Differentiable Perturb-and-Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder

Caio Corro **Ivan Titov**
ILCC, School of Informatics, University of Edinburgh
ILLC, University of Amsterdam
c.f.corro@uva.nl ititov@inf.ed.ac.uk

Abstract

Human annotation for syntactic parsing is expensive, and large resources are available only for a fraction of languages. A question we ask is whether one can leverage abundant unlabeled texts to improve syntactic parsers, beyond just using the texts to obtain more generalisable lexical features (i.e. beyond word embeddings). To this end, we propose a novel latent-variable generative model for semi-supervised syntactic dependency parsing. As exact inference is intractable, we introduce a differentiable relaxation to obtain approximate samples and compute gradients with respect to the parser parameters. Our method (Differentiable Perturb-and-Parse) relies on differentiable dynamic programming over stochastically perturbed edge scores. We demonstrate effectiveness of our approach with experiments on English, French and Swedish.

1 Introduction

A dependency tree is a lightweight syntactic structure exposing (possibly labeled) bi-lexical relations between words (Tesnière, 1959; Kaplan and Bresnan, 1982), see Figure 1. This representation has been widely studied by the NLP community leading to very efficient state-of-the-art parsers (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017; Ma and Hovy, 2017), motivated by the fact that dependency trees are useful in downstream tasks such as semantic parsing (Reddy et al., 2016; Marcheggiani and Titov, 2017), machine translation (Ding and Palmer, 2005; Bastings et al., 2017), information extraction (Culotta and Sorensen, 2004; Liu et al., 2015), question answering (Cui et al., 2005) and even as a filtering method for constituency parsing (Kong et al., 2015), among others.

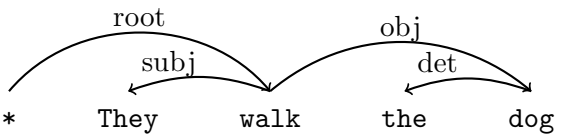


Figure 1: Dependency tree example: each arc represents a labeled relation between the head word (the source of the arc) and the modifier word (the destination of the arc). The first token is a fake root word.

Unfortunately, syntactic annotation is a tedious and expensive task, requiring highly-skilled human annotators. Consequently, even though syntactic annotation is now available for many languages, the datasets are often small. For example, 31 languages in the Universal Dependency Treebank[1], the largest dependency annotation resource, have fewer than 5,000 sentences, including such major languages as Vietnamese and Telugu. This makes the idea of using unlabeled texts as an additional source of supervision especially attractive.

In previous work, before the rise of deep learning, the semi-supervised parsing setting has been mainly tackled with two-step algorithms.

[1] http://universaldependencies.org/

On the one hand, feature extraction methods first learn an intermediate representation using an unlabeled dataset which is then used as input to train a supervised parser (Koo et al., 2008; Yu et al., 2008; Chen et al., 2009; Suzuki et al., 2011). On the other hand, the self-training and co-training methods start by learning a supervised parser that is then used to label extra data. Then, the parser is retrained with this additional annotation (Sagae and Tsujii, 2007; Kawahara and Uchimoto, 2008; McClosky et al., 2006). Nowadays, unsupervised feature extraction is achieved in neural parsers by the means of word embeddings (Mikolov et al., 2013). The natural question to ask is whether one can exploit unlabeled data in neural parsers beyond only inducing generalizable word representations.

Our method can be regarded as semi-supervised Variational Auto-Encoder, VAE (Kingma et al., 2014). Specifically, we introduce a probabilistic model (Section 3) parametrized with a neural network (Section 4). The model assumes that a sentence is generated conditioned on a latent dependency tree. Dependency parsing corresponds to approximating the posterior distribution over the latent trees within this model, achieved by the encoder component of VAE, see Figure 2a. The parameters of the generative model and the parser (i.e. the encoder) are estimated by maximizing the likelihood of unlabeled sentences. In order to ensure that the latent representation is consistent with treebank annotation, we combine the above objective with maximizing the likelihood of gold parse trees in the labeled data.

Training a VAE via backpropagation requires marginalization over the latent variables, which intractable for dependency trees. In this case, previous work proposed approximate training methods, mainly differentiable Monte-Carlo estimation (Kingma and Welling, 2013; Rezende et al., 2014) and score function estimation, e.g. REINFORCE (Williams, 1992). However, REINFORCE is known to suffer from high variance (Mnih and Gregor, 2014). Therefore, we propose an approximate differentiable Monte-Carlo approach that we call Differentiable Perturb-and-Parse (Section 5). The key idea is that we can obtain a differentiable relaxation of an approximate sample by (1) perturbing scores of candidate dependency edges and (2) performing structured argmax inference with differentiable dynamic programming, relying on the perturbed scores. In this way we bring together ideas of perturb-and-map inference (Papandreou and Yuille, 2011; Maddison et al., 2017) and continuous relaxation for dynamic programming (Mensch and Blondel, 2018). We evaluate our semi-supervised parser on English, French and Swedish and show improvement over a comparable supervised baseline (Section 6).

Our main contributions can be summarized as follows:

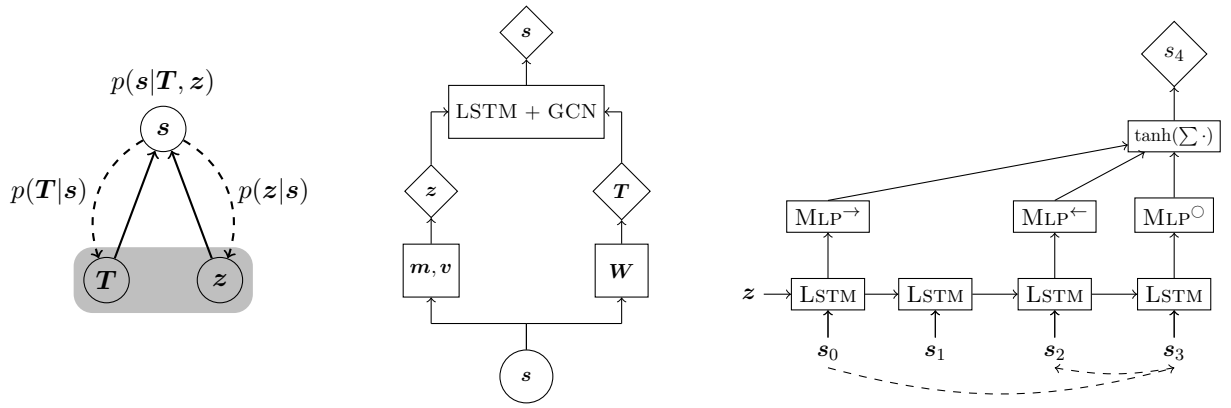
- we introduce a variational autoencoder for semi-supervised dependency parsing;
- we propose the Differentiable Perturb-and-Parse method for its estimation;
- we demonstrate the effectiveness of the approach on three different languages.

