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

Corro & Titov, ICLR 2019

Tom Effland

April 24, 2019

Overview

- First, we'll discuss the core idea of the paper, relaxed perturb-and-MAP, abstracting over parsing-specific details - this is what can actually be of use to the class.

- Then we can discuss the idea's application to parsing, if people care.
(But we still won't discuss the Eisner algorithm.)

General Treatment

Problem Statement

In NLP, we often want model some discrete structure given an input observation. (Let's call this inference)

- Observation $x \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots$,
  (e.g., $\mathcal{X} =$ set of variable length discrete sequences in vocab)
- Inferred structure $y \in \mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots$,
  (e.g., $\mathcal{Y} =$ set discrete label sequences, discrete segmentation, grammar derivation)
- Want to learn: $p_\phi(y|x) : \mathcal{X} \to \Delta_{\mathcal{Y}}$
- Want to predict: $\hat{y} \leftarrow \arg\max_{y' \in \mathcal{Y}} p_\phi(y'|x)$

Modeling Inference: Motivating Approaches

What are ways people approach this?

Why not just use a tractable generative model? (e.g., HMM or PCFG)
$p_\theta(y_{1:M}|x) \propto p_\theta(y_{1:M}, x)$

- They are too restrictive in the modeling assumptions
- $\Rightarrow$ They underperform, discriminative (conditional) models work better

Ok, use directed (locally normalized) conditional model: $p_\phi(y_{1:M}|x) = \prod_{i=1}^{M} p_\phi(y_i|y_{<i}, x)$

- No longer need independence assumptions on inputs (think naive bayes vs. logistic regression) or outputs for that matter
- But there's the problem when predicting structures:
  directed conditional models have a limited ability for later decisions to revise earlier ones, especially with beam-search

Modeling Inference: CRFs

Conditional Random Fields: Structure is influenced bidirectionally

- If your model decoding order doesn't reflect a causal process, undirected model is probably more appropriate

Instead of local normalization:

$$p_\phi(y_{1:M}|x) = \prod_{i=1}^{M} p_\phi(y_i|y_{<i}, x) = \prod_{i=1}^{M} \frac{\exp\{\phi(y_i|y_{<i}, x)\}}{\sum_{y_i'} \exp\{\phi(y_i'|y_{<i}, x)\}}$$

Global normalization:

$$p_\phi(y_{1:M}|x) = \frac{\prod_{i=1}^{M} \exp\{\phi(y_i|y_{<i}, x)\}}{\underbrace{\sum_{y' \in \mathcal{Y}} \prod_{i=1}^{M} \exp\{\phi(y_i'|y_{<i}, x)\}}_{Z_\phi(x)}}$$

When ϕ factor graph for y is a tree, $Z_\phi(x)$ is computable in polynomial time with dynamic programming (e.g., forward-backward, **sum-product**)

Semi-Supervised Learning

For semi-supervised learning, generative models are an attractive solution for learning on additional unsupervised data

- Principled: optimize marginal likelihood
- Prior can impose regularization
- Appropriate generative model can provide useful signal for inference

Embed our CRF inference model as the amortized approximate posterior in an VI setup! New setup, unsupervised case:

$$p_\theta(x)p_\theta(y|x), \quad q_\phi(y|x) \leftarrow p_\phi(y|x)$$
$$\log p_\theta(x) \geq \mathbb{E}_{y\sim q_\phi}[\log p_\theta(x|y)] - KL(q_\phi||p_\theta(y))$$

One MAJOR problem though, the usual one:

- What about $\nabla_\phi \mathbb{E}_{y\sim q_\phi}[\log p_\theta(x|y)]$?
- REINFORCE is often very poorly behaved in these situations

Remember Gumbel-Softmax?

Can draw a sample from a categorical with

$$\tilde{y} = \arg\max_{y'} \{\log \pi_{y'} + \gamma_{y'}\}, \quad \gamma_{y'} \sim \mathcal{G}(0,1)$$

and can draw a "relaxed" sample with

$$\tilde{y}_r = \frac{\exp\{\log \pi_{y'} + \gamma_{y'}\}}{\sum_{y'} \exp\{\log \pi_{y'} + \gamma_{y'}\}}, \quad \gamma_{y'} \sim \mathcal{G}(0,1)$$

Relaxed Perturb-and-MAP

(Gumbel-Softmax for tractable CRFs)

Can draw a sample from a CRF using Perturb-and-MAP [Papandreou and Yuille '11]

$$\tilde{y} = \arg\max_{y \in \mathcal{Y}} q_{\phi+\tilde{\gamma}}(y|x)$$

Gradient of log partition function is the joint distribution [Eisner '16, Mencsh and Blondel '18]

$$\nabla \log Z_\phi(x) = q_\phi(y|x)$$

and it's zero temperature limit is the MAP estimate (as one-hots)

$$\nabla \log Z_\phi(x;\tau) = q_\phi(y|x;\tau) \xrightarrow{\tau\to 0} \arg\max_{y\in\mathcal{Y}} q_\phi(y|x)$$

