w, \quad T \in \mathcal{P}, \ w \in \Sigma.
\]

A probabilistic context-free grammar (PCFG) consists of a grammar \( G \) and rule probabilities \( \pi = \{\pi_r\}_{r \in \mathcal{R}} \) such that \( \pi_r \) is the probability of the rule \( r \). Letting \( T_G \) be the set of all parse trees of \( G \), a PCFG defines a probability distribution over \( t \in T_G \) via \( p_\pi(t) = \Pi_{r \in \mathcal{R}} \pi_r \) where \( \mathcal{R} \) is the set of rules used in the derivation of \( t \). It also defines a distribution over string of terminals \( x \in \Sigma^* \) via

\[
p_\pi(x) = \sum_{t \in T_G(x)} p_\pi(t),
\]

where \( T_G(x) = \{ t | yield(t) = x \} \), i.e. the set of trees \( t \) such that \( t \)'s leaves are \( x \). We will use \( p_\pi(t | x) \triangleq p_\pi(t \mid yield(t) = x) \) to denote the posterior distribution over latent trees given the observed sentence \( x \).

Parameterization

The standard way to parameterize a PCFG is to simply associate a scalar to each rule \( \pi_r \) with the constraint that they form valid probability distributions, i.e. each nonterminal is associated with a fully-parameterized categorical distribution over its rules. This direct parameterization is algorithmically convenient since the M-step in the EM algorithm (Dempster et al., 1977) has a closed form. However, there is a long history of work showing that it is difficult to learn meaningful grammars from natural language data with this parameterization (Charniak, 1992).\(^2\) Successful approaches to unsupervised parsing have therefore modified the model/learning objective by guiding potentially unrelated rules to behave similarly.

Recognizing that sharing among rule types is beneficial, we propose a neural parameterization where rule probabilities are based on distributed representations. We associate embeddings with each symbol, introducing input embeddings \( w_N \) for each symbol \( N \) on the left side of a rule (i.e. \( N \in \{S\} \cup \mathcal{N} \cup \mathcal{P} \)). For each rule type \( r \), \( \pi_r \) is parameterized as follows,

\[
\begin{align*}
\pi_{S \rightarrow A} &= \frac{\exp(u_A^T f_1(w_S))}{\sum_{A' \in \mathcal{N}} \exp(u_{A'}^T f_1(w_S))}, \\
\pi_{A \rightarrow BC} &= \frac{\exp(u_{BC}^T w_A)}{\sum_{B'C' \in M} \exp(u_{B'C'}^T w_A)}, \\
\pi_{T \rightarrow w} &= \frac{\exp(u_{w}^T f_2(w_T))}{\sum_{w' \in \Sigma} \exp(u_{w'}^T f_2(w_T))},
\end{align*}
\]

where \( M \) is the product space \( (\mathcal{N} \cup \mathcal{P}) \times (\mathcal{N} \cup \mathcal{P}) \), and \( f_1, f_2 \) are MLPs with two parameter layers (see appendix A.1 for the full parameterization). We will use \( E_G = \{w_N | N \in \{S\} \cup \mathcal{N} \cup \mathcal{P} \} \) to denote the set of input symbol embeddings for a grammar \( G \), and \( \lambda \) to refer to the parameters of the neural network used to obtain the rule probabilities. A graphical model-like illustration of the neural PCFG is shown in Figure 1 (left).

It is clear that the neural parameterization does not change the underlying probabilistic assumptions. The difference between the two is analogous to the difference between count-based vs. feed-forward neural language models, where feed-forward neural language models make the same Markov assumptions as the count-based models but are able to take advantage of shared, distributed representations.

3 Compound PCFGs

A compound probability distribution (Robbins, 1951) is a distribution whose parameters are themselves random variables. These distributions generalize mixture models to the continuous case, for example in factor analysis which assumes the following generative process,

\[
z \sim \mathcal{N}(0, I) \quad x \sim \mathcal{N}(Wz, \Sigma).
\]

Compound distributions provide the ability to model rich generative processes, but marginalizing over the latent parameter can be computationally expensive unless conjugacy can be exploited.

\(^1\)Since we will be inducing a grammar directly from words, \( \mathcal{P} \) is roughly the set of part-of-speech tags and \( \mathcal{N} \) is the set of constituent labels. However, to avoid issues of label alignment, evaluation is only on the tree topology.

\(^2\)In preliminary experiments we were indeed unable to learn linguistically meaningful grammars with this PCFG.
In this work, we study compound probabilistic context-free grammars whose distribution over trees arises from the following generative process: we first obtain rule probabilities via

$$z \sim p_\gamma(z), \quad \pi_z = f_\lambda(z, E_G),$$

where $p_\gamma(z)$ is a prior with parameters $\gamma$ (spherical Gaussian in this paper), and $f_\lambda$ is a neural network that concatenates the input symbol embeddings with $z$ and outputs the sentence-level rule probabilities $\pi_z$,

$$
\pi_{z,S \rightarrow A} \propto \exp(\mathbf{u}_A^T f_1([w_S; z])), \\
\pi_{z,A \rightarrow BC} \propto \exp(\mathbf{u}_{BC}^T [w_A; z]), \\
\pi_{z,T \rightarrow w} \propto \exp(\mathbf{u}_{w}^T f_2([w_T; z]),
$$

where $[w; z]$ denotes vector concatenation. Then a tree/sentence is sampled from a PCFG with rule probabilities given by $\pi_z$,

$$t \sim \text{PCFG}(\pi_z), \quad x = \text{yield}(t).$$

This can be viewed as a continuous mixture of PCFGs, or alternatively, a Bayesian PCFG with a prior on sentence-level rule probabilities parameterized by $z, \lambda, E_G$. Importantly, under this generative model the context-free assumptions hold conditioned on $z$, but they do not hold unconditionally. This is shown in Figure 1 (right) where there is a dependence path through $z$ if it is not conditioned upon. Compound PCFGs give rise to a marginal distribution over parse trees $t$ via

$$p(t) = \int p(t | z)p_\gamma(z) \, dz,$$

where $p(t | z) = \prod_{r \in t} \pi_{z,r}$. The subscript in $\pi_{z,r}$ denotes the fact that the rule probabilities depend on $z$. Compound PCFGs are clearly more expressive than PCFGs as each sentence has its own set of rule probabilities. However, it still assumes a tree-based generative process, making it possible to learn latent tree structures.

