Probing the Learning Capabilities of RNN Seq2seq Models
Zhengxiang Wang
Department of Linguistics, Stony Brook University

Learning formal languages has emerged as ideal proxy tasks for evaluating the expressive power and generalization capacity of neural networks in recent years (Grefenstette et al., 2015; Bhat-tamishra et al., 2020; Delétang et al., 2022), for example, from automata-theoretic perspectives (Merrill, 2019; Ayache et al., 2019; Peng et al., 2018).

The paper studies the capabilities of Recurrent-Neural-Network sequence to sequence (RNN seq2seq) models in learning four deterministic string-to-string transduction tasks: (A) identity; (B) reversal; (C) total reduplication; (D) input-specified reduplication. For a given string \( w \in \Sigma^* \), \( f_A(w) = w \), \( f_B(w) = \overline{w} \), \( f_C(w) = ww \), and \( f_D(w, @^n) = ww^n \), where \( \overline{w} \) denotes the reverse of \( w \) and @ a special instruction symbol whose number of occurrence (i.e., \( n \)) signals the number of copies to make for \( w \). These transductions are traditionally well studied under finite state transducers (FSTs) and attributed with varying complexity (Filiot and Reynier, 2016; Dolatian and Heinz, 2020; Rawski et al., 2023). We are interested in understanding how well the three major types of RNN seq2seq models (i.e., SRNN, GRU, LSTM), with and without attention, learn four transduction tasks and factors that affect the trained models’ generalization abilities.

For the experiments, we set the alphabet \( \Sigma \) to be the 26 lowercase English letters. We randomly sampled from \( \Sigma^* \) strings of lengths 1-30 as the input sequences, with the target sequences obtained by applying the four deterministic functions that represent the tasks. Models were trained on input sequences of lengths 6-15 and evaluated on both unseen in-distribution examples of same length range and unseen out-of-distribution examples of unseen lengths (for all functions) or unseen instruction symbol number (only for \( f_D \)). To make the results comparable across models and across tasks, the input sequences and the training and evaluation conditions were deliberately set identical for every model trained and evaluated.

![Figure 1: Full-sequence accuracy per input length on unseen test examples across the four tasks for the three types of RNN seq2seq models.](image)

Fig 1 shows a sketch of the main results on a per-input-length level. We find that RNN seq2seq models are only able to approximate a mapping that fits training or in-distribution data, but not to learn the underlying data generation functions. Attention helps significantly, but does not solve the out-of-distribution generalization limitation. RNN variants and task complexity also play a role in the results. Our results show that total reduplication is more complex than identity, which is more complex than reversal, for attention-less models to learn. We argue that this is best understood in terms of complexity hierarchies of formal languages.
as opposed to complexity hierarchies of string transductions, which treats reversal as a function strictly more complex than identity.

References


