Enhancing Unsupervised Generative Dependency Parser with Contextual Information

Wenjuan Han · Yong Jiang · Kewei Tu
School of Information Science and Technology, ShanghaiTech University

Task: Unsupervised Dependency Parsing

Unsupervised Dependency parsing: induce a grammar from a set of unannotated sentences.

Previous Approaches

1. DMV (Klein and Manning, 2004) is a widely used generative model. – The dependency tree is decomposed into factorized grammar rules – A scalar probability is associated to each rule

2. NDMV (Jiang et al., 2016): predict rule probabilities using a neural network parameterized by Θ.

Cons of DMV and NDMV: strong conditional independence assumption which prohibits usage of global features of the entire sentence

Discriminative Neural Dependency Model with Valence

We rely on not only local information but also a sentence representation s when predicting rule probabilities, breaking the conditional independence assumption. The model can be regarded as an autoencoder model: Encoder: Bi-LSTM with parameters Φ

– Transform the sentence w to a hidden representation s

Decoder: conditional NDMV with parameters Θ

– Generate the dependency tree z as well as the POS tag sequence x conditioned on the hidden representation s

How to parse? => CYK algorithm when P_θ(s) is calculated.

Learning objective function

Log conditional likelihood (We may replace summation with maximization so that it becomes conditional Viterbi likelihood):

\[
J(Θ,Φ) = \frac{1}{N} \sum_{i=1}^{N} \log f(Θ,Φ(w^{(i)}|x^{(i)}))
\]

\[
\log P_Θ(\mathbf{x}|w) = \sum_{\mathbf{z} \in \mathcal{Z}(\mathbf{x})} P_Θ(\mathbf{x}|\mathbf{z}) P_Θ(\mathbf{z}|w)
\]

Learning Algorithm

EM algorithm to maximize the lower-bound of objective (q(z)) is an auxiliary distribution:

\[
Q(Θ,Φ) = \log P_Θ(\mathbf{x}|w) - KL(q(\mathbf{z})||P_Θ(\mathbf{x}|\mathbf{z},w))
\]

– E-step: compute the expected counts E_q(0, x, z) based on the optimal q which is set as

\[
q(z) = P_Θ(\mathbf{z}|\mathbf{x},w)
\]

– M-step: back-propagate the following objective into the parameters Θ, Φ.

\[
Q(Θ,Φ) = \sum_{r,z} \log p_r(\mathbf{z}|w) E_q(0, r, x, z) - \text{Constant}
\]

where r ranges over all the grammar rules

Variational Variant for D-NDMV

The aforementioned method to generate s is deterministic. We propose a variational variant where generating s is a sampling process and follows a Gaussian prior:

Learning Algorithm: Learning is almost the same as the deterministic variant except for a variational posterior distribution q and an additional KL term −KL(q(z)||p(z)) in the objective

Note: variational variant shares the same formulation of the encoder with VAE.

Experiments and Analysis

Figure 1. The rule probability predicted by D-NDMV given the first sentence is indeed significantly larger than that given the second sentence, which demonstrates the positive impact of conditioning rule probability prediction on the sentence embedding.

Analysis:

We analyze how the contextual information influence grammar rule probabilities. We examine the rule VBZ → JJ with valence 0 in two selected sentences (left) and the whole dataset (right).

Figure 2. Two distributions over the rule probability when the rule is used in the gold parse vs. when the rule is applicable to parsing the sentence but is not used in the gold parse.