Notation An upper case letter denotes an unordered set (e.g. S). A bold lower case letters (e.g. $\boldsymbol{s}$) denotes a (column) vector and a subscripted lower case letter (e.g. s_i) its value at row i. Similarly, a bold upper case letter (e.g. $\boldsymbol{T}$) stands for a matrix and a subscripted lower case letter (e.g. $t_{i,j}$) for its value at row i and column j. Greek letters ϕ and θ are used for parameters of probability distributions, e.g. the parameters of the neural network that computes the mean and variance of a Gaussian distribution. Greek letter τ denotes the scalar temperature hyperparameter.

2 Dependency parsing

A dependency is a bi-lexical relation between a head word (the source) and a modifier word (the target), see Figure 1. The set of dependencies of a sentence defines a tree-shaped structure.[2] In the parsing problem, we aim to compute the dependency tree of a given sentence.

[2] Semantic dependencies can have a more complex structure, e.g. words with several heads. However, we focus on syntactic dependencies only.



(a) Probabilistic model (b) Computation Graph (c) Decoder

Figure 2: **(a)** Illustration of our probabilistic model with random variables $\boldsymbol{s}$, $\boldsymbol{T}$ and $\boldsymbol{z}$ for sentences, dependency trees and sentence embeddings, respectively. The gray area delimits the latent space. Solid arcs denote the generative process, dashed arcs denotes posterior distributions over the latent variables. **(b)** Stochastic computation graph. **(c)** Illustration of the decoder when computing the probability distribution of s_4, the word at position 4. Dashed arcs at the bottom represent syntactic dependencies between word at position 4 and previous positions. At each step, the LSTM takes as input an embedding of the previous word (s_0 is a special start-of-sentence symbol). Then, the GCN combines different outputs of the LSTM by transforming them with respect to their syntactic relation with the current position. Finally, the probability of s_4 is computed via the softmax function.

Formally, we define a sentence as a sequence of tokens (words) from vocabulary W. We assume a one-to-one mapping between W and integers $1 \dots |W|$. Therefore, we write a sentence of length $n+1$ as a vector of integers $\boldsymbol{s} = [s_0, s_1, \dots, s_n]^\top$, with $1 \leq s_i \leq |W|$ and where s_0 is a special root symbol. A dependency tree of sentence $\boldsymbol{s}$ is a matrix of booleans $\boldsymbol{T} \in \{0,1\}^{(n+1)\times(n+1)}$ with $t_{h,m} = 1$ meaning that word s_h is the head of word s_m in the dependency tree.

More specifically, a dependency tree $\boldsymbol{T}$ is the adjacency matrix of a directed graph with $n+1$ vertices $v_0 \dots v_n$. A matrix $\boldsymbol{T}$ is a valid dependency tree if and only if this graph is a v_0-rooted spanning arborescence,[3] i.e. the graph is connected, each vertex has at most one incoming arc and the only vertex without incoming arc is v_0. A dependency tree is projective if and only if, for each arc $v_h \to v_m$, if $h < m$ (resp. $m < h$) then there exists a path with arcs T from v_h to each vertex v_k such that $h < k < m$ (resp. $m < k < h$). Intuitively, this means that we can draw the dependency tree above the sentence without crossing arcs.

Given a sentence $\boldsymbol{s}$, an arc-factored dependency parser computes the dependency tree $\boldsymbol{T}$ which maximizes a weighting function $f(\boldsymbol{T}; \boldsymbol{W}) = \sum_{h,m} t_{h,m} w_{h,m}$, where $\boldsymbol{W}$ is a matrix of dependency (arc) weights. This problem can be solved with a $\mathcal{O}(n^2)$ time complexity (Tarjan, 1977; McDonald et al., 2005). If we restrict $\boldsymbol{T}$ to be a projective dependency tree, then the optimal solution can be computed with a $\mathcal{O}(n^3)$ time complexity (Eisner, 1996). Restricting the search space to projective trees is appealing for treebanks exhibiting this property (either exactly or approximately): they enforce a structural constraint that can be beneficial for accuracy, especially in a low-resource scenario. Moreover, using a more restricted search space of potential trees may be

[3] Tree refers to the linguistic structure whereas arborescence refers to the graph structure.

especially beneficial in a semi-supervised scenario: with a more restricted space a model is less likely to diverge from a treebank grammar and capture non-syntactic phenomena. Finally, Eisner's (1996) algorithm can be described as a deduction system (Pereira and Warren, 1983), a framework that unifies many parsing algorithms. As such, our methodology could be applied to other grammar formalisms. For all these reasons, in this paper, we focus on projective dependency trees only.

3 Generative model

In the previous section, we explained how the projective dependency tree of maximum weight can be computed given the matrix $\boldsymbol{W}$. We now turn to the learning problem, i.e. estimation of the matrix $\boldsymbol{W}$.

We assume that we have access to a set of i.i.d. labeled sentences $\bar{S} = \{\langle \boldsymbol{s}, \boldsymbol{T} \rangle, \dots\}$ and a set of i.i.d. unlabeled sentences $\tilde{S} = \{\boldsymbol{s}, \dots\}$. In order to incorporate unlabeled data in the learning process, we introduce a generative model where the dependency tree is latent (Subsection 3.1). As such, we can maximize the likelihood of observed sentences even if the ground-truth dependency tree is unknown. We learn the parameters of this model using a variational Bayes approximation (Subsection 3.2) augmented with a discriminative objective on labeled data (Subsection 3.3).

3.1 Generative story

Under our probabilistic model, a sentence $\boldsymbol{s}$ is generated from a continuous sentence embedding $\boldsymbol{z}$ and with respect to a syntactic structure $\boldsymbol{T}$. We formally define the generative process of a sentence of length n as:

$$\boldsymbol{T} \sim p(\boldsymbol{T}|n) \quad \boldsymbol{z} \sim p(\boldsymbol{z}|n) \quad \boldsymbol{s} \sim p(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z}, n)$$

This Bayesian network is shown in Figure 2a. In order to simplify notation, we omit conditioning on n in the following. $\boldsymbol{T}$ and $\boldsymbol{z}$ are latent variables and $p(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})$ is the conditional likelihood of observations. The true distribution underlying the observed data is unknown, so we have to learn a model $p_\theta(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})$ parametrized by θ that best fits the given samples:

$$\theta = \arg\max_\theta \sum_{\boldsymbol{s}} \log p_\theta(\boldsymbol{s}) \tag{1}$$

Then, the posterior distribution of latent variables $p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$ models the probability of underlying representations (including dependency trees) with respect to a sentence. This conditional distribution can be written as:

$$p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s}) = \frac{p_\theta(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z}) p(\boldsymbol{T}) p(\boldsymbol{z})}{p_\theta(\boldsymbol{s})} \tag{2}$$

In the next subsection, we explain how these two quantities can be estimated from data.

