Hierarchy or Heuristic?

Examining Methods for Understanding Syntax in Recurrent Neural Networks

Isabel Papadimitriou

Stanford University
isabelvp@stanford.edu

Overview

How do LSTMs perform so well in language tasks? Do they process inputs in a similar way to the human language system? Recent research has focused on understanding if and how LSTMs encode linguistic syntax. We

- Replicate the experiments in McCoy et al [1]
- Demonstrate that the framework and metrics used to probe the LSTMs do not necessarily imply global syntactic awareness
- Provide further experiments to build upon McCoy et al's framework

Introduction

The success of LSTMs in NLP tasks, along with the opacity of end-to-end systems, lead to the question of what makes LSTMs so capable with linguistic data. The way in which this is investigated in much of the literature probes is through testing whether models successfully ignore subordinate clauses. If a model knows to fill in the singular "is" in the sentence "The boy petting the cats ___ happy", this means that the model somehow encodes "petting the cats" as separate from the overall sentence "the boy is happy" – a sign of syntactic understanding.

Through replicating and examining McCoy et al's results, we observed that models can perform well on identifying a subordinate clause as a separate unit **while in fact failing in higher-level grammatical awareness**. We propose more thorough criteria for defining hierarchical structural awareness: models should not only be able to recognize which clauses are separate, but also to embed relationships between them

Replication

We replicated the experiments in McCoy et al [1]. We constructed data from a simple grammar, and asked a model to form a question by fronting the main auxiliary verb:

(1) The girl in the red shirt will love these cats → Will the girl in the red shirts love these cats?

The training set consisted of examples where the main auxiliary was also the one closest to the subject noun, such as (1) above. The model was then tested on generalisation data where this was not the case as in:

(2) The girl who might visit will love these cats → Will the girl who might visit love these cats?

This tests whether LSTMs develop a more general, rule even when given linearly-explainable data

Replication Results

The performance of the LSTM is heavily contingent on what percentage of the RCs in the training data are simple 3-word RCs ("who can swim") vs more complicated ("who my cat likes"). This suggests that the accuracy when trained on mostly simple RCs is due to heuristics like counting.

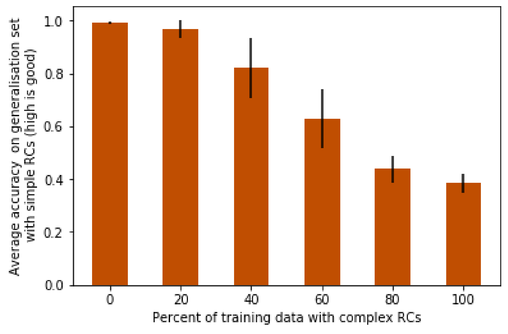

Figure 1: Accuracy of models trained on varying complexity of data, calculated by whether the models fronted the correct auxiliary for 1,000 inputs. Each average is across five different random training initializations.

Analysis of Replication Results

When we looked at the outputs of the models, we saw clear cues that even the best-performing models did not have a hierarchical understanding of grammar as we would define it. The model consistently performed a transformation of the following type:

(3) The girl **who might visit** will love these cats → Will the girl love these cats **who might visit**

This shows that the model has enough structural awareness to know that "who might visit" is a separate clause, and was not distracted by the other auxiliary "might". However, it does not have a more global awareness of the hierarchical relationship that the RC had with its parent clause.

Further Probing

To access the models internal state, we examined its probabilities for grammatical and ungrammatical sentences

Input	The cat that can swim will catch the fish
Correct output	Will the cat that can swim catch the fish?
Grammatical output, wrong	Will the cat catch the fish that can swim?
Ungrammatical output 1	*Will the cat catch that can swim the fish?
Ungrammatical output 2	*Will the cat catch the that can swim fish?

Table 1: An example entry in the scoring set

All of the models assigned a significantly higher probability to the correct output than to the two ungrammatical outputs ($p < 0.01$), even though they had never seen anything of the form of the correct output. This suggests that there is in fact a latent (though limited) grammatical representation

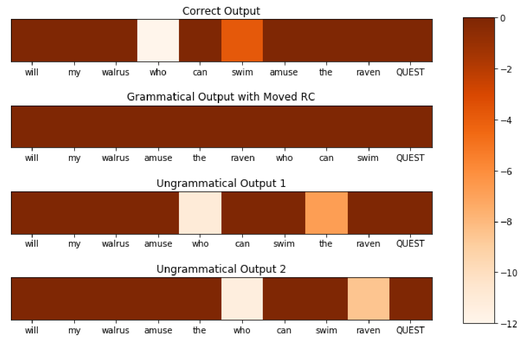

Figure 2: Log probability of each word in outputs, conditioned on the input "My walrus who can swim will amuse the ravens". The introduction of the RC ("who") is consistently surprising to the model except in the case of the RC on the object, suggesting that the model did not gain a global representation where the subject and object two instances of the same structure.

Conclusion

We argue that showing long-distance dependency awareness is not equivalent to showing hierarchical grammar awareness in a model. Using the experimental framework laid out by McCoy et al, we showed that though the models were often successfully ignoring the RC on the subject, they did not exhibit awareness of the RC's place in the phrase structure as a whole. However, we also take a step in the direction probing for more global hierarchical knowledge, and see that the models do in fact have a more global structural sensitivity. We provide a more robust theoretical and methodological framework to understand syntactic awareness in RNNs than only focusing on long-range dependencies.

References

[1] R. Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *CoRR*, abs/1802.09091, 2018.