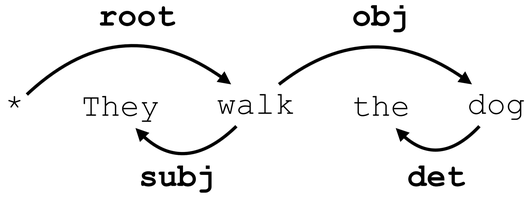
Differentiable Perturb-and-Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder

Caio Corro, **Ivan Titov**

ILCC, School of Informatics, University of Edinburgh ILLC, University of Amsterdam

Dependency Tree

Syntactic structure that is useful in downstream tasks but annotation is expensive: datasets are small for many major languages (e.g. Vietnamese).



T: Tree described as an adjacency matrix

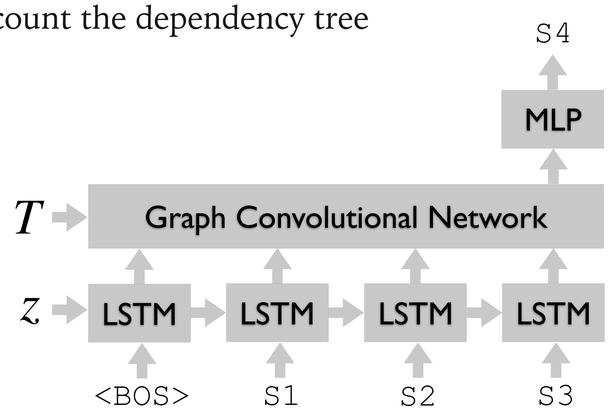
$\mathcal{T}(s)$: set of trees compatible with sentence s

W: matrix of arc weights computed with a NN

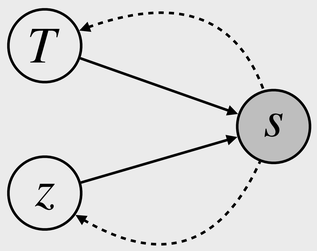
$$\underset{T \in \mathcal{T}(s)}{argmax} \sum_{h,m} T_{h,m} \times W_{h,m}$$

Syntactic language Model

Auto-regressive model that takes into account the dependency tree



Semi-Supervised Variatonal Auto-Encoder



We assume a sentence is generated from a latent tree and a latent embedding:

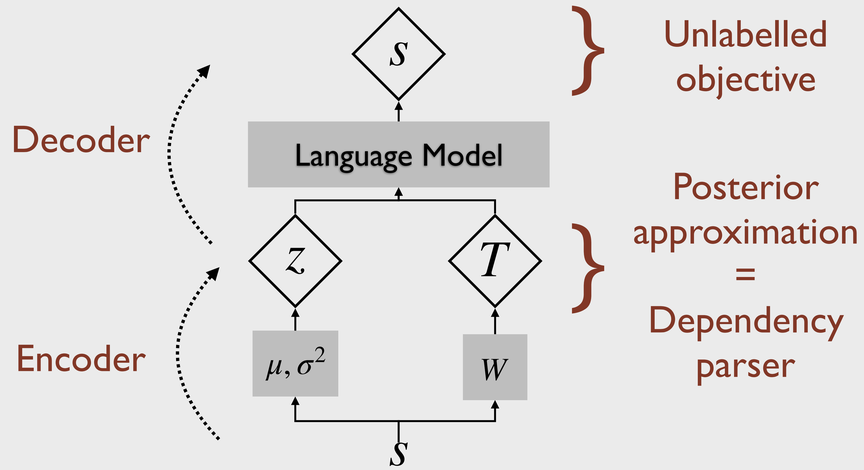
$$p(s) = \sum_T \int p(s, T, z)\, dz$$

We perform variational inference by jointly learning a distribution which is intended to be close to the posterior:

$$KL\left[q(T, z \mid s) \,\|\, p(T, z \mid s)\right] \simeq 0$$

$$\mathscr{L} = -\sum_j \log p(s_j)$$ } Unsupervised auto-encoder loss [Kingma & Welling, 2013]

$$-\sum_i \log q(T_i \mid s_i)$$ } Supervised loss [Kingma et al., 2014]



Perturb-and-Parse

As exact marginalisation over dependency trees is intractable, we introduce a reparametrization for differentiable Monte-Carlo estimation.

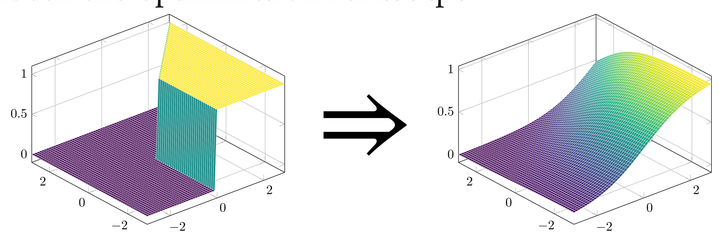
$$G \sim \mathcal{G}(0,1)$$

$$\widetilde{W} = W + G$$ } Arc weight perturbation with Gumbel noise [Papandreou & Yuille, 2011]

$$\underset{T \in \mathcal{T}(s)}{argmax} \sum_{h,m} T_{h,m} \times \widetilde{W}_{h,m}$$ } Solved with dynamic programming [Eisner, 1996]

Differentiable Dynamic Programming

The dynamic programming approach for parsing relies on recursive calls to the *one-hot-argmax* op, which introduces ill-defined derivatives during the backward pass. We replace *one-hot-argmax* ops with *softmax* ops to smooth the optimization landscape.



[Maddison et al., 2017; Goyal et al., 2018]

Acknowledgments

NWO VIDI 639.022.518
ERC BroadSem 678254

Experimental results

	English	French	Swedish
Supervised	88.79 / 84.74	84.09 / 77.58	86.59 / 78.95
VAE with z	89.39 / 85.44	84.43 / 77.89	86.92 / 80.01
VAE without z	**89.50 / 85.48**	**84.69 / 78.49**	**86.97 / 79.80**

Unlabeled/labeled attachment scores. Only 10% of the data is labeled (~4000 sentences for English)

	Supervised prec/recall	Semi-sup. prec/recall
(root)	93.46 / **89.30**	93.84 / **92.41**
> 7	**72.47** / 83.26	**78.72** / 83.11

The main improvement is observed on root word identification and long distance dependencies (arcs crossing at least 7 words)