3.2 Variational Inference and Variational Auto-Encoders

Computations in Equation 1 and Equation 2 require marginalization over the latent variables:

$$p_\theta(\boldsymbol{s}) = \sum_{\boldsymbol{T}} \int p_\theta(\boldsymbol{s}, \boldsymbol{T}, \boldsymbol{z})\, d\boldsymbol{z}$$

which is intractable in general. Variational Inference, VI (Jordan et al., 1999; Wainwright et al., 2008) tackles this problem by introducing a variational distribution $q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$ which is intended to be similar to $p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$. More formally, we want $\mathrm{KL}\left[q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s}) \| p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})\right]$ to be as small as possible, where KL is the Kulback-Leibler (KL) divergence. Then, the following equality holds:

$$\begin{aligned}
&\log p(\boldsymbol{s}) - \mathrm{KL}\left[q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s}) \| p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})\right] \\
&= \mathbb{E}_{q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})}[\log(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})] - \mathrm{KL}[q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})|p(\boldsymbol{T}, \boldsymbol{z})]
\end{aligned}$$

The KL divergence is always positive, therefore:

$$\begin{aligned}
\log p(\boldsymbol{s}) \geq\, & \mathbb{E}_{q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})}[\log(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})] \\
& - \mathrm{KL}[q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})|p(\boldsymbol{T}, \boldsymbol{z})] \\
=\, & \tilde{\mathcal{E}}_{\theta,q}(\boldsymbol{s})
\end{aligned} \tag{3}$$

where the right-hand side is called the Evidence Lower Bound (ELBO). By maximizing the ELBO term, the divergence $\mathrm{KL}\left[q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s}) \| p_\theta(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})\right]$ is implicitly minimized. Therefore, we define a surrogate objective, replacing the objective in Equation 1:

$$\theta = \arg\max_\theta \sum_{\boldsymbol{s}} \max_q \tilde{\mathcal{E}}_{\theta,q}(\boldsymbol{s}) \tag{4}$$

Earlier VI approaches assumed that variational distributions are computed for each data point individually, for example, using some form of message passing (Jordan et al., 1999). Nowadays, VI is typically amortized (Kingma and Welling, 2013; Rezende et al., 2014). The distribution $q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$ is computed by a neural network parameterized by ϕ: $q_\phi(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$. This implies that the maximization problem in Equation 4 is further replaced with

$$\theta, \phi = \arg\max_{\theta,\phi} \sum_{\boldsymbol{s}} \tilde{\mathcal{E}}_{\theta,\phi}(\boldsymbol{s}), \tag{5}$$

where $\tilde{\mathcal{E}}_{\theta,\phi}(\boldsymbol{s})$ stands for $\tilde{\mathcal{E}}_{\theta,q_\phi}(\boldsymbol{s})$.

The ELBO contains the non-trivial term $\mathbb{E}_{q(\boldsymbol{T},\boldsymbol{z}|\boldsymbol{s})}[\log p(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})]$. During training, Monte-Carlo method provides a tractable and unbiased estimation of the expectation. Note that a single sample from $q(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s})$ can be understood as encoding the observation into the latent space, whereas sampling a sentence from the latent space can be understood as decoding. Therefore, this approach is typically referred to as Variational Auto-Encoding. However, training a VAE requires the sampling process to be differentiable. In the case of the sentence embedding, we follow the usual setting and define $q(\boldsymbol{z}|\boldsymbol{s})$ as a diagonal Gaussian: backpropagation through the the sampling process $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{S})$ can be achieved through the reparametrization trick (Kingma and Welling, 2013; Rezende et al., 2014). Unfortunately, this approach cannot be applied to dependency tree sampling $\boldsymbol{T} \sim q(\boldsymbol{T}|\boldsymbol{s})$. We tackle this issue in Section 5.

3.3 Semi-supervised learning

VAEs are a convenient approach for semi-supervised learning (Kingma et al., 2014) and have been successfully applied in NLP (Kočiský et al., 2016; Xu et al., 2017; Zhou and Neubig, 2017; Yin et al., 2018). In this scenario, we are given the dependency structure of a subset of the observations, i.e. $\boldsymbol{T}$ is an observed variable. Then, the supervised ELBO term is defined as:

$$\begin{aligned}\tilde{\mathcal{E}}_{\theta,\phi}(\boldsymbol{s}) = &\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s})}[\log(\boldsymbol{s}|\boldsymbol{T}, \boldsymbol{z})] \\ &- \mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{s})|p(\boldsymbol{z})]\end{aligned} \tag{6}$$

Note that our end goal is to estimate the posterior ditribution over dependency trees $q_\phi(\boldsymbol{T}|\boldsymbol{s})$, i.e. the dependency parser, which does not appear in the supervised ELBO. We want to explicitly use the labeled data in order to learn the parameters of this parser. This can be achieved by adding a discriminative training term to the overall loss.[4]

The loss function for training a semi-supervised VAE is:[5]

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(\bar{S}, \tilde{S}) = &- \sum_{\boldsymbol{s},\boldsymbol{T}\in\bar{S}} \log q_\phi(\boldsymbol{T}|\boldsymbol{s}) \\ &- \sum_{\boldsymbol{s},\boldsymbol{T}\in\bar{S}} \bar{\mathcal{E}}_{\theta,\phi}(\boldsymbol{s}, \boldsymbol{T}) - \sum_{\boldsymbol{s}\in\tilde{S}} \tilde{\mathcal{E}}_{\theta,\phi}(\boldsymbol{s})\end{aligned} \tag{7}$$

where the first term is the standard loss for supervised learning of log-linear models (Johnson et al., 1999; Lafferty et al., 2001). Due to numerical instability, this term does not take into account the structure.[6]

4 Neural Parametrization

In this section, we describe the neural parametrization of the encoder distribution q_ϕ (Subsection 4.1) and the decoder distribution p_θ (Subsection 4.2). A visual representation is given on Figure 2b.

4.1 Encoder

We factorize the encoder as $q_\phi(\boldsymbol{T}, \boldsymbol{z}|\boldsymbol{s}) = q_\phi(\boldsymbol{T}|\boldsymbol{s})q_\phi(\boldsymbol{z}|\boldsymbol{s})$. The categorical distribution over dependency trees is parametrized by a log-linear model (Lafferty et al., 2001) where the weight of an arc is given by the neural network of Kiperwasser and Goldberg (2016):[7]

$$\begin{aligned}\boldsymbol{W} &= \textsc{DepWeights}(\boldsymbol{s}) \\ q_\phi(\boldsymbol{T}|\boldsymbol{s}) &= \frac{\exp(\sum_{i,j} w_{i,j} t_{i,j})}{\sum_{\boldsymbol{T}'} \exp(\sum_{i,j} w_{i,j} t'_{i,j})}\end{aligned}$$

[4] This term can be equivalently regarded as a form of data-dependent prior on the posterior distribution, see Section 3.1.2 in Kingma et al. (2014).