Our motivation for the compound PCFG is based on the observation that for grammar induction, first-order context-free assumptions are generally made not because they represent an adequate model of natural language, but because they allow for tractable training. Higher-order PCFGs can introduce dependencies between children and ancestors/siblings through, for example, vertical/horizontal Markovization (Johnson, 1998; Klein and Manning, 2003). However such dependencies complicate training due to the rapid increase in the number of rules. Under this view, we can interpret the compound PCFG as a restricted version of some higher-order PCFG where a child can depend on its ancestors and siblings through a shared latent vector. We hypothesize that this

---

4 A piece of evidence for the misspecification of first-order PCFGs as a statistical model of natural language is that if one pretrains a first-order PCFG on supervised data and continues training with the unsupervised objective (i.e. log marginal likelihood), the resulting grammar deviates significantly from the supervised initial grammar while the log marginal likelihood improves (Johnson et al., 2007). Similar observations have been made for part-of-speech induction with Hidden Markov Models (Meria1do, 1994).

5 Another interpretation of the compound PCFG is to view it as a vectorized version of indexed grammars (Aho, 1968), which extend CFGs by augmenting nonterminals with additional index strings that may be inherited or modified during derivation. Compound PCFGs instead equip nonterminals with a continuous vector that is always inherited.

---

Under the Bayesian PCFG view, $p_\gamma(z)$ is a distribution over $z$ (a subset of the prior), and thus a hyperprior.
dependence among siblings is especially useful in grammar induction from words, where (for example) if we know that watched is used as a verb then the noun phrase is likely to be a movie.

In contrast to the usual Bayesian treatment of PCFGs which places priors on global rule probabilities (Kurihara and Sato, 2006; Johnson et al., 2007; Wang and Blunsom, 2013), the compound PCFG assumes a prior on local, sentence-level rule probabilities. It is therefore closely related to the Bayesian grammars studied by Cohen et al. (2009) and Cohen and Smith (2009), who also sample local rule probabilities from a logistic normal prior for training dependency models with valence (DMV) (Klein and Manning, 2004).

**Inference in Compound PCFGs** The expressivity of compound PCFGs comes at a significant challenge in learning and inference. Letting $\theta = \{E_\theta, \lambda\}$ be the parameters of the generative model, we would like to maximize the log marginal likelihood of the observed sentence

$$\log p_\theta(x).$$

In the neural PCFG the log marginal likelihood $\log p_\theta(x) = \log \sum_{t \in T_G(x)} p_\theta(t)$ can be obtained by summing out the latent tree structure using the inside algorithm (Baker, 1979), which is differentiable and thus amenable to gradient-based optimization. In the compound PCFG, the log marginal likelihood is given by

$$\log p_\theta(x) = \log \left( \int \sum_{t \in T_G(x)} p_\theta(t \mid z) p_\gamma(z) \, dz \right).$$

Notice that while the integral over $z$ makes this quantity intractable, when we are condition on $z$, we can tractably perform the inner summation as before using the inside algorithm. We therefore resort to collapsed amortized variational inference. We first obtain a sample $z$ from a variational posterior distribution (given by an amortized inference network), then perform the inner marginalization conditioned on this sample. The evidence lower bound ELBO($\theta, \phi; x$) is then given by,

$$\mathbb{E}_{q_\phi(z \mid x)} \left[ \log p_\theta(x \mid z) \right] - \text{KL}[q_\phi(z \mid x) \parallel p_\gamma(z)],$$

and we can calculate $p_\theta(x \mid z) = \sum_{t \in T_G(x)} p(t \mid z)$ with the inside algorithm given a sample $z$ from a variational posterior $q_\phi(z \mid x)$. For the variational family we use a diagonal Gaussian where the mean/log-variance vectors are given by an affine layer over max-pooled hidden states from an LSTM over $x$. We can obtain low-variance estimators for the gradient $\nabla_{\theta, \phi} \text{ELBO}(\theta, \phi; x)$ by using the reparameterization trick for the expected reconstruction likelihood and the analytical expression for the KL term (Kingma and Welling, 2014).

We remark that under the Bayesian PCFG view, since the parameters of the prior (i.e. $\theta$) are estimated from the data, our approach can be seen as an instance of empirical Bayes (Robbins, 1956).6

**MAP Inference** After training, we are interested in comparing the learned trees against an annotated treebank. This requires inferring the most likely tree given a sentence, i.e. $\text{argmax}_t p_\theta(t \mid x)$. For the neural PCFG we can obtain the most likely tree by using the Viterbi version of the inside algorithm (CKY algorithm). For the compound PCFG, the argmax is intractable to obtain exactly, and hence we estimate it with the following approximation,

$$\text{argmax}_t \int p_\theta(t \mid x, z) p_\gamma(z \mid x) \, dz \approx \text{argmax}_t p_\theta(t \mid x, \mu_\phi(x)),$$

where $\mu_\phi(x)$ is the mean vector from the inference network. The above approximates the true posterior $p_\theta(z \mid x)$ with $\delta(z - \mu_\phi(x))$, the Dirac delta function at the mode of the variational posterior.7 This quantity is tractable to estimate as in the PCFG case. Other approximations are possible: for example we could use $q_\phi(z \mid x)$ as an importance sampling distribution to estimate the first integral. However we found the above approximation to be efficient and effective in practice.

### 4 Experimental Setup

**Data** We test our approach on the Penn Treebank (PTB) (Marcus et al., 1993) with the standard splits (2-21 for training, 22 for validation, 23 for test) and the same preprocessing as in recent works (Shen et al., 2018, 2019), where we discard punctuation, lowercase all tokens, and take the top 10K most frequent words as the vocabulary. This setup is more challenging than traditional setups, which usually experiment on shorter sentences and use gold part-of-speech tags.

We further experiment on Chinese with version 5.1 of the Chinese Penn Treebank (CTB) (Xue et al., 2005), with the same splits as in Chen and Manning (2014). On CTB we also remove punctuation and keep the top 10K word types.

---

6See Berger (1985) (chapter 4), Zhang (2003), and Cohen (2016) (chapter 3) for further discussion on compound models and empirical Bayes.

7Since $p_\theta(t \mid x, z)$ is continuous with respect to $z$, we have $\int p_\theta(t \mid x, z) \delta(z - \mu_\phi(x)) \, dz = p_\theta(t \mid x, \mu_\phi(x))$. 

