Poverty of the Stimulus with CHILDES: Parameter search 1

I did this 'pre-search' mainly to convince myself that the values I was going to be using for the grid search were reasonable. My main concern was whether the number of layers would change the performance of the model.

I didn't do any rigorous testing, since I just wanted to get a feel of the relative importance of the parameters, and how they interacted.

1 Hyperparameter pre-search

In Table 1 we can see that as the number of layers changes, the resulting perplexity doesn't change too dramatically.

I also wanted to check if the number of layers would interact with the hidden size. In Table 2 the differences between the test set perplexities of models with different hidden size is pretty much the same no matter the number of layers. For example, the difference in performance between a model with 2 layers and 100 hidden units and one with 2 layers and 200 hidden units is close to the difference in performance between a model with 1 layer and 100 hidden units and a model with 1 layer and 200 hidden units.

Note that I didn't vary the embedding size with the hidden size. That was simply a mistake on my part.

I did do further tests. The first of those was with models with 1600 hidden units. I wanted to check if this would improve the model over 800 hidden units. It turns out it did not, but that may have been due to the fact that the embedding size was still 200. Then I co-varied hidden size and batch size, and

ppl	layers	hidden/ embedding size	dropout rate	learning rate
42.08	1	200	0.0	20
39.97	2	200	0.2	20
40.01	3	200	0.2	20
40.35	4	200	0.2	20
41.26	5	200	0.2	20

Table 1: varying layers

ppl	layers	hidden size	embedding size	dropout rate	learning rate
47.27	1	100	200	0.0	20
42.08	1	200	200	0.0	20
36.84	1	400	200	0.0	20
35.32	1	800	200	0.0	20
46.92	2	100	200	0.2	20
39.97	2	200	200	0.2	20
36.51	2	400	200	0.2	20
35.38	2	800	200	0.2	20
46.79	3	100	200	0.2	20
40.01	3	200	200	0.2	20
36.20	3	400	200	0.2	20
35.63	3	800	200	0.2	20

Table 2: varying layers and hidden size

then dropout and learning rate. These results weren't too interesting, so I won't take the time to report them in detail here.