On the Bigger-is-Worse Nature of Pre-Trained Language Model Surprisal

Byung-Doh Oh William Schuler

The Ohio State University

oh.531@osu.edu
arxiv.org/abs/2212.12131

PAPER

Introduction

- Expectation-based theories of sentence processing [4, 7] posit that processing difficulty is driven by predictability, but make no strong claims about the probability distribution
- As such, surprisal estimates calculated from language models (LMs) with different architectures have been evaluated on their ability to predict measures of comprehension difficulty [11, 5]
- Earlier work [3, 12] observed ***negative*** relationship between LM perplexity and fit to reading times
- Pre-trained GPT-2 LM [9] shows ***positive*** relationship between LM perplexity and fit to reading times [8]

Replication Study: Evaluation on Human Reading Times

- Regression fit to reading times of the Natural Stories SPR corpus [2] and the Dundee eye-tracking corpus [6]
- Baseline predictors: word length, word position within sentence (both SPR and ET), saccade length, whether previous word was fixated (ET only)
- Surprisal predictors calculated using variants of GPT-2 [9], GPT-Neo [1], OPT [13] with different sizes

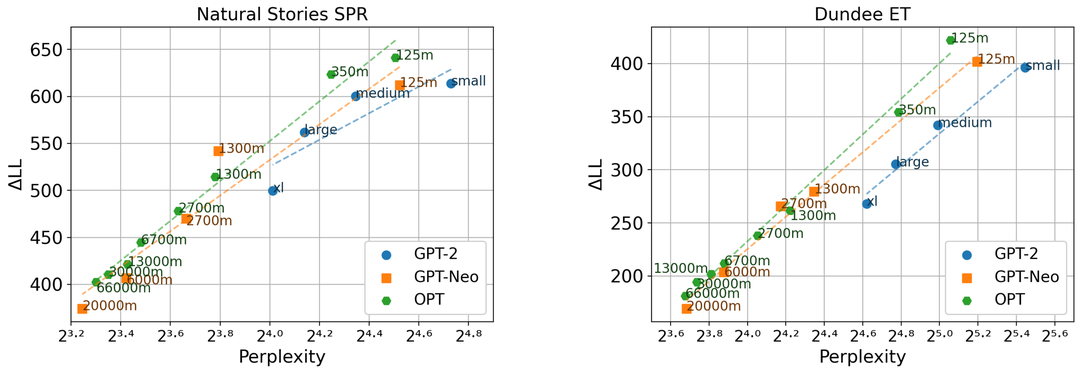

Figure 1: Improvements in regression model log-likelihood from including each surprisal predictor on the exploratory partition. The dotted least-squares regression lines had a slope significantly greater than 0 at $p < 0.05$ level according to a one-tailed t-test, with the exception of the regression line for GPT-2 on Natural Stories ($p = 0.07$).

- Results show a ***strictly monotonic***, ***positive*** log-linear relationship between perplexity and fit to reading times

Post-hoc Residual Analysis: Linguistic Phenomena Underlying the Trend

- Each data point in both corpora was associated with various word- and sentence-level properties [10]
- For every corpus-LM combination, subsets with the largest differences in SE between models were identified
- Data points in each subset were further separated according to whether the regression model underpredicted or overpredicted the target reading times

[1] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. *Zenodo*, 2021.
[2] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*, 55:63–77, 2021.
[3] A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proc. CMCL*, pages 10–18, 2018.
[4] J. Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proc. NAACL*, pages 1–8, 2001.
[5] J. Hale, C. Dyer, A. Kuncoro, and J. Brennan. Finding syntax in human encephalography with beam search. In *Proc. ACL*, pages 2727–2736, 2018.
[6] A. Kennedy, R. Hill, and J. Pynte. The Dundee Corpus. In *Proc. ECEM*, 2003.
[7] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.
[8] B.-D. Oh, C. Clark, and W. Schuler. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5(777963), 2022.
[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.
[10] C. Shain, M. van Schijndel, and W. Schuler. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*, 2018.
[11] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.
[12] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. P. Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proc. CogSci*, pages 1707–1713, 2020.
[13] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open pre-trained Transformer language models. *arXiv preprint*, 2022.