Hyperparameters Our PCFG uses 30 nonterminals and 60 preterminals, with 256-dimensional symbol embeddings. The compound PCFG uses 64-dimensional latent vectors. The bidirectional LSTM inference network has a single layer with 512 dimensions, and the mean and the log variance vector for $q_\phi(z|\mathbf{x})$ are given by max-pooling the hidden states of the LSTM and passing it through an affine layer. Model parameters are initialized with Xavier uniform initialization. For training we use Adam (Kingma and Ba, 2015) with $\beta_1 = 0.75$, $\beta_2 = 0.999$ and learning rate of 0.001, with a maximum gradient norm limit of 3. We train for 10 epochs with batch size equal to 4. We employ a curriculum learning strategy (Bengio et al., 2009) where we train only on sentences of length up to 30 in the first epoch, and increase this length limit by 1 each epoch. This slightly improved performance and similar strategies have used in the past for grammar induction (Spitkovsky et al., 2012). During training we perform early stopping based on validation perplexity. To mitigate against overfitting to PTB, experiments on CTB utilize the same hyperparameters from PTB.

Baselines and Evaluation We observe that even on PTB, there is enough variation in setups across prior work on grammar induction to render a meaningful comparison difficult. Some important dimensions along which prior works vary include, (1) lexicalization: earlier work on grammar induction generally assumed gold (or induced) part-of-speech tags (Klein and Manning, 2004; Smith and Eisner, 2004; Bod, 2006; Snyder et al., 2009), while more recent works induce grammar directly from words (Spitkovsky et al., 2013; Shen et al., 2018); (2) use of punctuation: even within papers that induce a grammar directly from words, some papers employ heuristics based on punctuation as punctuation is usually a strong signal for start/end of constituents (Segner, 2007; Ponvert et al., 2011; Spitkovsky et al., 2013), some train with punctuation (Jin et al., 2018; Drozdov et al., 2019; Kim et al., 2019), while others discard punctuation altogether for training (Shen et al., 2018, 2019); (3) train/test data: some works do not explicitly separate out train/test sets (Reichart and Rappoport, 2010; Golland et al., 2012) while some do (Huang et al., 2012; Parikh et al., 2014; Htut et al., 2018). Maintaining train/test splits is less of an issue for unsupervised structure learning, however in this work we follow the latter and separate train/test data. (4) evaluation: for unlabeled $F_1$, almost all works ignore punctuation (even approaches that use punctuation during training typically ignore them during evaluation), but there is some variance in discarding trivial spans (width-one and sentence-level spans) and using corpus-level versus sentence-level $F_1$. In this paper we discard trivial spans and evaluate on sentence-level $F_1$ per recent work (Shen et al., 2018, 2019).

Given the above, we mainly compare our approach against two recent, strong baselines with open source code: Parsing Predict Reading Network (PRPN) (Shen et al., 2018) and Ordered Neurons (ON) (Shen et al., 2019). These approaches train a neural language model with gated attention-like mechanisms to induce binary trees, and achieve strong unsupervised parsing performance even when trained on corpora where punctuation is removed. Since the original results were on both language modeling and grammar induction, their hyperparameters were presumably tuned to do well on both and thus may not be optimal for just unsupervised parsing. We therefore tune the hyperparameters of these baselines for unsupervised parsing only (i.e. on validation $F_1$).

5 Results and Discussion

Table 1 shows the unlabeled $F_1$ scores for our models and various baselines. All models soundly outperform right branching baselines, and we find that the neural PCFG/compound PCFG are strong models for grammar induction. In particular the compound PCFG outperforms other models by an appreciable margin on both English and Chinese. We again note that we were unable to induce meaningful grammars through a traditional PCFG with the scalar parameterization. See appendix A.2 for the full results (including corpus-level $F_1$) broken down by sentence length.

Table 2 analyzes the learned tree structures. We compare similarity as measured by $F_1$ against

---

8 However, we used $F_1$ against validation trees on PTB to select some hyperparameters (e.g. grammar size), as is sometimes done in grammar induction. Hence our PTB results are arguably not fully unsupervised in the strictest sense of the term. The hyperparameters of the PRPN/ON baselines are also tuned using validation $F_1$ for fair comparison.

9 Corpus-level $F_1$ calculates precision/recall at the corpus level to obtain $F_1$, while sentence-level $F_1$ calculates $F_1$ for each sentence and averages across the corpus.

10 https://github.com/yikangshen/PRPN

11 https://github.com/yikangshen/Ordered-Neurons

12 The training perplexity was much higher than in the neural case, indicating significant optimization issues.
contrast, both PRPN and ON perform as well as an π of an NLM by additionally conditioning the terminal language model (LM) obtains 86.2 (Table 3).

While induced trees for downstream tasks

**Induced Trees for Downstream Tasks** While the compound PCFG has fewer independence assumptions than the neural PCFG, it is still a more constrained model of language than standard neural language models (NLM) and thus not competitive in terms of perplexity: the compound PCFG obtains a perplexity of 196.3 while an LSTM language model (NLM) obtains 86.2 (Table 3).

In contrast, both PRPN and ON perform as well as an gold, left, right, and “self” trees (top), where self $F_1$ score is calculated by averaging over all 6 pairs obtained from 4 different runs. We find that PRPN is particularly consistent across multiple runs. We also observe that different models are better at identifying different constituent labels, as measured by label recall (Table 2, bottom). While left as future work, this naturally suggests an ensemble approach wherein the empirical probabilities of constituents (obtained by averaging the predicted binary constituent labels from the different models) are used either to supervise another model or directly as potentials in a CRF constituency parser. Finally, all models seemed to have some difficulty in identifying SBAR/VP constituents which typically span more words than NP constituents.

| Model               | PTB Mean | PTB Max | CTB Mean | CTB Max |
|---------------------|----------|---------|----------|---------|
| PRPN (Shen et al., 2018) | 37.4     | 38.1    | –        | –       |
| ON (Shen et al., 2019)     | 47.7     | 49.4    | –        | –       |
| URNNG (Kim et al., 2019) | –        | 45.4    | –        | –       |
| DIORA (Drozdov et al., 2019) | –        | 58.9    | –        | –       |
| Left Branching        | 8.7      | 9.7     |          |         |
| Right Branching       | 39.5     | 20.0    |          |         |
| Random Trees          | 19.2     | 19.5    | 15.7     | 16.0    |
| PRPN (tuned)          | 47.3     | 47.9    | 30.4     | 31.5    |
| ON (tuned)            | 48.1     | 50.0    | 25.4     | 25.7    |
| Neural PCFG           | 50.8     | 52.6    | 25.7     | 29.5    |
| Compound PCFG         | 55.2     | 60.1    | 36.0     | 39.8    |
| Oracle Trees          | 84.3     | 81.1    |          |         |