[5] Remember that we want to maximize the ELBO terms, hence the negated terms.

[6] That is the discriminative term considers heads independently.

[7] We reimplemented the exact same architecture as in the code distributed with the paper.

The sentence embedding model is specified as a diagonal Gaussian parametrized by a LSTM, similarly to the seq2seq framework (Sutskever et al., 2014; Bowman et al., 2016):

$$\begin{aligned} \boldsymbol{m}, \log \boldsymbol{v}^2 &= \textsc{EmbParams}(\boldsymbol{s}) \\ q_\phi(\boldsymbol{z}|\boldsymbol{s}) &= \mathcal{N}(\boldsymbol{z}|\boldsymbol{m}, \boldsymbol{v}) \end{aligned}$$

Where $\boldsymbol{m}$ and $\boldsymbol{v}$ are mean and variance vectors, respectively.[8]

4.2 Decoder

We use an autoregressive decoder that combines a LSTM, in order to take into account the history of generated words, with a Graph Convolutional Network, GCN (Kipf and Welling, 2016; Marcheggiani and Titov, 2017), in order to take into account syntactic dependencies.

The hidden state of the LSTM is initialized with latent variable $\boldsymbol{z}$ (the sentence embedding). Then, at each step $1 \leq i \leq n$, an embedding associated with word at position $i-1$ is fed as input. A special start-of-sentence symbol embedding is used at the first position.

Let $\boldsymbol{o}^i$ be the hidden state of the LSTM at position i. The standard seq2seq architecture uses this vector to predict the word at position i. Instead, we transform it in order to take into account the syntactic structure described by the latent variable $\boldsymbol{T}$. Due to the autoregressive nature of the decoder, we can only take into account dependencies $t_{h,m}$ such that $h < i$ and $m < i$. Before being fed to the GCN, the output of the LSTM is fed to distinct multi-layer perceptrons[9] that characterize syntactic relations: if s_h is the head of s_i, $\boldsymbol{o}^h$ is transformed with $\textsc{Mlp}^{\curvearrowright}$, if s_m is a modifier of s_i, $\boldsymbol{o}^m$ is transformed with $\textsc{Mlp}^{\curvearrowleft}$, and lastly o^i is transformed with $\textsc{Mlp}^{\circlearrowleft}$. Formally, the GCN is defined as follows:

$$\begin{aligned} \boldsymbol{g}^i = \tanh \Bigg(& \textsc{Mlp}^{\circlearrowleft}(\boldsymbol{o}^i) \\ & + \sum_{h=0}^{i-1} t_{h,i} \times \textsc{Mlp}^{\curvearrowright}(\boldsymbol{o}^h) \\ & + \sum_{m=0}^{i-1} t_{i,m} \times \textsc{Mlp}^{\curvearrowleft}(\boldsymbol{o}^m) \Bigg) \end{aligned}$$

The output vector $\boldsymbol{g}^i$ is then used to estimate the probability of word s_i. The neural architecture of the decoder is illustrated on Figure 2c.

5 Differentiable Perturb-and-Parse

Encoder-decoder architectures are usually straightforward to optimize with the backpropagation algorithm (Linnainmaa, 1976; LeCun et al., 2012) using any autodiff library. Unfortunately, our VAE contains stochastic nodes that can not be differentiated efficiently as marginalization is too expensive or intractable (see Figure 2b for the list of stochastic nodes in our computation graph). Kingma and Welling (2013) and Rezende et al. (2014) proposed to rely on a Monte-Carlo estimation of the gradient. This approximation is differentiable because the sampling process is moved out of the backpropagation path. For example, sampling from our diagonal Gaussian random variable can be re-expressed as:

$$\begin{aligned} \boldsymbol{m}, \log \boldsymbol{v}^2 &= \textsc{EmbParams}(\boldsymbol{s}) \\ \boldsymbol{e} &\sim \mathcal{N}(0, 1) \\ \boldsymbol{z} &= \boldsymbol{m} + \boldsymbol{v} \times \boldsymbol{e} \end{aligned} \tag{8}$$

As such, $\boldsymbol{e} \sim \mathcal{N}(0, 1)$ is an input of the neural network for which we do not need to compute partial derivatives. This technique is called the reparametrization trick.

In this section, we introduce our Differentiable Perturb-and-Parse operator to cope with the distribution over dependency trees. Firstly, in Subsection 5.1, we propose an approximate sampling process by computing the best parse tree with respect to independently perturbed arc weights. Secondly, we propose a differentiable surrogate of the parsing algorithm in Subsection 5.2.

[8] The covariance matrix can be reduced to a vector as we restrict it to be diagonal.

[9] Distinct means that $\textsc{Mlp}^{\curvearrowright}$, $\textsc{Mlp}^{\curvearrowleft}$ and $\textsc{Mlp}^{\circlearrowleft}$ have different parameters.

5.1 Perturb-and-Parse

Sampling from a categorical distributions can be achieved through the Gumbel-Max trick (Gumbel, 1954; Maddison et al., 2014). Randomly generated Gumbel noise is added to the log-probability of every element of the sample space. Then, the sample is simply the element with maximum *perturbed* log-probability. Let $\boldsymbol{d} \in \triangle^k$ be a random variable taking values in the corner of the unit-simplex of dimension k with probability:

$$p(\boldsymbol{d} \in \triangle^k) = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{d})}{\sum_{\boldsymbol{d}' \in \triangle^k} \exp(\boldsymbol{w}^\top \boldsymbol{d}')}$$

where $\boldsymbol{w}$ is a vector of weights. Sampling $\boldsymbol{d} \sim p(\boldsymbol{d})$ can be re-expressed as follows:

$$\begin{aligned} \boldsymbol{g} &\sim \mathcal{G}(0,1) \\ \boldsymbol{d} &= \arg\max_{\boldsymbol{d} \in \triangle^k} (\boldsymbol{w} + \boldsymbol{g})^\top \boldsymbol{d} \end{aligned}$$

where $\mathcal{G}(0,1)$ is the Gumbel distribution. Sampling $\boldsymbol{g} \sim \mathcal{G}(0,1)$ is equivalent to setting $g_i = -\log(-\log u_i))$ where $u_i \sim \text{Uniform}(0,1)$. The sampling process is outside the backpropagation, similar to the reparametrization in Equation 8.

Unfortunately, this reparametrization is difficult to apply when the discrete variable can take an exponential number of values as in Markov Random Fields (MRF). Papandreou and Yuille (2011) proposed an approximate sampling process: each components is perturbed independently. Then, standard MAP inference algorithm computes the sample. This technique is called perturb-and-map.

