A NEURAL NETWORK MODEL OF ADAPTATION IN READING

Marten van Schijndel (Johns Hopkins University) & Tal Linzen (Johns Hopkins University)
vanksy@jhu.edu

Humans rapidly adapt their lexical and syntactic expectations to match the statistics of the current linguistic context (e.g., Fine et al., 2013). Computational word prediction models (language models) that adapt to the current context make more accurate predictions (e.g., Kuhn & de Mori, 1990). Combining these two research traditions, we propose a simple adaptive neural language model, and show that adaptation improves our predictions of human reading times.

Our baseline model is a long short-term memory (LSTM) language model trained on 90 million words of English Wikipedia articles. For adaptation, at the end of each new sentence, we update the parameters of the model based on its errors in predicting that single sentence. We tested the model on the Natural Stories Corpus (Futrell et al., 2017), which has 10 narratives with self-paced reading times from 181 English speakers.

Linguistic accuracy: We first measured how well the model predicts upcoming words. We use the standard measure of perplexity; this measure is lower when the model assigns higher probabilities to the words that in fact occurred. Adaptation over the test corpus dramatically improved test perplexity compared to a non-adaptive version of the model (86.99 vs 141.49).

Fit to reading times: We next tested whether our adaptive language model is a better model of human expectations than a non-adaptive one. We adapted the model to each story independently[1] and used its surprisal at each word to predict the corresponding self-paced reading times. Adaptive surprisal was predictive of reading times over a linear mixed model baseline containing non-adaptive surprisal[2] ($p < 0.001$), and its presence caused non-adaptive surprisal to no longer be a significant predictor (Table 1). This result indicates that this model more closely represents human expectations than non-adaptive language models.

Does the model adapt its syntax? To test whether the model adapts its lexical predictions, its syntactic predictions, or both, we generated 200 pairs of dative sentences, each with a prepositional object (PO) variant (*The boy threw the ball to the dog*) and a double object (DO) variant (*The boy threw the dog the ball*). We shuffled 100 PO items into 1000 filler items from Wikipedia and adapted the model to these 1100 sentences. We then froze the weights of the adapted model and tested its predictions for two types of sentences: the PO counterparts of the DO sentences used during adaptation, and 100 sentences that had the same syntax as those used during adaptation (DO) but shared no content words with them. We then repeated the experiment with the role of DO and PO reversed. This process was repeated 10 times each for PO and DO with different critical items and filler sentences. We found that the model adapted more strongly to vocabulary choice than syntax but was sensitive to both (Figure 1).

Overall, adaptation greatly improved the language model's accuracy and RT predictions. This improvement was due not only to lexical but also syntactic adaptation.

	$\hat{\beta}$	$\hat{\sigma}$	t
Sentence position	0.29	0.53	0.5
Word length	6.42	1.00	6.4
Surprisal	-0.89	0.68	-1.3
Adaptive surprisal	8.77	0.68	13.0

Table 1: Fixed effect self-paced reading regression coefficients.

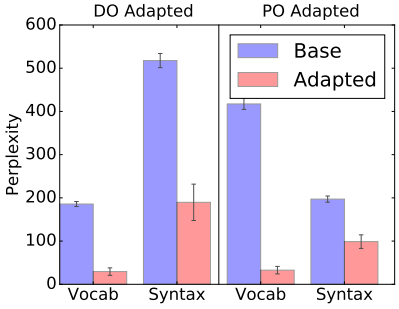

Figure 1: Lexical vs syntactic adaptation. (Note that the base model prefers PO.)

[1] After each story, the model reverts to the initial language model and must restart adaptation on the next story.

[2] The baseline is as follows: RT $\sim$ word_length + sentence_position + non-adaptive_surprisal + (1|word) + (1 + word_length + sentence_position + non-adaptive_surprisal + adaptive_surprisal |subject)