Table 1: Unlabeled sentence-level $F_1$ scores on PTB and CTB test sets. Top shows results from previous work while the rest of the results are from this paper. Mean/Max scores are obtained from 4 runs of each model with different random seeds. Oracle is the maximum score obtainable with binarized trees, since we compare against the non-binarized gold trees per convention. Results with † are trained on a version of PTB with punctuation, and hence not strictly comparable to the present work. For URNNG/DIORA, we take the parsed test set provided by the authors from their best runs and evaluate $F_1$ with our evaluation setup, which ignores punctuation.

Table 2: (Top) Mean $F_1$ similarity against Gold, Left, Right, and Self trees. Self $F_1$ score is calculated by averaging over all 6 pairs obtained from 4 different runs. (Bottom) Fraction of ground truth constituents that were predicted as a constituent by the models broken down by label (i.e. label recall).

LSTM LM while maintaining good unsupervised parsing performance.

We thus experiment to see if it is possible to use the induced trees to supervise a more flexible generative model that can make use of tree structures—namely, recurrent neural network grammars (RNNG) (Dyer et al., 2016). RNNGs are generative models of language that jointly model syntax and surface structure by incrementally generating a syntax tree and sentence. As with NLMs, RNNGs make no independence assumptions, and have been shown to outperform NLMs in terms of perplexity and grammaticality judgment when trained on gold trees (Kuncoro et al., 2018; Wilcox et al., 2019). We take the best run from each model and parse the training set, and use the induced trees to supervise an RNNG for each model using the parameterization from Kim et al. (2019). We are also interested in syntactic evaluation of our models, and for this we utilize the framework and dataset from Marvin and Linzen (2018), where a model is presented two minimally different sentences such as:

the senators near the assistant are old

*the senators near the assistant is old

and must assign higher probability to grammatical sentence. Additionally, Kim et al. (2019) report perplexity improvements by fine-tuning an RNNG trained on gold trees with the unsupervised RNNG (URNNG)—whereas the RNNG is is trained to maximize the joint likelihood $\log p(x, t)$, the URNNG maximizes a lower bound on the log marginal likelihood $\log \sum_{t} p(x, t)$ with a structured inference network that approximates the true

\[ \frac{1}{\lambda} \log \sum_{\tau} \prod_{w \in \tau} \frac{\exp(u_{w} f_{\lambda}(w; x; h_{\lambda}))}{\sum_{u_{w}} \exp(u_{w} f_{\lambda}(w; x; h_{\lambda}))} \]

where $h_{\lambda}$ is the hidden state from an LSTM over $x_{<t}$. However the unsupervised parsing performance was far worse ($\approx 25 F_1$ on the PTB).

---

13 We did manage to almost match the perplexity of an NLM by additionally conditioning the terminal probabilities on previous history, i.e. $p_{w, \tau \rightarrow \tau'} \propto \exp(u_{w} f_{\lambda}(w; x; h_{\lambda}))$ where $h_{\lambda}$ is the hidden state from an LSTM over $x_{<t}$. However the unsupervised parsing performance was far worse ($\approx 25 F_1$ on the PTB).

14 The train/test $F_1$ was similar for all models.

15 https://github.com/harvardnlp/urnng
Table 3: Results from training RNNGs on induced trees from various models (Induced RNNG) on the PTB. Induced URNNG indicates fine-tuning with the URNNG objective. We show perplexity (PPL), grammaticality judgment performance (Syntactic Eval.), and unlabeled $F_1$. PPL/$F_1$ are calculated on the PTB test set and Syntactic Eval. is from Marvin and Linzen (2018)’s dataset. Results on top do not make any use of annotated trees, while the bottom two results are trained on binarized gold trees. Note that the perplexity numbers here are not comparable to standard results on the PTB since our models are generative model of sentences and hence we do not carry information across sentence boundaries.

 posterior. We experiment with a similar approach where we fine-tune RNNGs trained on induced trees with URNNGs. We perform early stopping for both RNNG and URNNG based on validation perplexity. See appendix A.3 for further details regarding the experimental setup.

The results are shown in Table 3. For perplexity, RNNGs trained on induced trees (Induced RNNG in Table 3) are unable to improve upon an LSTM LM, in contrast to the supervised RNNG which does outperform the LSTM language model (Table 3, bottom). For grammaticality judgment however, the RNNG trained with compound PCFG trees outperforms the LSTM LM despite obtaining worse perplexity, and performs on par with the RNNG trained on gold trees. Fine-tuning with the URNNG results in improvements in perplexity and grammaticality judgment across the board (Induced URNNG in Table 3). We also obtain large improvements on unsupervised parsing as measured by $F_1$, with the fine-tuned URNNGs outperforming the respective original models. This is potentially due to an ensembling effect between the original model and the URNNG’s structured inference network, which is parameterized as a neural CRF constituency parser (Durrett and Klein, 2015; Liu et al., 2018).

Model Analysis We analyze our best compound PCFG model in more detail. Since we induce a full set of nonterminals in our grammar, we can analyze the learned nonterminals to see if they can be aligned with linguistic constituent labels. Figure 2 visualizes the alignment between induced and gold labels, where for each nonterminal we show the empirical probability that a predicted constituent of this type will correspond to a particular linguistic constituent in the test set, conditioned on its being a correct constituent (for reference we also show the precision). We observe that some of the induced nonterminals clearly align to linguistic nonterminals. Further results, including preterminal alignments to part-of-speech tags, are shown in appendix A.4.

19While left as future work, it is possible to use the compound PCFG itself as an inference network. Also note that the $F_1$ scores for the URNNGs in Table 3 are optimistic since we selected the best-performing runs of the original models based on validation $F_1$ to parse the training set.

