Comparison of Structural and Neural Language Models as Surprisal Estimators

Byung-Doh Oh (oh.531@osu.edu), Christian Clark, and William Schuler (The Ohio State University)

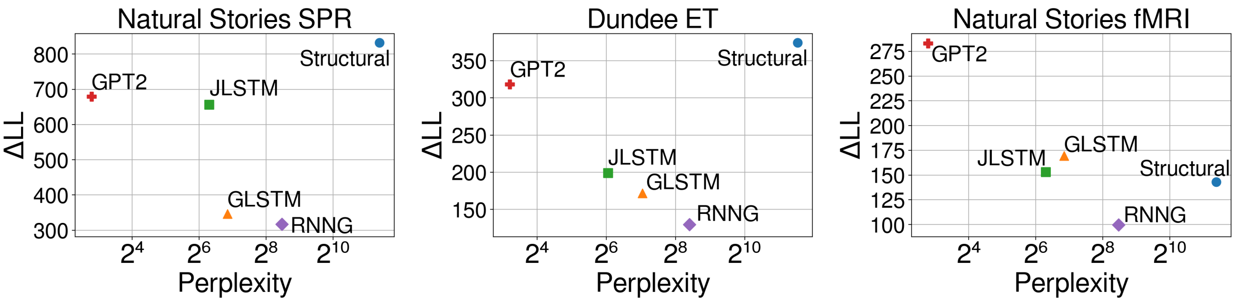
A recent trend in computational psycholinguistics has been to use large pretrained neural language models (NLMs) to generate surprisal estimates [2, 5, 8]. Although there is some evidence that NLM surprisal is predictive of human behavioral responses [2, 5], there has been very little work (see [4]) comparing its predictive power to that of surprisal from structural parser-based processing models. In this study, we conduct regression analyses on three different datasets and demonstrate that surprisal estimates from a sentence processing model informed by syntactic and morphological structure contribute to substantially better fits than those from widely-used pretrained NLMs [3, 6, 4, 9] on self-paced reading and eye-tracking data, but not on fMRI data.

To this end, we use a non-recurrent neural extension of a left-corner parser [12] that has a character-based model for estimating word generation probabilities at preterminal nodes. The proposed model defines a process of generating words w_t from underlying lemmas x_t and morphological rules r_t, which allows the processing model to capture the predictability of given word forms in a fine-grained manner.

In order to evaluate the quality of surprisal estimates from the sentence processing model informed by syntactic and morphological structure (*Structural Model*) as well as those from widely used pretrained NLMs (*GLSTM* [3], *JLSTM* [6], *RNNG* [4], *GPT2* [9]), linear mixed-effects regression analyses were conducted to evaluate model fit in terms of log-likelihood improvement on top of a baseline regression model. To this end, surprisal predictors for the Natural Stories self-paced reading corpus [1] and the Dundee eye-tracking corpus [7] were calculated from the structural model and the pretrained NLMs. The baseline predictors included were word length, word position, and unigram surprisal for Natural Stories, and word length, word position, and saccade length for Dundee. All predictors were z-transformed prior to fitting, and all surprisal predictors were spilled over by one position. All regression models included by-subject random slopes for all fixed effects. The results show that on both corpora, surprisal from the structural model made the biggest contribution to model fit in comparison to surprisal from the pretrained NLMs (Figures 1a and 1b, difference between structural model and other models significant with $p < 0.001$ by a permutation test). This finding, despite the fact that the pretrained NLMs were trained on much larger datasets (Table 1) and also show lower perplexities on test data, suggests that the structural model may provide a more human-like account of processing difficulty and may suggest a larger role of morphology, phonotactics, and orthographic complexity than was previously thought.

Additionally, to examine whether a similar tendency is observed in brain responses, we analyzed the time series of blood oxygenation level-dependent (BOLD) signals identified using functional magnetic resonance imaging (fMRI) with continuous-time deconvolutional regression (CDR; [11]). For this experiment, we used the fMRI data of the language network used in [10], which were collected from 78 subjects that listened to a recorded version of the Natural Stories Corpus. Similarly, a baseline CDR model and a series of CDR models that include each surprisal estimate were fitted to BOLD measures. The baseline predictors included were the index of current fMRI sample, unigram surprisal, and the deconvolutional intercept. Subsequently, the contribution of each surprisal estimate was examined by calculating the improvement in regression log-likelihood. The results show that in contrast to self-paced reading and eye-tracking data, surprisal from *GPT2* made the biggest contribution to regression model fit (Figure 1c, difference between *GPT2* and other models significant with $p < 0.001$ by a permutation test, other comparisons not significant).

Taken together, these results suggest that sentence processing is not purely driven by accurate next-word prediction that large NLMs are capable of. In addition, the differential contribution of surprisal from the structural model suggests that latency-based measures and blood oxygenation levels may capture different aspects of processing difficulty.



(a) Baseline log-likelihood: -17485.2 (b) Baseline log-likelihood: -60807.5 (c) Baseline log-likelihood: -269825.1

Figure 1: Perplexity measures from each model, and improvements in regression model log-likelihood from including surprisal estimates from each model. The perplexity of the structural model and the RNNG model is higher partly because they are optimized to predict a joint distribution over words and parse trees.

Model	Training corpus
GPT2 [9]	>1B tokens
JLSTM [6]	∼800M tokens
GLSTM [3]	∼80M tokens
RNNG [4]	∼950k tokens
Structural Model	∼950k tokens

Table 1: The training corpus size for each model.

References

[1] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. The Natural Stories Corpus. In *LREC*, pages 76–82, 2018.

[2] A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *CMCL*, pages 10–18, 2018.

[3] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. Colorless green recurrent networks dream hierarchically. In *NAACL-HLT*, pages 1195–1205, 2018.

[4] J. Hale, C. Dyer, A. Kuncoro, and J. Brennan. Finding syntax in human encephalography with beam search. In *ACL*, pages 2727–2736, 2018.

[5] Y. Hao, S. Mendelsohn, R. Sterneck, R. Martinez, and R. Frank. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *CMCL*, pages 75–86, 2020.

[6] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *CoRR*, 2016.

[7] A. Kennedy, R. Hill, and J. Pynte. The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

[8] G. Prasad, M. van Schijndel, and T. Linzen. Using priming to uncover the organization of syntactic representations in neural language models. In *CoNLL*, pages 66–76, 2019.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *ArXiv*, 2019.

[10] C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, and E. Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138, 2019.

[11] C. Shain and W. Schuler. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *PsyArXiv*, 2019.

[12] M. van Schijndel, A. Exley, and W. Schuler. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540, 2013.