Poverty of the Stimulus with CHILDES: Supplementary Materials

Aditya Yedetore

CHILDES data pre-processing

All extraneous marks from the CHILDES NA-Eng corpora were removed.

training, validation, and test split 20% of the files in the CHILDES Treebank (Valian, Soderstrom, Brown, Suppes) were randomly selected and placed in the test set, for purposes of creating a test set for fine-tuning. 3% of the remaining files (by number of non-child utterances) were allocated to the validation and test sets each. The rest were allocated to the training set.

For the fine tuning data set, I

Hyper-parameters and further model details

LSTM For LSTMs I explored the following hyper-parameters for a total of 144 models.

1. layers: 2
2. hidden and embedding size: 200, 800
3. batch size: 20, 80
4. dropout rate: 0.0, 0.2, 0.4, 0.6
5. learning rate: 5.0, 10.0, 20.0
6. random seed: 1001, 1002, 1003 (...)

Each had it's own random seed, which ranged from 1001 to 1144.

The 5 LSTM models with the lowest perplexities after 40 training epochs are reported in Table 1.

nlayers	nhidden/embed	lr	batch_size	dropout	seed	test loss	test ppl
2	800	20	80	0.4	1135	3.25	25.70
2	800	10	20	0.4	1095	3.25	25.84
2	800	5	20	0.4	1093	3.26	25.98
2	800	10	80	0.4	1131	3.26	26.06
2	800	20	20	0.4	1097	3.26	26.13

seed	valid ppl	test ppl
1093	7.10	7.29
1094	6.43	6.57
1095	6.61	6.74
1096	6.89	6.97
1097	5.96	6.14
1131	7.22	7.30
1133	6.86	7.09
1134	6.64	6.76
1135	7.20	7.35

1 Fine Tuning

The models were fine tuned for seq2seq using the yes-no question data from the CHILDES Treebank. The LSTMs were fine tuned on the concatenation of the declarative form of a yes-no sentence and it's question form. The hidden states were reset after each declarative-question pair.

The test set was comprised completely of questions from files that were excluded from the pre-training. Both the test set and the validation set were chosen randomly, and were of similar sizes, so we can expect the performance over those sets to be comparable barring any over-fitting.

Performance on the held out questions

2 Results

The evaluation sets were generated using a CFG, with vocabulary of words commonly found in the training data.

seed	1000 test	1000 gen	10000 test	10000 gen
1093	0.898	0.026	0.8983	0.0313
1094	0.924	0.017	0.9221	0.0166
1095	0.979	0.053	0.9846	0.0395
1096	0.944	0.082	0.9455	0.0778
1097	0.978	0.049	0.983	0.0512
1131	0.997	0.030	0.9923	0.0313
1133	1	0.020	1	0.0162
1134	0.987	0.022	0.9887	0.0214
1135	0.994	0.079	0.9932	0.0860