20As a POS induction system, the many-to-one performance of the compound PCFG using the preterminals is 68.0. A similarly-parameterized compound HMM with 60 hidden states (an HMM is a particularly type of PCFG) with 60 states obtains 63.2. This is still quite a bit lower than the state-of-the-art (Tran et al., 2016; He et al., 2018; Stratos, 2019), though comparison is confounded by various factors such as preprocessing (e.g. we drop punctuation). A neural PCFG/HMM obtains 68.2 and 63.4 respectively.

16Under our RNNG parameterization, the LSTM LM is equivalent to an RNNG trained with right branching trees.

17Kuncoro et al. (2018, 2019) also observe that models that achieve lower perplexity do not necessarily perform better on syntactic evaluation tasks.

18Li et al. (2019) similarly obtain improvements by refining a model trained on induced trees on classification tasks.
Table 4: For each query sentence (bold), we show the 5 nearest neighbors based on cosine similarity, where we take the representation for each sentence to be the mean of the variational posterior.

We next analyze the continuous latent space. Table 4 shows nearest neighbors of some sentences using the mean of the variational posterior as the continuous representation of each sentence. We qualitatively observe that the latent space seems to capture topical information. We are also interested in the variation in the leaves due to $z$ when the variation due to the tree structure is held constant. To investigate this, we use the parsed dataset to obtain pairs of the form $(\mu_\phi(x^{(n)}), t_j^{(n)})$, where $t_j^{(n)}$ is the $j$-th subtree of the (approximate) MAP tree $t^{(n)}$ for the $n$-th sentence. Therefore each mean vector $\mu_\phi(x^{(n)})$ is associated with $|x^{(n)}| - 1$ subtrees, where $|x^{(n)}|$ is the sentence length. Our definition of subtree here ignores terminals, and thus each subtree is associated with many mean vectors. For a frequently occurring subtree, we perform PCA on the set of mean vectors that are associated with the subtree to obtain the top principal component. We then show the constituents that had the 5 most positive/negative values for this top principal component. The top 5 constituents with the most negative values are shown in the top left part of Table 5. For example, a particularly common subtree—associated with 180 unique constituents—is given by

\[(NT-04 (T-13 w_1)) (NT-12 (NT-20 (NT-20 (NT-07 (T-05 w_2)))) (T-45 w_3)) (T-35 w_4) (T-40 w_5) (T-22 w_6)).\]

The top 5 constituents with the most negative values are shown in the top left part of Table 5. We find that the leaves $[w_1, \ldots, w_6]$, which form a 6-word constituent, vary in a regular manner as $z$ is varied. We also observe that root of this subtree (NT-04) aligns to prepositional phrases (PP) in Figure 2, and the leaves in Table 5 (top left) are indeed mostly PP. However, the model fails to identify $(T-40 w_5) (T-22 w_6)$ as a constituent in this case (as well as well in the bottom right example). See appendix A.5 for more examples. It is possible that the model is utilizing the subtrees to capture broad template-like structures and then using $z$ to fill them in, similar to recent works that also train models to separate “what to say” from “how to say it” (Wiseman et al., 2018; Peng et al., 2019; Chen et al., 2019a,b).

**Limitations** We report on some negative results as well as important limitations of our work. While distributed representations promote parameter sharing, we were unable to improve the above modifications could eventually lead to improvements with the right setup.

Relatedly, the models were quite sensitive to parameterization (e.g. it was important to use residual layers for $f_1$, $f_2$), grammar size, and optimization method. Finally, despite vectorized GPU implementations, training was significantly more expensive (both in terms of time and memory) than NLM-based grammar induction systems due to the $O(|R||x|^3)$ dynamic program, which makes our approach potentially difficult to scale. However, given that we did not exhaustively explore the full space of possible parameterizations, the above modifications could eventually lead to improvements with the right setup.

21 It is also possible to take advantage of pretrained word embeddings by using them to initialize output word embeddings or directly working with continuous emission distributions (Lin et al., 2015; He et al., 2018).
promising results on inducing parse trees directly. Approaches to grammar induction have shown cascaded applications of finite state models. Povtvert et al. (2011) induce parse trees through an incremental algorithm to unsupervised subtrees to parse new sentences. Seginer (2007) devised parsing by using random subsets of training data. Van Cauwenberge (2011) considered a nonparametric-style approach to unsupervised acquisition of a grammar and learning the parameters is achieved through PCFGs, which modulate rule probabilities from generative models. This work explores grammar induction with compound PCFGs, which modulate rule probabilities with per-sentence continuous latent variables. The latent vector induces marginal dependencies beyond the traditional first-order context-free assumptions within a tree-based generative process, leading to improved performance. The collapsed PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006; Cohen et al., 2012), which extend PCFGs to the latent variable setting by splitting nonterminal symbols into latent subsymbols. In particular, latent vector grammars (Zhao et al., 2018) and compositional vector grammars (Socher et al., 2013) also employ continuous vectors within their grammars. CCGs (Erk and Kübler, 2007), which extend PCFGs to the latent variable setting by splitting nonterminal symbols into latent subsymbols, was the basis for much subsequent work. This work explores grammar induction with compound PCFGs, which modulate rule probabilities with per-sentence continuous latent vectors. The latent vector induces marginal dependencies beyond the traditional first-order context-free assumptions within a tree-based generative process, leading to improved performance. The collapsed PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006; Cohen et al., 2012), which extend PCFGs to the latent variable setting by splitting nonterminal symbols into latent subsymbols. In particular, latent vector grammars (Zhao et al., 2018) and compositional vector grammars (Socher et al., 2013) also employ continuous vectors within their grammars. However, these extensions have been employed for learning supervised parsers on annotated treebanks, in contrast to the unsupervised setting of the current work.

### 6 Related Work

Grammar induction has a long and rich history in natural language processing. Early work on grammar induction with pure unsupervised learning was mostly negative (Lari and Young, 1990; Carroll and Charniak, 1992; Charniak, 1993), though Pereira and Schabes (1992) reported some success on partially bracketed data. Clark (2001) and Klein and Manning (2002) reported some of the first successful statistical approaches to grammar induction. In particular, the constituent-context model (CCM) of Klein and Manning (2002), which explicitly models both constituents and dis-PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006; Cohen et al., 2012), which extend PCFGs to the latent variable setting by splitting nonterminal symbols into latent subsymbols. In particular, latent vector grammars (Zhao et al., 2018) and compositional vector grammars (Socher et al., 2013) also employ continuous vectors within their grammars. However, these extensions have been employed for learning supervised parsers on annotated treebanks, in contrast to the unsupervised setting of the current work.

