WHAT DO RECURRENT NEUTAL NETWORK GRAMMARS LEARN ABOUT SYNTAX? BY NIKITA DATCENKO

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

26.06.2017

Fitting a Recurrent Neural Network Grammar (RNNG) to data is a way to test or confirm some aspect of a theory. Neural networks are capable of representing larger classes of hypotheses than traditional probabilistic methods. RNNGs obtain state-of-the-art parsing and language modeling performance. A Recurrent Neural Network is one type of neural networks, which performs the same task for every element in the sequence with the output being depended on the previous computations. In theory RNNs can make use of information in arbitrarily long sequences, however practically they are limited to looking a few steps back. Which is why LSTM (long short-term memory) is used, because long-term dependencies can be captured, basically it is a different way of computing the hidden state. A RNNG defines a joint p-robability distribution over string terminals and phrase-structure nonterminals and is represented in 3 sets: a set of nonterminals, a set of terminals and a set of all model parameters. The 3 different actions that are used by RNNG are:

- Introduces an open non-terminal symbol onto the stack
- Generates a terminal symbol and places it on the stack and buffer
- A composition function is executed, yielding a composed representation that is pushed onto the stack.

As a result, the stack-only RNNG achieves the best performance and the ablating the stack is most harmful.

An attention mechanism can be viewed as a method for making the RNN work better by letting the network know where to look as it is performing its task. There are 2 theories about phrasal representation:

- Phrasal representation are strongly determined by a privileged lexical head (head of a phrase is the word that determines the syntactic category
- Does phrase-internal material wholly determine the representation of a phrase (endocentric) or does nonterminal relabeling of a constituent introduces new information (exocentric)?

In noun phrases the model pays the most attention to the rightmost noun and assigns near-zero attention on determiners and possessive determiners, while also paying nontrivial attention weights to the adjectives. In the case of conjunctions of multiple noun phrases, the model consistently picks the conjunction as the head. The attention weights on simple verb phrases are peaked around the noun phrase instead of the verb.

As a result, nonterminal category labels add small amount of information compared to puterly edocentric representation, which does not make much difference in performance. In summary, RNNGs without access to nonterminal infromation during training are used to support the hypothesis that phrasal representation are largely endocentric. Even though the model is learning something similar to heads, the attention vectors are not completely peaked around a single component.