Arc-factored dependency parsing can be express as a MRF where variable nodes represent arcs, singleton factors weight arcs and a fully connected factor forces the variable assignation to describe a valid dependency tree (Smith and Eisner, 2008). Therefore, we can apply the perturb-and-map method to dependency tree sampling:[10]

$$\begin{aligned} \boldsymbol{W} &= \textsc{EmbParams}(\boldsymbol{s}) \\ \boldsymbol{P} &\sim \mathcal{G}(0,1) \\ \boldsymbol{T} &= \textsc{Eisner}(\boldsymbol{W} + \boldsymbol{P}) \end{aligned}$$

The (approximate) Monte-Carlo estimation of the expectation in Equation 3 is then defined as:[11]

$$\begin{aligned} &\mathbb{E}_{q_\phi(\boldsymbol{T}|\boldsymbol{s})} \left[\log p_\theta(\boldsymbol{s}|\boldsymbol{T})\right] \\ &\simeq \log p_\theta(\boldsymbol{s}|\textsc{Eisner}(\boldsymbol{W} + \boldsymbol{P})) \end{aligned}$$

where $\simeq$ denotes a Monte-Carlo estimation of the gradient and $\boldsymbol{P} \sim \mathcal{G}(0,1)$ is sampled in the last line. Therefore, the sampling process is outside the backpropagation path. Unfortunately, the Eisner algorithm is built using One-Hot-Argmax operations that have ill-defined partial derivatives. We propose to replace them with a differentiable surrogate in the next section.

5.2 Differentiable Parsing Algorithm

We now propose a continuous relaxation of the projective dependency algorithm. We start with a brief outline of the algorithm using the parsing-as-deduction formalism, restricting this presentation to the minimum needed to describe our continuous relaxation. We refer the reader to Eisner (1996) for an in-depth presentation.

The parsing-as-deduction formalism provides an unified presentation of many parsing algorithms (Pereira and Warren, 1983; Shieber et al., 1995). In this framework, a parsing algorithm is defined as a deductive system, i.e. as a set of axioms, a goal item and a set of deduction rules. Each deduced item represents a sub-analysis of the input. Regarding implementation, the common way is to rely on dynamic programming: items are deduced in a bottom-up fashion, from smaller sub-analyses to large ones. To this end, intermediate results are stored in a global chart.

[10] Alternatively, it is possible to sample from the set of projective dependency trees by running the inside-out algorithm (Eisner, 2016). However, it is then not straightforward to formally derive a path derivative gradient estimation.

[11] We remove variable $\boldsymbol{z}$ to simplify notation.

Algorithm 1 This function search the best split point for constructing an element given its span. $\boldsymbol{b}$ is a one-hot vector such that $b_{i-k} = 1$ iff k is the best split position.

1: **function** DEDUCE-URIGHT$(i, j, \boldsymbol{W})$
2: $\quad \boldsymbol{s} \leftarrow$ null-initialized vector of size $j - i$
3: $\quad$ **for** $i \leq k < j$ **do**
4: $\quad\quad s_{i-k} \leftarrow [i \triangleright k] + [k+1 \triangleleft j] + w_{j,i}$
5: $\quad \boldsymbol{b} \leftarrow$ ONE-HOT-ARGMAX$(\boldsymbol{s})$
6: $\quad$ BACKPTR$[i \square j] \leftarrow \boldsymbol{b}$
7: $\quad$ WEIGHT$[i \square j] \leftarrow \boldsymbol{b}^\top \boldsymbol{s}$

For projective dependency parsing, the algorithm builds a chart whose items are of the form $[i \square j]$, $[i \square\!\!\triangleleft j]$, $[i \triangleright j]$ and $[i \triangleleft j]$ that represent sub-analyses from word i to word j. An item $[i \triangleright j]$ (resp. $[i \square j]$) represents a sub-analysis where every word $s_k, i \leq k \leq j$ is a descendant of s_i and where s_j cannot have any other modifier (resp. can have). The two other types are defined similarly for descendants of word s_j. In the first stage of the algorithm, the maximum weight of items are computed (deduced) in a bottom-up fashion. For example, the weight WEIGHT$[i \square j]$ is defined as the maximum of WEIGHT$[i \triangleright k]$ + WEIGHT$[k+1 \triangleleft j]$, $\forall k$ s.t. $i \leq k < j$, plus $w_{i,j}$ because $[i \square j]$ assumes a dependency with head s_i and modifier s_j. In the second stage, the algorithm retrieves arcs whose scores have contributed to the optimal objective. Part of the pseudo-code for the first and second stages are given in Algorithm 1 and Algorithm 2, respectively. Note that, usually, the second stage is implemented with a linear time complexity but we cannot rely on this optimization for our continuous relaxation.

This algorithm can be thought of as the construction of a computational graph where WEIGHT, BACKPTR and CONTRIB are sets of nodes (variables). This graph includes ONE-HOT-ARGMAX operations that are not differentiable (see line 5 in Algorithm 1). This operation takes as input a vector of weights $\boldsymbol{v}$ of size k and return a one-hot vector $\boldsymbol{o}$ of the same size

Algorithm 2 If item $[i \square j]$ has contributed the optimal objective, this function sets $t_{i,j}$ to 1. Then, it propagates the contribution information to its antecedents.

1: **function** BACKTRACK-URIGHT$(i, j, \boldsymbol{T})$
2: $\quad t_{i,j} \leftarrow$ CONTRIB$[i \square j]$
3: $\quad \boldsymbol{b} \leftarrow$ BACKPTR$[i \square j]$
4: $\quad$ **for** $i \leq k < j$ **do**
5: $\quad\quad$ CONTRIB$[i \triangleright k] \xleftarrow{+} b_{i-k} t_{i,j}$
6: $\quad\quad$ CONTRIB$[k+1 \triangleleft j] \xleftarrow{+} b_{i-k} t_{i,j}$

with $o_i = 1$ if and only if v_i is the element of maximum value:[12]

$$o_i = \mathbb{1}[\forall 1 \leq j \leq k, j \neq i : v_i > v_j]$$

We follow a recent trend in differentiable approximation of the ONE-HOT-ARGMAX function and replace it with the PEAKED-SOFTMAX operator (Jang et al., 2017; Maddison et al., 2017; Goyal et al., 2017; Goyal et al., 2018; Mensch and Blondel, 2018):

$$o_i = \frac{\exp(1/\tau\ v_i)}{\sum_{1 \leq j \leq k} \exp(1/\tau\ v_j)}$$

where $\tau > 0$ is a temperature hyperparameter controlling the smoothness of the relaxation: when $\tau \to \infty$ the relaxation becomes equivalent to ONE-HOT-ARGMAX. With this update, the parsing algorithm is fully differentiable. Note, however, that outputs are not valid dependency trees anymore. Indeed, then an output matrix $\boldsymbol{T}$ contains continuous values that represent soft-selection of arcs.

