Probing the Learning Capabilities of RNN Seq2seq Models

Zhengxiang Wang

zhengxiang.wang@stonybrook.edu

Department of Linguistics, Stony Brook University

Introduction

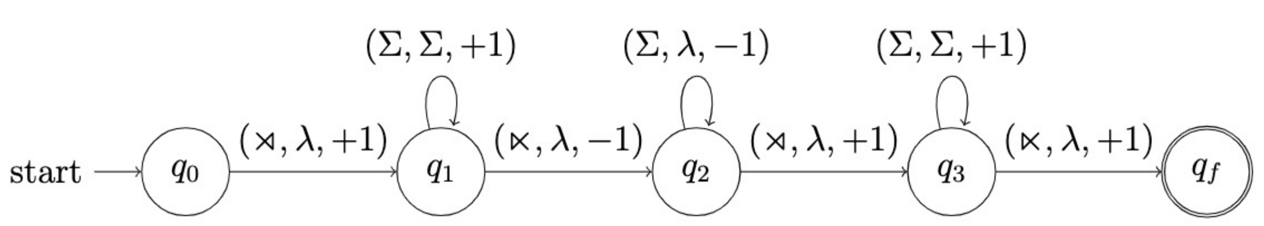
The paper studies the capabilities of Recurrent-Neural-Network sequence to sequence (RNN seq2seq) models in learning four deterministic transduction tasks of varying complexity and that can be described as learning alignments. Two main questions are:

- **Question 1**: how well do RNN seq2seq models generalize to unseen in-distribution and out-of-distribution examples?
- **Question 2**: What are the possible factors that impact trained models' generalization abilities?

Four Transduction Tasks

- **Identity** (*f*: w → w). Ex: abc → abc
- **Reversal** (*f*: w → w^R). Ex: abc → cba
- **Total Reduplication** (*f*: w → ww). Ex: abc → abcabc
- **Input-specified Reduplication** (*f*: w@^{n} → ww^n). Ex:
  - abc@ → abcabc
  - abc@@ → abcabcabc
  - abc@@@ → abcabcabcabc

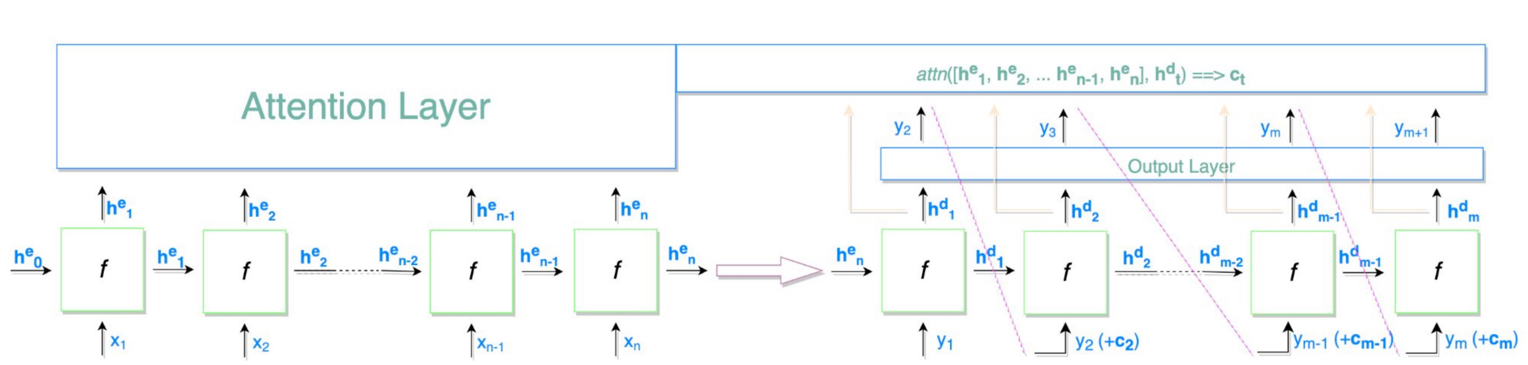
Increasing complexity under Finite State Transducer (FST)