So we have that the gradient of the perturbed partition function converges to a sample as the temp approaches zero

$$\nabla \log Z_{\phi+\tilde{\gamma}}(x;\tau) = q_{\phi+\tilde{\gamma}}(y|x;\tau) \xrightarrow{\tau\to 0} \arg\max_{y\in\mathcal{Y}} q_{\phi+\tilde{\gamma}}(y|x) = \tilde{y}$$

Takeaway: Perturb and temper potentials, then run inference
$\Rightarrow$ Marginals are a relaxed sample from the CRF

Application to Dependency Parsing

Dependency Parsing

Dependency grammar is a formalism of *syntax* for how words modify each other in a sentence

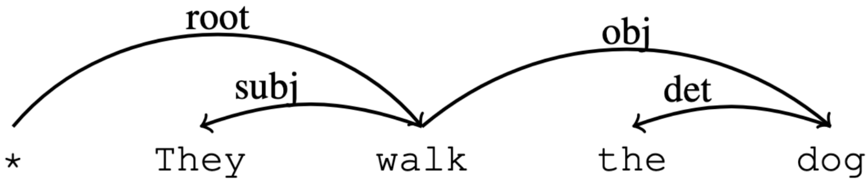


Figure: Example dependency structure

It can be represented as an adjacency matrix A (ignoring labels) where columns $A_{\cdot,j}$ sum to 1.

An entry at $A_{i,j} = 1$ if the edge $x_i \to x_j$ exists.

Trees are also *projective* – no crossing edges.

Inference Model

Model can be viewed as a CRF, with a potential for each cell in the matrix plus a special fully connected factor that ensures the tree constraints.

The model potential scores for some valid tree T are

$$q_\phi(T|x) = \frac{\exp\{\phi(T, W(x))\}}{\sum_{T' \in \mathcal{T}} \exp\{\phi(T', W(x)\}}, \quad \phi(T, W) = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} T_{ij} W_{ij}(x)$$

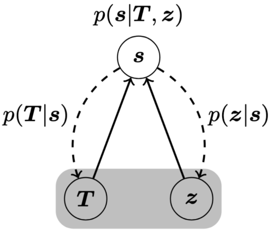
Projectivity of the tree implies that the argmax and marginals can be inferred in $O(n^3)$ (Eisner's Algorithm)

Also have a latent sentence vector $q_\phi(z|x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$ from sentence encoding

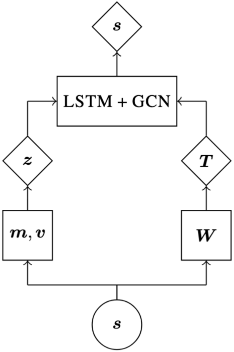
Generation

Assume a generative model, for known sentence length n :

- $z \sim p(z|n) = \mathcal{N}(0, I_d)$
- $T \sim p(T|n)$ ▷ Uniform distribution of rooted projective tree matrices
- $x_{1:n} \sim p_\theta(x_{1:n}|z, T, n) = \prod_{i=1}^{n} p_\theta(x_i|x_{<i}, z, T_{\leq i, \leq i})$



(a) Probabilistic model



(b) Computation Graph

Semi-Supervised Learning

They use the standard Semi-supervised VAE objective:

$$\mathcal{J}_L(\theta, \phi; x, T) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|T, z)]}_{\mathbb{E}_\epsilon[\log p_\theta(x|T, z_\phi(x, \epsilon))]} - \alpha_z KL(q_\phi(z|x)||p(z)) + \log q_\phi(T|x)$$

$$\mathcal{J}_U(\theta, \phi; x) = \underbrace{\mathbb{E}_{q_\phi(z,T|x)}[\log p_\theta(x|T, z)]}_{\mathbb{E}_{P,\epsilon}[\log p_\theta(x|T_\phi(x, P; \tau), z_\phi(x, \epsilon))]} - \alpha_z KL(q_\phi(z|x)||p(z)) - \alpha_T KL(q_\phi(T|x)||p(T))$$

$$\mathcal{L}(\theta, \phi; \mathcal{D}_L, \mathcal{D}_U) = \mathbb{E}_{(x,T)\sim\mathcal{D}_L}[\mathcal{J}_L] + \mathbb{E}_{(x)\sim\mathcal{D}_U}[\mathcal{J}_U]$$

Note: Strange balancing of objectives – OK, due to a combo of the datasets not being too heavily imbalanced towards $\mathcal{D}_U$ and the small KL weights reducing the impact of unsupervised regularization

Experiments

Test the state-of-the-art parsing architecture on three standard datatsets:

	Labeled	Unlabeled
English	3984	35848
French	1476	13280
Swedish	4880	5331

Figure: Dataset info.

	English	French	Swedish
Supervised	88.79 / 84.74	84.09 / 77.58	86.59 / 78.95
VAE w. z	89.39 / 85.44	84.43 / 77.89	86.92 / 80.01
VAE w/o 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

Figure: Results: Edge Precision / Recall. Considerable improvement from unlabeled data, approaches fully supervised performance w/ 10% of the data

Worth noting: have to set KL weight for T to 0 and z to $.1$