Constituency Parsing with a Self-Attentive Encoder

Nikita Kitaev and Dan Klein

Overview

State-of-the-art parsing results with self-attention and a novel method for factoring position and content information. Achieves over 95 F1 on English!

Decoder

Structure
The score of a tree factors over labelled spans:

\[ s(T) = \sum_{i,j \in T} s(i,j; t) \]

A feed-forward network computes span scores:

\[ s(i,j; t) = M_{\text{Self-LayerNorm}}(x^*, x_i, x_j) + c_1 + c_2 \]

where \( x = Y^T \cdot Y_{i,j} - Y_{i,j} \cdot Y_{i,j} \)

Objective
Train to have the correct objective

A feed-forward network over labelled spans:

Structure
Vinyals et al. (2015)
Gaddy et al. (2018)
Cross and Huang (2016)
Choe and Charniak (2016)

Results

English
Single Model, WSJ Only
Vinyls et al. (2015)
Cross and Huang (2016)
Gaddy et al. (2018)
Lam et al. (2017)
Ours with CharLSTM

Multi-Model / External
Vinyls et al. (2015)
Cross and Huang (2016)
Choe and Charan (2018)
Liu and Zhang (2017)
Fried et al. (2017)
Ours with ELMo
2x Ours with ELMo

Multilingual (SPMRL)

What Helps?

Long-Distance Context Matters
8 layers of self-attention with window size 10 have an effective receptive field of size 80. But the full, un-windowed model is still better!

Code and models are publicly available
Includes integration with popular NLP toolkits: spaCy and NLTK

Sample Usage (with spaCy)

Sample Usage (with NLTK)

More at: github.com/nikitakit/self-attentive-parser