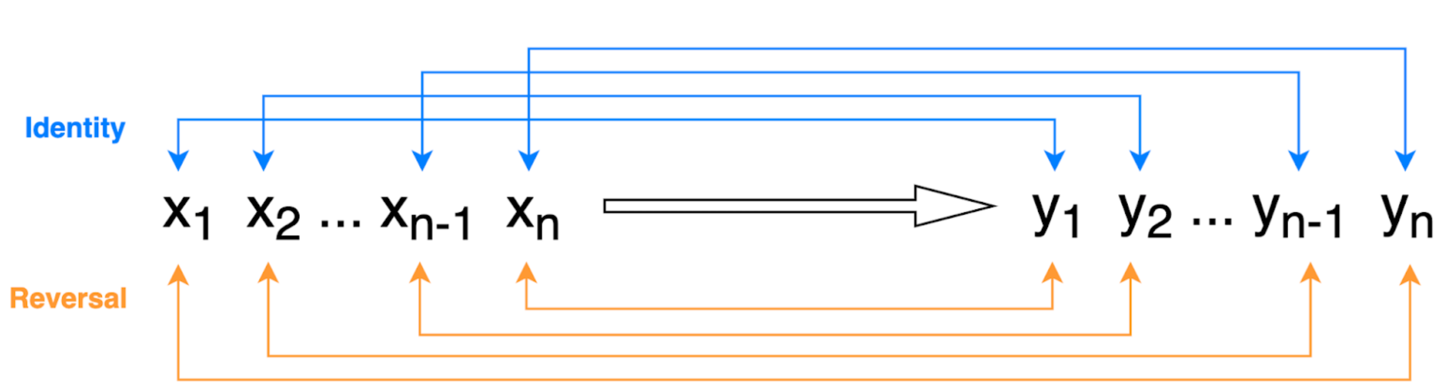
Ex: 2-way FST for modelling Total Reduplication

RNN Seq2seq Models

- **RNN general formula: ht = f(ht-1, xt)**
- **RNN seq2seq architecture**



- **Difference between FSTs and RNN seq2seq models:**
  - FSTs: read and write for every input symbol
  - RNN seq2seq: read everything before writing anything
- **Learning input-target alignments**



Experimental Setups

- **Data**
  - Identical input sequences from all datasets across all tasks
  - Input lengths 6-15 for train/dev/test, 1-5 & 16-30 for gen set. Four are disjoint.
  - Test set: in-distribution examples; gen set: out-of-distribution examples
- **Models**

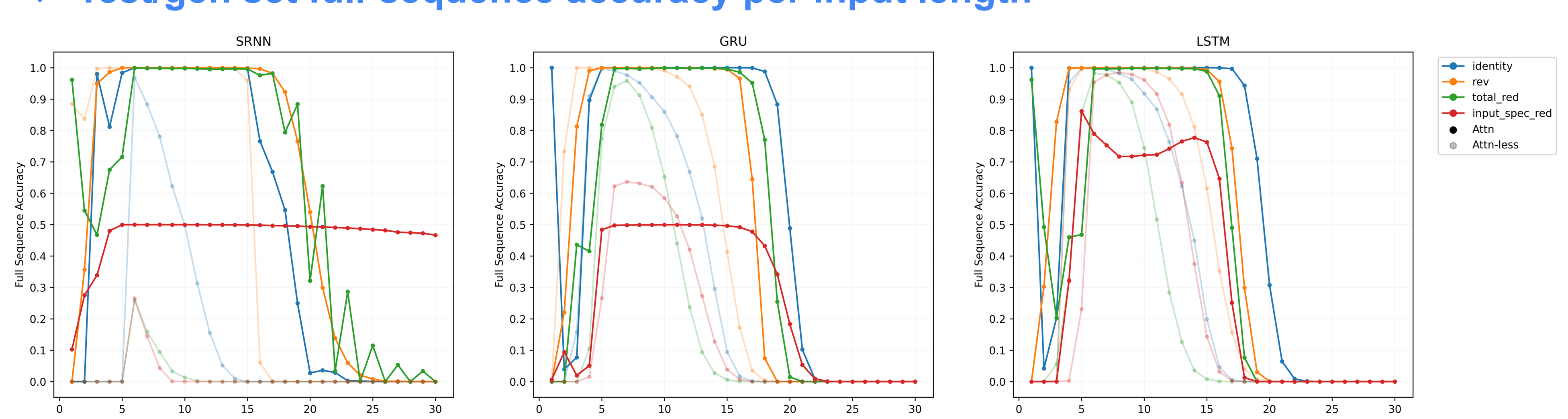
RNN	Attention	Param #	lr (Adam)	Hidden size	Embd size	Max Epoch #
SRNN	True	1,466,396	0.0005	512	128	500
SRNN	False	1,204,252				
GRU	True	3,305,500				
GRU	False	2,519,068				
LSTM	True	4,225,052				
LSTM	False	3,176,476				

