Latent Constituents via Self-Attention

July 25, 2018

1 Introduction

2 Problem

We would like to learn a generative model over source sentences $\mathbf{x} = \{x_0, x_1, \ldots\}$, using a distribution over latent trees $\mathbf{z}$. (Shen et al., 2018) Yin et al. (2018)

3 PRPN

(Shen et al., 2018)

- $d_t \in \mathbb{R}^+$ is a score which is used to rank a token's propensity to be the beginning of a constituent, or how high up in the tree it should be. These d_t are used to gate self attention by preventing a token from attending past another token i if $d_i > d_t$.
-

Comment on Shen et al. (2018)'s figures, namely 1 through 3: a ternary tree is not exactly possible under their model (regardless of τ). The way attention is parameterized, if $tau = 0$ then it would be better to view the tree as a fully left-branching binary tree.

4 Model

4.1 Generative Model

Parameterize with $d_t \sim \mathcal{N}(\mu_t, \sigma_t)$. Reparameterize comparisons with gumbel softmax? Leave all self attentions as is?

- $p(z)$
- $p(x|z)$

5 Training and Inference

References

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. Neural language modeling by jointly learning syntax and lexicon. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkgOLb-0W`.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018. URL `https://arxiv.org/abs/1806.07832v1`.