5.3 Discussion

Dynamic programs for parsing have been studied as abstract algorithms that can be instantiated with different semirings (Goodman, 1999). For example, computing the weight of the best parse requires the $\langle \mathbb{R}, \max, + \rangle$ semiring. This semiring can be augmented with set-valued operations to retrieve the best derivation. However, a straightforward implementation would

[12] We assume that there are no ties in the weights, which is very likely to happen because (1) we use randomly initialized deep neural networks and (2) weights are perturbed using random Gumbel noise.

have a $\mathcal{O}(n^5)$ space complexity: for each item in the chart, we also need to store the set of arcs. Under this formalism, the backpointer trick is a method to implicitly constructs these sets and maintain the optimal $\mathcal{O}(n^3)$ complexity. Our continuous relaxation replaces the max operator with a smooth surrogate and the set values with an expectation over sets. Unfortunately, $\langle \mathbb{R}, \text{PEAKED-SOFTMAX} \rangle$ is not a commutative monoid, therefore the semiring analogy is not transposable. Mensch and Blondel (2018) introduced a similar approach for tagging with the Viterbi algorithm.

6 Experiments

We ran a series of experiments on 3 different languages to test our method for semi-supervised dependency parsing: English, French and Swedish.

6.1 Corpora

English We use the Stanford Dependency conversion (De Marneffe and Manning, 2008) of the Penn Treebank (Marcus et al., 1993) with the usual section split: 02-21 for training, 22 for development and 23 for testing. In order to simulate our framework under a low-resource setting, the annotation is kept for 10% of the training set only.[13]

French We use a similar setting with the French Treebank version distributed for the SPMRL 2013 shared task and the provided train/dev/test split (Abeillé et al., 2000; Seddah et al., 2013).

Swedish We use the Talbanken dataset (Nivre et al., 2006) which contains two written text parts: the professional prose part (P) and the high school students' essays part (G). We drop the annotation of (G) in order to use this section as unlabeled data. We split the (P) section in labeled train/dev/test using a pseudo-randomized scheme.[14]

[13] A labeled sentence is the sentence which has an index (in the training set) modulo 10 equal to zero.

[14] We follow the splitting scheme of Hall et al. (2006) but fix section 9 as development instead of k-fold cross-validation. Sentence i is allocated to section i mod 10. Then, section 1-8 are used for training, section 9 for dev and section 0 for test.

	Labeled	Unlabeled
English	3984	35848
French	1476	13280
Swedish	4880	5331

Table 1: Number of labeled and unlabeled instances in each dataset.

The size of each dataset is reported in Table 1. Note that the setting is especially challenging for Swedish: the amount of unlabeled data we use here barely exceeds that of labeled data.

6.2 Hyper-parameters of the network

In order to ensure that we do not bias our model for the benefit of the semi-supervised scenario, we use the same parameters as Kiperwasser and Goldberg (2016) for the parser. Also, we did not performed any language-specific parameter selections. This makes us hope that our method can be applied to other languages with little extra effort.

Encoder: word embeddings We concatenate trainable word embeddings of size 100 with external word embeddings.[15] We use the word-dropout settings of Kiperwasser and Goldberg (2016). For English, external embeddings are pre-trained with the structured skip n-gram objective (Ling et al., 2015).[16] For French and Swedish, we use the Polyglot embeddings (Al-Rfou et al., 2013).[17] We stress out that no part-of-speech tag is used as input in any part of our network.

Encoder: dependency parser The dependency parser is built upon a two-stack BiLSTM with a hidden layer size of 125 (i.e. the output at each position is of size 250). Each dependency is then weighted using a single-layer perceptron with a tanh activation function. Arc label prediction rely on a similar setting, we re-

[15] The external embeddings are not updated when training our network.

[16] We use the pre-trained embeddings distributed by Dyer et al. (2015).

[17] We use the pre-trained embeddings distributed at `https://sites.google.com/site/rmyeid/projects/polyglot`

fer to the reader to Kiperwasser and Goldberg (2016) for more information about the parser's architecture.

Encoder: sentence embedding The sentence is encoded into a fixed size vector with a simple left-to-right LSTM with an hidden size of 100. The hidden layer at the last position of the sentence is then fed to two distinct single-layer perceptrons, with an output size of 100 followed by a piecewise tanh activation function, that computes means and standard deviations of the diagonal Gaussian distribution.

Decoder The decoder use fixed pre-trained embeddings only. The recurrent layer of the decoder is a LSTM with an hidden layer size of 100. $\text{MLP}^{\curvearrowright}$, $\text{MLP}^{\curvearrowleft}$ and $\text{MLP}^{\bigcirc}$ are all single-layer perceptrons with an output size of 100 and without activation function.

Training We train our network using stochastic gradient descent for 30 epochs using Adadelta (Zeiler, 2012) with default parameters as provided by the Dynet library (Neubig et al., 2017). In the semi-supervised scenario, we alternate between labeled and unlabeled instances. The temperature of the PEAKED-SOFTMAX operator is fixed to $\tau = 1$. We run the code on a NVIDIA Titan X GPU. For English, the supervised parser took 1.5 hours to train while the semi-supervised parser without sentence embedding, which sees 2 times more instances per epoch, took 3.5 hours to train.[18]

Loss function Previous work has shown that learned latent structures tend to differ from linguistic syntactic structures (Kim et al., 2017; Williams et al., 2018). Therefore, we encourage the VAE to rely on latent structures close to the targeted ones by bootstrapping the training procedure with labeled data only. In the first two epochs, we train the network with the discriminative loss only. Then, for the next two epochs, we add the supervised ELBO term (Equation 6). Finally, after the 6th epoch, we also add the unsupervised ELBO term (Equation 3). Moreover, we follow a common practice for VAEs: we scale down the KL-divergence of priors (Bowman et al., 2016; Miao et al., 2017; Yin et al., 2018). Regarding the distribution over sentence embeddings, we use a 0.01 weight for its KL-divergence with the prior. We experimented using a flat distribution for the dependency tree prior but did not manage to make the training stable in this setting. Therefore, we simply use a weight of 0. Lastly, we scale the gradient flowing from the decoder to the encoder by 0.1. This is similar to downweighting the generative part of the objective but additionally ensures that the decoder both becomes sufficiently accurate early in training and is fast to adapt to any changes in the distribution of the latent variables.

6.3 Parsing results

For each dataset, we train under the supervised and the semi-supervised scenario. Moreover, in the semi-supervised setting, we experiment with and without latent sentence embedding $\boldsymbol{z}$. For the sake of completeness, we also report results with with the same settings as Kiperwasser and Goldberg (2016), i.e. trained with a structured hinge loss (Taskar et al., 2005). Parsing results are summarized in Table 2. Note that our VAE does not take into account arc-labels, therefore they are learned with the discriminative loss only.

We observe a score increase in all three languages. Moreover, we observe that VAE performs slightly better without latent sentence embedding. We assume this is due to the fact that dependencies are more useful when no information leaks in the decoder through $\boldsymbol{z}$. Interestingly, we observe an improvement, albeit smaller, even on Swedish, where we used very limited amount of unlabeled data. We note that training with structured hinge loss gives stronger results than our supervised baseline. In order to maintain the probabilistic interpretation of our model, we did not include a similar term in our model.