Results

- **Aggregate full-sequence accuracy (%) with best results in bold**

		Attentional			Attention-less		
Task	**Dataset**	**SRNN**	**GRU**	**LSTM**	**SRNN**	**GRU**	**LSTM**
Identity	Train	100.00	100.00	100.00	69.74	98.26	100.00
	Test	99.97	**100.00**	**100.00**	42.82	70.46	**77.57**
	Gen	25.52	**37.41**	36.37	0.00	**10.41**	10.01
Rev	Train	100.00	100.00	100.00	100.00	100.00	100.00
	Test	**99.98**	99.87	99.88	**99.55**	88.46	92.85
	Gen	**40.14**	23.54	25.79	**23.89**	19.72	12.42
Total Red	Train	100.00	100.00	99.99	15.22	90.57	93.51
	Test	99.71	**99.77**	99.64	5.60	50.76	**55.17**
	Gen	**42.34**	23.23	20.31	0.00	4.39	**6.18**
Input-spec Red	Train	99.98	100.00	100.00	13.51	100.00	100.00
	Test	**99.94**	99.76	99.66	9.08	72.67	**81.15**
	Gen	**35.98**	10.58	18.32	0.00	4.55	**15.81**
Average	Train	100.00	100.00	100.00	49.62	97.21	98.37
	Test	**99.90**	99.85	99.79	39.27	70.59	**76.68**
	Gen	**35.99**	23.69	25.20	5.97	9.77	**11.11**

- **Test/gen set full-sequence accuracy per input length**



Discussion and Conclusion

- **Generalization abilities**: models tend to only learn a mapping that fits the training or in-distribution data, but not the underlying data generation functions
- **Attention**: helps significantly, but does not solve the out-of-distribution generalization problem
- **Task complexity**: Total reduplication > Identity > Reversal, attested only for attention-less models, but not input specified reduplication & attentional models

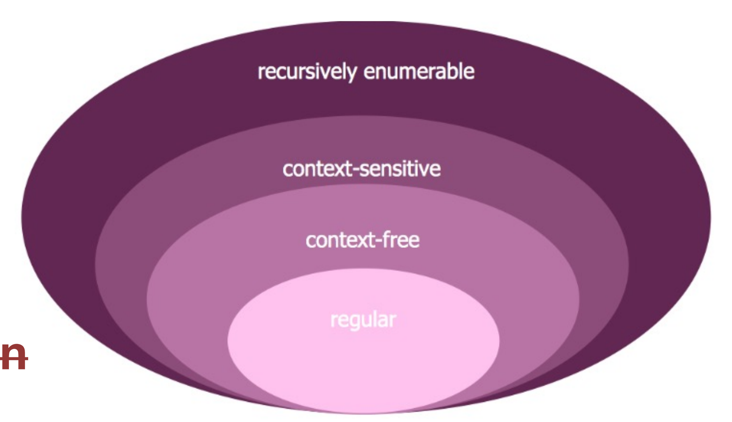
Complexity Hypothesis

Language recognition viewpoint

- Reversal → w#w^R (Context Free)
- Identity → w#w (Context Sensitive)
- Total Red → w#ww (Context Sensitive)
- Input-spec Red → w#ww^n (>= Context Sensitive)

Increasing complexity under Chomsky Hierarchy

The results are better understood from complexity hierarchy of **formal languages**, instead of that of ~~**string transduction**~~



Future Works

- **Experiments at a larger scale**
  - ✓ A wider range of training and evaluation input lengths for all tasks
  - ✓ Worth further testing whether the proposed task complexity hierarchies apply for input-specified reduplication and attentional models with more proper experimental setups
- **Models with other configurations**
  - ✓ Bidirectional encoder
  - ✓ Multi-layered RNNs in the encoder and decoder
  - ✓ Different variants of attention

Selected References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. **Neural machine translation by jointly learning to align and translate.** In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- Thang Luong, Hieu Pham, and Christopher D. Manning. **Effective approaches to attention- based neural machine translation**. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. **Sequence to sequence learning with neural networks**. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Wein- berger, editors, Advances in Neural Information Processing Systems, volume 27. Cur- ran Associates, Inc., 2014.
- Jonathan Rawski, Hossep Dolatian, Jeffrey Heinz, and Eric Raimy. **Regular and polyregular theories of reduplication**. Glossa: a journal of general linguistics, 8(1), 2023.