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; Bhattamishra 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 $w \in \Sigma^*$, $f_A(w) = w$, $f_B(w) = \overleftarrow{w}$, $f_C(w) = ww$, and $f_D(w, @^n) = ww^n$, where $\overleftarrow{w}$ 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 Σ to be the 26 lowercase English letters. We randomly sampled from Σ^* 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.

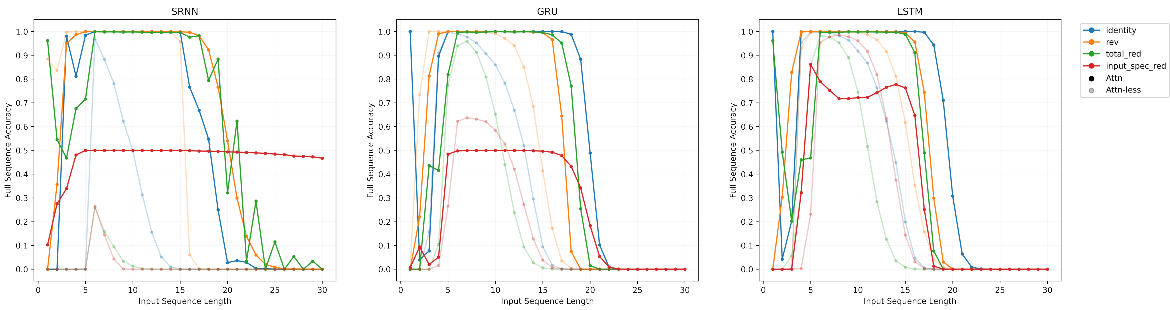

Figure 1: Full-sequence accuracy per input length on unseen test examples across the four tasks for the three types of RNN seq2seq models.

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

Stéphane Ayache, Rémi Eyraud, and Noé Goudian. Explaining black boxes on sequential data using weighted automata. In Olgierd Unold, Witold Dyrka, and Wojciech Wieczorek, editors, *Proceedings of The 14th International Conference on Grammatical Inference 2018*, volume 93 of *Proceedings of Machine Learning Research*, pages 81–103. PMLR, feb 2019. URL `https://proceedings.mlr.press/v93/ayache19a.html`.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the Ability and Limitations of Transformers to Recognize Formal Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.576. URL `https://aclanthology.org/2020.emnlp-main.576`.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the chomsky hierarchy, 2022. URL `https://arxiv.org/abs/2207.02098`.

Hossep Dolatian and Jeffrey Heinz. Computing and classifying reduplication with 2-way finite-state transducers. *Journal of Language Modelling*, 8(1):179–250, Sep. 2020. doi: 10.15398/jlm.v8i1.245. URL `https://jlm.ipipan.waw.pl/index.php/JLM/article/view/245`.

Emmanuel Filiot and Pierre-Alain Reynier. Transducers, logic and algebra for functions of finite words. *ACM SIGLOG News*, 3(3):4–19, aug 2016. doi: 10.1145/2984450.2984453. URL `https://doi.org/10.1145/2984450.2984453`.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/b9d487a30398d42ecff55c228ed5652b-Paper.pdf`.

William Merrill. Sequential neural networks as automata, 2019. URL `https://arxiv.org/abs/1906.01615`.

Hao Peng, Roy Schwartz, Sam Thomson, and Noah A. Smith. Rational recurrences. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1214, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1152. URL `https://aclanthology.org/D18-1152`.

Jonathan Rawski, Hossep Dolatian, Jeffrey Heinz, and Eric Raimy. Regular and polyregular theories of reduplication. *Glossa: a journal of general linguistics*, 8(1), 2023. doi: 10.16995/glossa.8885. URL `https://doi.org/10.16995/glossa.8885`.