We also analyze the results in English with respect to dependency lengths, see Table 3.[19] We

[18] To achieve fast training in the semi-supervised scenario, we optimized the implementation of the continuous relaxation of the Eisner. Similarly to Mensch and Blondel (2018), we observed a significant loss of efficiency when implementing it with vanilla Dynet.

[19] We used the evaluation script from the SPMRL

	English	French	Swedish
Supervised	88.79 / 84.74	84.09 / 77.58	86.59 / 78.95
VAE w. $\boldsymbol{z}$	89.39 / 85.44	84.43 / 77.89	86.92 / 80.01
VAE w/o $\boldsymbol{z}$	89.50 / 85.48	84.69 / 78.49	86.97 / 79.80
Kipperwasser & Goldberg	89.88 / 86.49	84.30 / 77.83	86.93 / 80.12

Table 2: Parsing results: unlabeled attachment score / labeled attachment score. We also report results with the parser of Kiperwasser and Goldberg (2016) which uses a different discriminative loss for supervised training.

Distance	Supervised Re / Pr	Semi-supervised Re / Pr
(to root)	93.46 / **89.30**	93.84 / **92.41**
1	95.61 / 94.07	95.33 / 94.57
2	93.01 / 90.88	92.50 / 92.09
3 . . . 6	85.95 / 88.13	87.31 / 87.93
> 7	**72.47** / 83.26	**78.72** / 83.11

Table 3: Recall / Precision evaluation with respect to dependency lengths for the supervised parser and the best semi-supervised parser on the English test set. Bold numbers highlight the main differences.

observe that the semi-supervised parser tends to correct two kind of errors. Firstly, it makes fewer mistakes on root attachments, i.e. the recall is similar between the two parsers but the precision of the semi-supervised one is higher. We hypothesis that root attachment errors come at a high price in the decoder because there is only a small fraction of the vocabulary that is observed with this syntactic function. Secondly, the semi-supervised parser recovers more long distance relations, i.e. the recall for dependencies with a distance superior or equal to 7 is higher. Intuitively, we assume these dependencies are more useful in the decoder: for short distance dependencies, the LSTM efficiently captures the context of the word to predict, whereas this information could be vanishing for long distances, meaning the GCN has more more impact on the prediction.

2013 shared task: http://www.spmrl.org/spmrl2013-sharedtask.html

7 Related work

Dependency parsing in the low-ressource scenario has been of interest in the NLP community due to the expensive nature of annotation. On the one hand, transfer approaches learn a delexicalized parser for a resource-rich language which is then used to parse a low-resource one (Agić et al., 2016; McDonald et al., 2011). On the other hand, the grammar induction approach learns a dependency parser in an unsupervised manner. Klein and Manning (2004) introduced the first generative model that outperforms the right-branching heuristic in English. Close to our work, Cai et al. (2017) use an auto-encoder setting where the decoder tries to rebuild the source sentence. However, their decoder is unstructured (e.g. it is not auto-regressive).

Variational Auto-Encoders (Kingma and Welling, 2013; Rezende et al., 2014) have been investigated in the semi-supervised settings (Kingma et al., 2014) for NLP. Kočiský et al. (2016) learn a semantic parser where the latent variable is a discrete sequence of symbols. Zhou and Neubig (2017) successfully applied the variational method to semi-supervised morphological re-inflection where discrete latent variables represent linguistic features (e.g. tense, part-of-speech tag). Yin et al. (2018) proposed a semi-supervised semantic parser. Similarly to our model, they rely on a structured latent variable. However, all of these systems use either categorical random variables or the REINFORCE score estimator. To the best of our knowledge, no previous work used continuous relaxation of a dynamic programming latent variable in the VAE setting.

The main challenge is backpropagation

through discrete random variables. Maddison et al. (2017) and Jang et al. (2017) first introduced the Gumbel-Softmax operator for the categorical distribution. There are two issues regarding more complex discrete distributions.

Firstly, one have to build a reparametrization of the the sampling process. Papandreou and Yuille (2011) showed that low-order perturbations provide samples of good qualities for graphical models. While independent factor perturbation in a graphical model may not capture the underlying structure of the latter, Gane et al. (2014) proposed to learn perturbations that implicitly model the structure. Either way, the sampling process is reduced to a structured arg max with a perturbed objective function.

Secondly, one have to build a good differentiable surrogate to the structured arg max operator. Early work replaced the structured arg max with structured attention (Kim et al., 2017). However, computing the marginals over the parse forest is sensitive to numerical stability outside specific cases like non-projective dependency parsing (Liu and Lapata, 2018; Tran and Bisk, 2018). Moreover, we are interested in continuous surrogates that are more peaked toward the maximum. It is well-known that differentiating through parametrized linear programs can be achieved by introducing (non-linear) penalties that prevent inequality constraints activation (Gould et al., 2016). Mensch and Blondel (2018) studied this setting for dynamic program smoothing. They experiments with the linear-chain Viterbi and the DTW algorithm. Our approach is highly related but we describe a practical implementation with a different type of dynamic programs, i.e. using the parsing-as-deduction formalism. Peng et al. (2018) propose to replace the true gradient with a proxy that tries to satisfy constraints on a arg max operator via a projection. However, their approach is computationally expensive, so they remove the tree constraint on dependencies during back-propagation. A parallel line of work focuses on sparse structures that are differentiable (Martins and Astudillo, 2016; Niculae et al., 2018).

8 Conclusions

We presented a novel generative learning approach for semi-supervised dependency parsing. We model the dependency structure of a sentence as a latent variable and build a Variational Auto-Encoder. To effectively train the posterior distribution over dynamic programming latent variables, we build a continuous relaxation of the parsing algorithm.

This work could be extended to the unsupervised scenario. Prior over the latent dependency distribution could be used to introduce linguistic knowledge into the parser (Naseem et al., 2010; Noji et al., 2016).

Acknowledgments

We thank Diego Marcheggiani, Wilker Ferreira Aziz and Serhii Havrylov for their comments and suggestions. The project was supported by the the Dutch National Science Foundation (NWO VIDI 639.022.518) and European Research Council (ERC Starting Grant BroadSem 678254).

References

Abeillé et al., 2000 Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for french. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA).

Agić et al., 2016 Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Al-Rfou et al., 2013 Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bastings et al., 2017 Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Simaan. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pages 1947–1957. Association for Computational Linguistics.

Bowman et al., 2016 Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics.

Cai et al., 2017 Jiong Cai, Yong Jiang, and Kewei Tu. 2017. CRF autoencoder for unsupervised dependency parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1638–1643. Association for Computational Linguistics.

Chen et al., 2009 Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto, Yujie Zhang, and Hitoshi Isahara. 2009. Using short dependency relations from auto-parsed data for chinese dependency parsing. *ACM Transactions on Asian Language Information Processing*, pages 10:1–10:20.