### 7 Conclusion

This work explores grammar induction with compound PCFGs, which modulate rule probabilities with per-sentence continuous latent vectors. The latent vector induces marginal dependencies beyond the traditional first-order context-free assumptions within a tree-based generative process, leading to improved performance. The collapsed PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006; Cohen et al., 2012), which extend PCFGs to the latent variable setting by splitting nonterminal symbols into latent subsymbols. In particular, latent vector grammars (Zhao et al., 2018) and compositional vector grammars (Socher et al., 2013) also employ continuous vectors within their grammars. However, these extensions have been employed for learning supervised parsers on annotated treebanks, in contrast to the unsupervised setting of the current work.
Acknowledgments

We thank Phil Blunsom for initial discussions which seeded many of the core ideas in the present work. We also thank Yonatan Belinkov and Shay Cohen for helpful feedback, and Andrew Drozdov for providing the parsed dataset from their DIORA model. YK is supported by a Google Fellowship. AMR acknowledges the support of NSF 1704834, 1845664, AWS, and Oracle.

References

Alfred Aho. 1968. Indexed Grammars—An Extension of Context-Free Grammars. *Journal of the ACM*, 15(4):647–671.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. 2018. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. In *Proceedings of ICML*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. In *Proceedings of NIPS*.

James K. Baker. 1979. Trainable Grammars for Speech Recognition. In *Proceedings of the Spring Conference of the Acoustical Society of America*.

Joshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of ICML*.

James O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer.

Rens Bod. 2006. An All-Subtrees Approach to Unsupervised Parsing. In *Proceedings of ACL*.

Glenn Carroll and Eugene Charniak. 1992. Two Experiments on Learning Probabilistic Dependency Grammars from Corpora. In *AAAI Workshop on Statistically-Based NLP Techniques*.

Eugene Charniak. 1993. *Statistical Language Learning*. MIT Press.

Danqi Chen and Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of EMNLP*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019a. Controllable Paraphrase Generation with a Syntactic Exemplar. In *Proceedings of ACL*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019b. A Multi-task Approach for Disentangling Syntax and Semantics in Sentence Representations. In *Proceedings of NAACL*.

Alexander Clark. 2001. Unsupervised Induction of Stochastic Context Free Grammars Using Distributional Clustering. In *Proceedings of CoNLL*.

Shay B. Cohen. 2016. *Bayesian Analysis in Natural Language Processing*. Morgan and Claypool.

Shay B. Cohen, Kevin Gimpel, and Noah A Smith. 2009. Logistic Normal Priors for Unsupervised Probabilistic Grammar Induction. In *Proceedings of NIPS*.

Shay B. Cohen and Noah A Smith. 2009. Shared Logistic Normal Distributions for Soft Parameter Tying in Unsupervised Grammar Induction. In *Proceedings of NAACL*.

Shay B. Cohen, Karl Stratos, Michael Collins, Dean P. Foster, and Lyle Ungar. 2012. Spectral Learning of Latent-Variable PCFGs. In *Proceedings of ACL*.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Andrew Drozdov, Patrick Verga, Mohit Yadev, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive Auto-Encoders. In *Proceedings of NAACL*.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. 2019. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *Proceedings of ICLR*.

Greg Durrett and Dan Klein. 2015. Neural CRF Parsing. In *Proceedings of ACL*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent Neural Network Grammars. In *Proceedings of NAACL*.

Dave Golland, John DeNero, and Jakob Uszkoreit. 2012. A Feature-Rich Constituent Context Model for Grammar Induction. In *Proceedings of ACL*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Learning of Syntactic Structure with Invertible Neural Projections. In *Proceedings of EMNLP*.

Phu Mon Htut, Kyunghyun Cho, and Samuel R. Bowman. 2018. Grammar Induction with Neural Language Models: An Unusual Replication. In *Proceedings of EMNLP*.

Yun Huang, Min Zhang, and Chew Lim Tan. 2012. Improved Constituent Context Model with Features. In *Proceedings of PAACL*.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. Unsupervised Grammar Induction with Depth-bounded PCFG. In *Proceedings of TACL*.

Mark Johnson. 1998. PCFG Models of Linguistic Tree Representations. *Computational Linguistics*, 24:613–632.
Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov chain Monte Carlo. In Proceedings of NAACL.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. arXiv:1602.02410.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In Proceedings of AAAI.

Yoon Kim, Alexander M. Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised Recurrent Neural Network Grammars. In Proceedings of NAACL.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In Proceedings of ICLR.

Diederik P. Kingma, Tim Salimans, and Max Welling. 2016. Improving Variational Inference with Autoregressive Flow. arXiv:1606.04934.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In Proceedings of ICLR.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In Proceedings of ACL.

Dan Klein and Christopher Manning. 2002. A Generative Constituent-Context Model forImproved Grammar Induction. In Proceedings of ACL.

Dan Klein and Christopher Manning. 2004. Corpus-based Induction of Syntactic Structure: Models of Dependency and Constituency. In Proceedings of ACL.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In Proceedings of ACL.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better. In Proceedings of ACL.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. 2019. Scalable Syntax-Aware Language Models Using Knowledge Distillation. In Proceedings of ACL.

Kenichi Kurihara and Taisuke Sato. 2006. Variational Bayesian Grammar Induction for Natural Language. In Proceedings of International Colloquium on Grammatical Inference.

Karim Lari and Steve Young. 1990. The Estimation of Stochastic Context-Free Grammars Using the Inside-Outsides Algorithm. Computer Speech and Language, 4:35–56.

Bowen Li, Lili Mou, and Frank Keller. 2019. An Imitation Learning Approach to Unsupervised Parsing. In Proceedings of ACL.

Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. 2007. The Infinite PCFG using Hierarchical Dirichlet Processes. In Proceedings of EMNLP.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 2015. Unsupervised POS Induction with Word Embeddings. In Proceedings of NAACL.

Yang Liu, Matt Gardner, and Mirella Lapata. 2018. Structured Alignment Networks for Matching Sentences. In Proceedings of EMNLP.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19:313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In Proceedings of EMNLP.

Takuya Matsuzaki, Yusuke Miyao, and Junichi Tsujii. 2005. Probabilistic CFG with Latent Annotations. In Proceedings of ACL.

Bernard Merialdo. 1994. Tagging English Text with a Probabilistic Model. Computational Linguistics, 20(2):155–171.

Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using Left-corner Parsing to Encode Universal Structural Constraints in Grammar Induction. In Proceedings of EMNLP.

Ankur P. Parikh, Shay B. Cohen, and Eric P. Xing. 2014. Spectral Unsupervised Parsing with Additive TreeMetrics. In Proceedings of ACL.

Hao Peng, Ankur P. Parikh, Manaal Faruqui, Bhuvan Dhingra, and Dipanjan Das. 2019. Text Generation withExemplar-based Adaptive Decoding. In Proceedings of NAACL.

Fernando Pereira and Yves Schabes. 1992. Inside-Outside Reestimation from Partially Bracketed Corpora. In Proceedings of ACL.

Slav Petrov, Leon Barret, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In Proceedings of ACL.

Elis Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simpled Unsupervised Grammar Induction from Raw Text with Cascaded Finite State Methods. In Proceedings of ACL.

Ofir Press and Lior Wolf. 2016. Using the Output Embedding to Improve Language Models. In Proceedings of EACL.

Roi Reichart and Ari Rappoport. 2010. Improved Fully Unsupervised Parsing with Zoomed Learning. In Proceedings of EMNLP.
Danilo J. Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In Proceedings of ICML.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of ICML.

Herbert Robbins. 1951. Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pages 131–149. Berkeley: University of California Press.

Herbert Robbins. 1956. An Empirical Bayes Approach to Statistics. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, pages 157–163. Berkeley: University of California Press.

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning Character-level Representations for Part-of-Speech Tagging. In Proceedings of ICML.

Yoav Seginer. 2007. Fast Unsupervised Incremental Parsing. In Proceedings ofACL.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In Proceedings of ICLR.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks. In Proceedings of ICLR.

Noah A. Smith and Jason Eisner. 2004. Annealing Techniques for Unsupervised Statistical Language Learning. In Proceedings ofACL.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised Multilingual Grammar Induction. In Proceedings ofACL.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with Compositional Vector Grammars. In Proceedings of ACL.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2012. Three Dependency-and-Boundary Models for Grammar Induction. In Proceedings ofEMNLP-CoNLL.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2013. Breaking Out of Local Optima with Count Transforms and Model Recombination: A Study in Grammar Induction. In Proceedings ofEMNLP.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A Minimal Span-Based Neural Constituency Parser. In Proceedings of ACL.

Karl Stratos. 2019. Mutual Information Maximization for Simple and Accurate Part-of-Speech Induction. In Proceedings ofNAACL.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In Proceedings ofACL.

Ke Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. 2016. Unsupervised Neural Hidden Markov Models. In Proceedings of the Workshop on Structured Prediction for NLP.

Pengyu Wang and Phil Blunsom. 2013. Collapsed Variational Bayesian Inference for PCFGs. In Proceedings ofCoNLL.

Wenhui Wang and Baobao Chang. 2016. Graph-based Dependency Parsing with Bidirectional LSTM. In Proceedings ofACL.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. In Proceedings ofNAACL.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning Neural Templates for Text Generation. In Proceedings ofEMNLP.

Ji Xu, Daniel Hsu, and Arian Maleki. 2018. Benefits of Over-Parameterization with EM. In Proceedings ofNeurIPS.

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, 11:207–238.

Cun-Hui Zhang. 2003. Compound Decision Theory and Empirical Bayes Methods. The Annals of Statistics, 31:379–390.

Yanpeng Zhao, Liwen Zhang, and Kewei Tu. 2018. Gaussian Mixture Latent Vector Grammars. In Proceedings ofACL.

Xiaodan Zhu, Parinaz Sobhani, and Hongyu Guo. 2015. Long Short-Term Memory Over Tree Structures. In Proceedings ofICML.
A Appendix

A.1 Model Parameterization

Neural PCFG We associate an input embedding \( w_N \) for each symbol \( N \) on the left side of a rule (i.e. \( N \in \{S\} \cup \mathcal{N} \cup \mathcal{P} \)) and run a neural network over \( w_N \) to obtain the rule probabilities. Concretely, each rule type \( \pi_r \) is parameterized as follows,

\[
\begin{align*}
\pi_{S \rightarrow A} &= \frac{\exp\left(u_A^T f_1(w_S)\right)}{\sum_{A' \in \mathcal{N}} \exp\left(u_{A'}^T f_1(w_S)\right)}, \\
\pi_{A \rightarrow BC} &= \frac{\exp(u_{BC}^T w_A)}{\sum_{B'C' \in \mathcal{M}} \exp(u_{B'C'}^T w_{A'})}, \\
\pi_{T \rightarrow w} &= \frac{\exp(u_{w1}^T f_2(w_T))}{\sum_{w' \in \Sigma} \exp(u_{w'}^T f_2(w_T))},
\end{align*}
\]

where \( \mathcal{M} \) is the product space \((\mathcal{N} \cup \mathcal{P}) \times (\mathcal{N} \cup \mathcal{P})\), and \( f_1, f_2 \) are MLPs with two residual layers,

\[
\begin{align*}
f_i(x) &= g_{i,1}(g_{i,2}(W_i x)), \\
g_{i,j}(y) &= \text{ReLU}(V_{i,j} \text{ReLU}(U_{i,j} y) + y).
\end{align*}
\]

The bias terms for the above expressions (including for the rule probabilities) are omitted for notational brevity. In Figure 1 we use the following to refer to rule probabilities of different rule types,

\[
\begin{align*}
\pi_S &= \{ \pi_r \mid r \in L(S) \}, \\
\pi_N &= \{ \pi_r \mid r \in L(A), A \in \mathcal{N} \}, \\
\pi_P &= \{ \pi_r \mid r \in L(T), T \in \mathcal{P} \}, \\
\pi &= \pi_S \cup \pi_N \cup \pi_P,
\end{align*}
\]

where \( L(A) \) denotes the set of rules with \( A \) on the left hand side.

Compound PCFG The compound PCFG rule probabilities \( \pi_z \) given a latent vector \( z \),

\[
\begin{align*}
\pi_{z,S \rightarrow A} &= \frac{\exp(u_A^T f_1([w_S; z]))}{\sum_{A' \in \mathcal{N}} \exp(u_{A'}^T f_1([w_S; z]))}, \\
\pi_{z,A \rightarrow BC} &= \frac{\exp(u_{BC}^T [w_A; z])}{\sum_{B'C' \in \mathcal{M}} \exp(u_{B'C'}^T [w_{A'}; z])}, \\
\pi_{z,T \rightarrow w} &= \frac{\exp(u_{w1}^T f_2([w_T; z]))}{\sum_{w' \in \Sigma} \exp(u_{w'}^T f_2([w_T; z]))}.
\end{align*}
\]

Again the bias terms are omitted for brevity, and \( f_1, f_2 \) are as before where the first layer’s input dimensions are appropriately changed to account for concatenation with \( z \).

