Stack-based Multi-layer Attention for Transition-based Dependency Parsing

Zhirui Zhang¹, Shujie Liu², Mu Li², Ming Zhou², Enhong Chen¹
University of Science and Technology of China¹ and Microsoft Research Asia²

Abstract

Although sequence-to-sequence (seq2seq) network has achieved significant success in many NLP tasks such as machine translation and text summarization, simply applying this approach to transition-based dependency parsing cannot yield a comparable performance gain as in other state-of-the-art methods, such as stack-LSTM and head selection. In this paper, we propose a stack-based multi-layer attention model for seq2seq learning to better leverage structural linguistics information. In our method, two binary vectors are used to track the decoding stack in transition-based parsing, and multi-layer attention is introduced to capture multiple word dependencies in partial trees. We conduct experiments on PTB and CTB datasets, and the results show that our proposed model achieves state-of-the-art accuracy and significant improvement in labeled precision with respect to the baseline seq2seq model.

Background

Sequence-to-sequence Learning:
Follow the attention-based encoder-decoder architecture

Encoder
The encoder reads in the source sentence \( X = \{x_1, x_2, \ldots, x_T \} \) and transforms it into a sequence of hidden states \( h = \{h_1, h_2, \ldots, h_T \} \) using a bi-directional RNN

Decoder
The decoder uses another RNN to generate a corresponding target sequence \( Y = \{y_1, y_2, \ldots, y_T \} \) based on the hidden states \( h = \{h_1, h_2, \ldots, h_T \} \)

At each time \( i \), the conditional probability of target symbol \( y_i \) is computed by

\[
p(y_i | y_{<i}, h_i) = \text{softmax}(g(\text{emb}(y_i), z_i, c_i))
\]

Where \( z_i \) is the hidden state of the decoder and \( c_i \) is the source context vector

Attention Mechanism
In attention-based seq2seq model, the context vector \( c_i \) is a weighted sum of the hidden states \( h = \{h_1, h_2, \ldots, h_T \} \) with the coefficients \( a_{i,1}, a_{i,2}, \ldots, a_{i,T} \)

\[
a_{i,t} = \frac{\exp(e_{i,t})}{\sum_k \exp(e_{i,k})}
\]

\[
e_{i,t} = v^T \tanh(W_{d}z_{i,t} + U_{d}h_{t})
\]

Analysis

Table 1: Results of various state-of-the-art parsing systems on English dataset (PTB with Stanford Dependencies) and Chinese dataset (CTB). The numbers reported from different systems are taken from: Z&N11 (Zhang and Nivre, 2011); C&A14 (Chen and Manning, 2014); CoBiSA (Wiseman and Rush, 2016); DyER15 (Dyer et al., 2015); Weiss15 (Weiss et al., 2015); K&G16 (Kiperwasser and Goldberg, 2016); DENSE (Zhang et al., 2017).

Note that Dozat and Manning(2016) achieve 95.74 UAS and 89.30 UAS on PTB-SD and CTB datasets respectively. For ensemble, we train 4 models using the same network with different random initialization.

Table 2: Impact of \( f \) on English PTB dataset.

Table 3: Impact of the different components on English PTB dataset.