Cui et al., 2005 Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 400–407. ACM.

Culotta and Sorensen, 2004 Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

De Marneffe and Manning, 2008 Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Ding and Palmer, 2005 Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 541–548. Association for Computational Linguistics.

Dozat and Manning, 2017 Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 2017 International Conference on Learning Representations*.

Dyer et al., 2015 Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343. Association for Computational Linguistics.

Eisner, 1996 Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Eisner, 2016 Jason Eisner. 2016. Inside-outside and forward-backward algorithms are just backprop (tutorial paper). In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 1–17. Association for Computational Linguistics.

Gane et al., 2014 Andreea Gane, Tamir Hazan, and Tommi Jaakkola. 2014. Learning with maximum a-posteriori perturbation models. In *Artificial Intelligence and Statistics*, pages 247–256.

Goodman, 1999 Joshua Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25(4).

Gould et al., 2016 Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. 2016. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *CoRR*, abs/1607.05447.

Goyal et al., 2017 Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. Differentiable scheduled sampling for credit assignment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 366–371. Association for Computational Linguistics.

Goyal et al., 2018 Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. 2018. A continuous relaxation of beam search for end-to-end training of neural sequence models. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana, February.

Gumbel, 1954 Emil Julius Gumbel. 1954. *Statistical theory of extreme values and some practical applications: a series of lectures*. Number 33. US Govt. Print. Office.

Hall et al., 2006 Johan Hall, Joakim Nivre, and Jens Nilsson. 2006. Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 316–323. Association for Computational Linguistics.

Jang et al., 2017 Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with

gumbel-softmax. In *Proceedings of the 2017 International Conference on Learning Representations*.

Johnson et al., 1999 Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic "unification-based" grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Jordan et al., 1999 Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kaplan and Bresnan, 1982 Ronald M Kaplan and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. *Formal Issues in Lexical-Functional Grammar*, pages 29–130.

Kawahara and Uchimoto, 2008 Daisuke Kawahara and Kiyotaka Uchimoto. 2008. Learning reliability of parses for domain adaptation of dependency parsing. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Kim et al., 2017 Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. In *Proceedings of the 2017 International Conference on Learning Representations*.

Kingma and Welling, 2013 Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma et al., 2014 Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc.

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

Kipf and Welling, 2016 Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Klein and Manning, 2004 Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

Kočiský et al., 2016 Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. 2016. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1078–1087. Association for Computational Linguistics.

Kong et al., 2015 Lingpeng Kong, Alexander M. Rush, and Noah A. Smith. 2015. Transforming dependencies into phrase structures. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 788–798. Association for Computational Linguistics.

Koo et al., 2008 Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603. Association for Computational Linguistics.

Lafferty et al., 2001 John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*.

LeCun et al., 2012 Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. 2012. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Ling et al., 2015 Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304. Association for Computational Linguistics.

Linnainmaa, 1976 Seppo Linnainmaa. 1976. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160.

Liu and Lapata, 2018 Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Liu et al., 2015 Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng WANG. 2015. A dependency-based neural network for relation classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 285–290. Association for Computational Linguistics.

Ma and Hovy, 2017 Xuezhe Ma and Eduard Hovy. 2017. Neural probabilistic model for non-projective mst parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 59–69, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Maddison et al., 2014 Chris J. Maddison, Daniel Tarlow, and Tom Minka. 2014. A* Sampling. In *Advances in Neural Information Processing Systems 27*.

Maddison et al., 2017 Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*.

Marcheggiani and Titov, 2017 Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1507–1516. Association for Computational Linguistics.

Marcus et al., 1993 Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Martins and Astudillo, 2016 Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

McClosky et al., 2006 David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics.

McDonald et al., 2005 Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

McDonald et al., 2011 Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

Mensch and Blondel, 2018 Arthur Mensch and Mathieu Blondel. 2018. Differentiable dynamic programming for structured prediction and attention. In *In Proceedings of International Conference on Machine Learning*.

Miao et al., 2017 Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*.

Mikolov et al., 2013 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mnih and Gregor, 2014 Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning*.

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

Neubig et al., 2017 Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Niculae et al., 2018 Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. 2018. SparseMAP: Differentiable sparse structured inference. In *Proceedings of ICML 2018*.

Nivre et al., 2006 J. Nivre, J. Nilsson, and J. Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Noji et al., 2016 Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing

to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43. Association for Computational Linguistics.

Papandreou and Yuille, 2011 George Papandreou and Alan L Yuille. 2011. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 193–200. IEEE.

Peng et al., 2018 Hao Peng, Sam Thomson, and Noah A. Smith. 2018. Backpropagating through structured argmax using a spigot. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1863–1873. Association for Computational Linguistics.

Pereira and Warren, 1983 Fernando C. N. Pereira and David H. D. Warren. 1983. Parsing as deduction. In *21st Annual Meeting of the Association for Computational Linguistics*.

Reddy et al., 2016 Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

Rezende et al., 2014 Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun. PMLR.

Sagae and Tsujii, 2007 Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Seddah et al., 2013 Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182. Association for Computational Linguistics.

Shieber et al., 1995 Stuart M Shieber, Yves Schabes, and Fernando CN Pereira. 1995. Principles and implementation of deductive parsing. *The Journal of logic programming*, 24(1-2):3–36.

Smith and Eisner, 2008 David Smith and Jason Eisner. 2008. Dependency parsing by belief propagation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 145–156. Association for Computational Linguistics.

Sutskever et al., 2014 Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Suzuki et al., 2011 Jun Suzuki, Hideki Isozaki, and Masaaki Nagata. 2011. Learning condensed feature representations from large unsupervised data sets for supervised learning. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 636–641. Association for Computational Linguistics.

Tarjan, 1977 Robert Endre Tarjan. 1977. Finding optimum branchings. *Networks*, 7(1):25–35.

Taskar et al., 2005 Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM.

Tesnière, 1959 Lucien Tesnière. 1959. *Les éléments de Syntaxe structurale*. Editions Klincksieck.

Tran and Bisk, 2018 Ke Tran and Yonatan Bisk. 2018. Inducing grammars with and for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 25–35. Association for Computational Linguistics.

Wainwright et al., 2008 Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Williams et al., 2018 Adina Williams, Andrew Drozdov, and Samuel R Bowman. 2018. Do latent

tree learning models identify meaningful structure in sentences? *Transactions of the Association for Computational Linguistics*, 6:253–267.

Williams, 1992 Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Xu et al., 2017 Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *AAAI*, pages 3358–3364.

Yin et al., 2018 Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–765. Association for Computational Linguistics.

Yu et al., 2008 Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. 2008. Chinese dependency parsing with large scale automatically constructed case structures. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1049–1056. Coling 2008 Organizing Committee.

Zeiler, 2012 Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhou and Neubig, 2017 Chunting Zhou and Graham Neubig. 2017. Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 310–320. Association for Computational Linguistics.