Poverty of the Stimulus with CHILDES: Supplementary Materials

Aditya Yedetore

CHILDES data pre-processing

The data was taken from the CHILDES Eng-NA corpora, and extraneous transcription marks were removed. Utterances by the target children were also removed.

Pre-Training: train valid test split We randomly selected 20% of the files in the CHILDES Treebank (Valian, Soderstrom, Brown, Suppes) for the test set. We randomly allocated files containing 6% of the remaining utterances to the validation and test sets. The remainder (94%) was the training set.

To maintain local continuity, the training data was not shuffled.

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: 3 per parameter combination, unique for each LSTM

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

Table 1: Model Parameters

1 Fine Tuning

The LSTM models were fine tuned for seq2seq using the yes-no question data from the CHILDES Treebank. Each item in the training data was the concatenation of the declarative form of a yes-no sentence and the question itself. The test set was comprised of questions from files that were excluded from the pre-training.

The validation set was chosen randomly to match the size of the test set. The overall order of the decl-quest pairs were shuffled for the all the data.

While training, the hidden states of the LSTM were reset after each decl-quest pair. Back-propagation was through both the declarative and the question forms of the sentence.

Table 2 shows the performance of on the held out questions.

seed	valid ppl	test ppl	test acc.
1093	7.10	7.29	99.99
1094	6.43	6.57	99.99
1095	6.61	6.74	99.99
1096	6.89	6.97	99.99
1097	5.96	6.14	99.99
1131	7.22	7.30	99.99
1133	6.86	7.09	99.99
1134	6.64	6.76	99.99
1135	7.20	7.35	99.99

Table 2: Performance of top 9 LSTMs on held out questions. test acc is the percent accuracy of the first word in the question form of the declarative

2 Evaluation sets

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

`gen set CFG`

```
S -> NP_M_S VP_M_S | NP_M_P VP_M_P

NP_M_S -> Det N_S RC_S
NP_M_P -> Det N_P RC_P

NP_O -> Det N_S | Det N_P | Det N_S Prep Det N
NP_O -> Det N_P Prep Det N | Det N_S RC_S | Det N_P RC_P

N -> N_S | N_P

VP_M_S -> 'MAIN-AUX' Aux_S V_intrans
VP_M_S -> 'MAIN-AUX' Aux_S V_trans NP_O

VP_M_P -> 'MAIN-AUX' Aux_P V_intrans
VP_M_P -> 'MAIN-AUX' Aux_P V_trans NP_O

RC_S -> Rel Aux_S V_intrans | Rel Det N_S Aux_S V_trans
RC_S -> Rel Det N_P Aux_P V_trans | Rel Aux_S V_trans Det N
RC_P -> Rel Aux_P V_intrans | Rel Det N_S Aux_S V_trans
RC_P -> Rel Det N_P Aux_P V_trans | Rel Aux_P V_trans Det N

Det -> 'the' | 'some' | 'my' | 'your'
N_S -> 'dog' | 'girl' | 'boy' | 'animal'
N_P -> 'dogs' | 'girls' | 'boys' | 'animals'
V_intrans -> 'play' | 'sit' | 'fall' | 'talk' | 'sleep'
V_trans -> 'like' | 'want' | 'see' | 'eat'
Aux_P -> 'do' | 'did' | 'can' | 'would' | 'shall'
Aux_S -> 'does' | 'did' | 'can' | 'would' | 'shall'
Prep -> 'around' | 'with' | 'near'
Rel -> 'who' | 'that'


test set CFG.

S -> NP_M_S VP_M_S | NP_M_P VP_M_P

NP_M_S -> Det N_S | Det N_S Prep Det N
NP_M_P -> Det N_P | Det N_P Prep Det N

NP_O -> Det N_S | Det N_P | Det N_S Prep Det N
NP_O -> Det N_P Prep Det N | Det N_S RC_S | Det N_P RC_P

N -> N_S | N_P

VP_M_S -> 'MAIN-AUX' Aux_S V_intrans
```

```
VP_M_S -> 'MAIN-AUX' Aux_S V_trans NP_O

VP_M_P -> 'MAIN-AUX' Aux_P V_intrans
VP_M_P -> 'MAIN-AUX' Aux_P V_trans NP_O

RC_S -> Rel Aux_S V_intrans | Rel Det N_S Aux_S V_trans
RC_S -> Rel Det N_P Aux_P V_trans | Rel Aux_S V_trans Det N
RC_P -> Rel Aux_P V_intrans | Rel Det N_S Aux_S V_trans
RC_P -> Rel Det N_P Aux_P V_trans | Rel Aux_P V_trans Det N

Det -> 'the' | 'some' | 'my' | 'your'
N_S -> 'dog' | 'girl' | 'boy' | 'animal'
N_P -> 'dogs' | 'girls' | 'boys' | 'animals'
V_intrans -> 'play' | 'sit' | 'fall' | 'talk' | 'sleep'
V_trans -> 'like' | 'want' | 'see' | 'eat'
Aux_P -> 'do' | 'did' | 'can' | 'would' | 'shall'
Aux_S -> 'does' | 'did' | 'can' | 'would' | 'shall'
Prep -> 'around' | 'with' | 'near'
Rel -> 'who' | 'that'
```

3 Results

First word performance on 10000 randomly selected items from the CFG generated evaluation sets are in Table-3.

seed	test	gen
1093	89.8	3.13
1094	92.2	1.66
1095	98.4	3.95
1096	94.5	7.78
1097	98.3	5.12
1131	99.2	3.13
1133	100.0	1.62
1134	98.8	2.14
1135	99.3	8.60

Table 3: Performance of top 9 LSTMs. The proportion of the 10000 randomly selected gen set items for which the fine-tuned LSTM predicted the correct first word in the question form of the sentence after seeing the declarative form.