A.2 Corpus/Sentence \( F_1 \) by Sentence Length

For completeness we show the corpus-level and sentence-level \( F_1 \) broken down by sentence length in Table 6, averaged across 4 different runs of each model.

A.3 Experiments with RNNGs

For experiments on supervising RNNGs with induced trees, we use the parameterization and hyperparameters from Kim et al. (2019), which uses a 2-layer 650-dimensional stack LSTM (with dropout of 0.5) and a 650-dimensional tree LSTM (Tai et al., 2015; Zhu et al., 2015) as the composition function.

Concretely, the generative story is as follows: first, the stack representation is used to predict the next action (\textsc{shift} or \textsc{reduce}) via an affine transformation followed by a sigmoid. If \textsc{shift} is chosen, we obtain a distribution over the vocabulary via another affine transformation over the stack representation followed by a softmax. Then we sample the next word from this distribution and shift the generated word onto the stack using the stack LSTM. If \textsc{reduce} is chosen, we pop the last two elements off the stack and use the tree LSTM to obtain a new representation. This new representation is shifted onto the stack via the stack LSTM. Note that this RNNG parameterization is slightly different than the original from Dyer et al. (2016), which does not ignore constituent labels and utilizes a bidirectional LSTM as the composition function instead of a tree LSTM. As our RNNG parameterization only works with binary trees, we binarize the gold trees with right binarization for the RNNG trained on gold trees (trees from the unsupervised methods explored in this paper are already binary). The RNNG also trains a discriminative parser alongside the generative model for evaluation with importance sampling. We use a CRF parser whose span score parameterization is similar similar to recent works (Wang and Chang, 2016; Stern et al., 2017; Kitaev and Klein, 2018): position embeddings are added to word embeddings, and a bidirectional LSTM with 256 hidden dimensions is run over the input representations to obtain the forward and backward hidden states. The score \( s_{ij} \in \mathbb{R} \) for a constituent spanning the \( i \)-th and \( j \)-th word is given by,

\[
s_{ij} = \text{MLP}([\overrightarrow{h}_{j+1} - \overrightarrow{h}_i; \overrightarrow{h}_{i-1} - \overrightarrow{h}_j]),
\]

where the MLP has a single hidden layer with
Table 6: Average unlabeled $F_1$ for the various models broken down by sentence length on the PTB test set. For example WSJ-10 refers to $F_1$ calculated on the subset of the test set where the maximum sentence length is at most 10. Scores are averaged across 4 runs of the model with different random seeds. Oracle is the performance of binarized gold trees (with right branching binarization). Top shows sentence-level $F_1$ and bottom shows corpus-level $F_1$.

|               | Left Branching | Right Branching | Random Trees | PRPN (tuned) | ON (tuned) | Neural PCFG | Compound PCFG | Oracle |
|---------------|----------------|-----------------|--------------|--------------|------------|--------------|---------------|--------|
| WSJ-10        | 17.4           | 58.5            | 31.8         | 58.4         | 63.9       | 64.6         | 70.5          | 82.1   |
| WSJ-20        | 12.9           | 49.8            | 25.2         | 54.3         | 57.5       | 58.1         | 63.4          | 84.1   |
| WSJ-30        | 9.9            | 44.4            | 21.5         | 50.9         | 53.2       | 54.6         | 58.9          | 84.2   |
| WSJ-40        | 8.6            | 41.6            | 19.7         | 48.5         | 50.5       | 52.6         | 56.6          | 84.3   |
| WSJ-Full      | 8.7            | 39.5            | 19.2         | 47.3         | 48.1       | 50.8         | 55.2          | 84.3   |

ReLU nonlinearity followed by layer normalization (Ba et al., 2016).

For experiments on fine-tuning the RNNG with the unsupervised RNNG, we take the discriminative parser (which is also pretrained alongside the RNNG on induced trees) to be the structured inference network for optimizing the evidence lower bound. We refer the reader to Kim et al. (2019) and their open source implementation\textsuperscript{22} for additional details. We also observe that as noted by Kim et al. (2019), a URNNG trained from scratch on this version of PTB without punctuation failed to outperform a right-branching baseline.

The LSTM language model baseline is the same size as the stack LSTM (i.e. 2 layers, 650 hidden units, dropout of 0.5), and is therefore equivalent to an RNNG with completely right branching trees. For all models we share input/output word embeddings (Press and Wolf, 2016). Perplexity estimation for the RNNGs and the compound PCFG uses 1000 importance-weighted samples.

For grammaticality judgment, we modify the publicly available dataset from Marvin and Linzen (2018)\textsuperscript{23} to only keep sentence pairs that did not have any unknown words with respect to our PTB vocabulary of 10K words. This results in 33K sentence pairs for evaluation.

\textsuperscript{22}https://github.com/harvardnlp/urnng

\textsuperscript{23}https://github.com/BeckyMarvin/LM_syneval

A.4 Nonterminal/Preterminal Alignments

Figure 3 shows the part-of-speech alignments and Table 7 shows the nonterminal label alignments for the compound PCFG/neural PCFG.

A.5 Subtree Analysis

Table 8 lists more examples of constituents within each subtree as the top principal component is varied. Due to data sparsity, the subtree analysis is performed on the full dataset. See section 5 for more details.
Figure 3: Preterminal alignment to part-of-speech tags for the compound PCFG (top) and the neural PCFG (bottom).
Table 7: Analysis of label alignment for nonterminals in the compound PCFG (top) and the neural PCFG (bottom). Label alignment is the proportion of correctly-predicted constituents that correspond to a particular gold label. We also show the predicted constituent frequency and accuracy (i.e. precision) on the right. Bottom line shows the frequency in the gold trees.
Table 8: For each subtree (shown at the top of each set of examples), we perform PCA on the variational posterior mean vectors that are associated with that particular subtree and take the top principal component. We then list the top 5 constituents that had the lowest (left) and highest (right) principal component values.