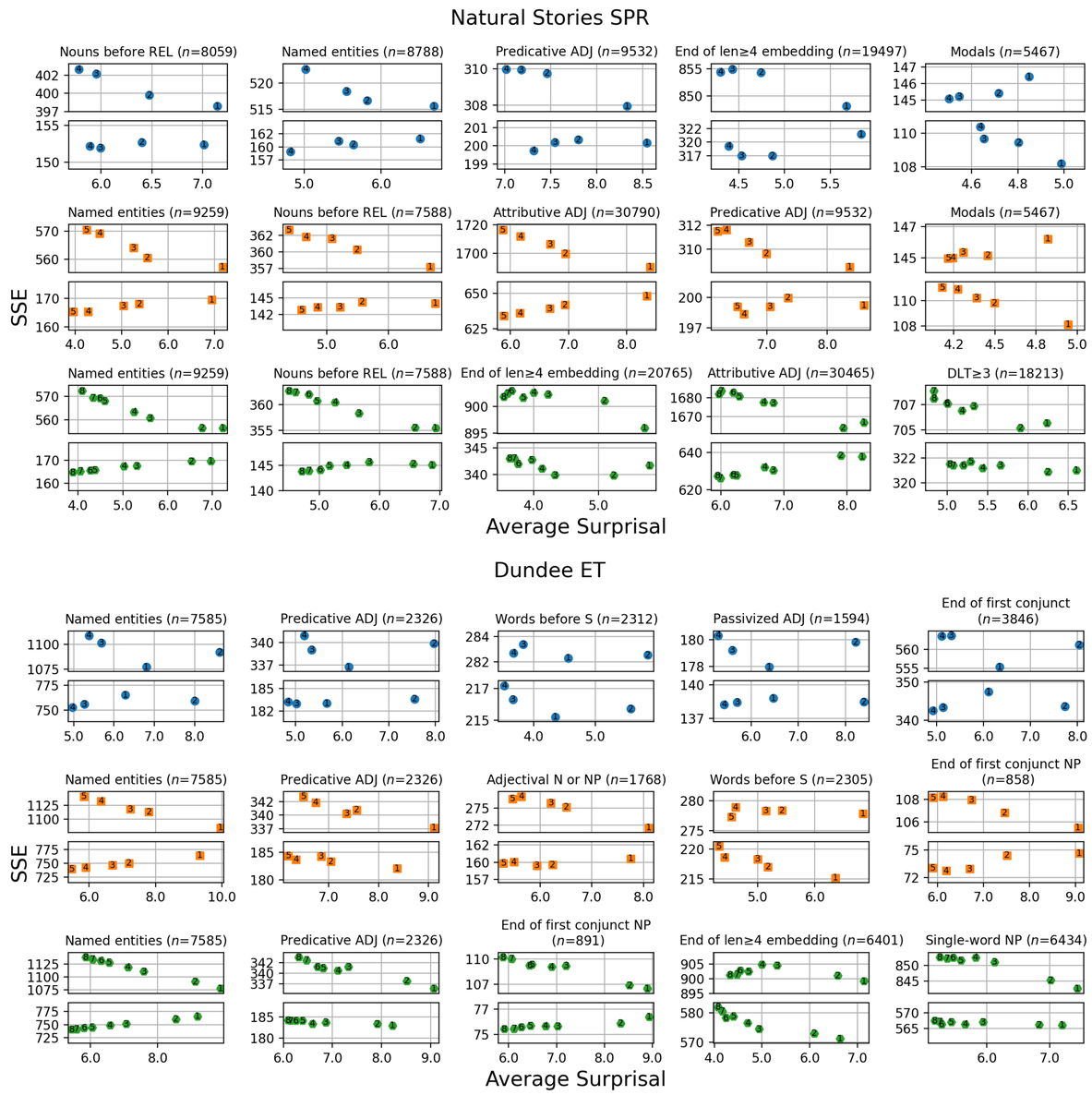

Figure 2: Average surprisal from each GPT-2, GPT-Neo, and OPT model variant (top, middle, and bottom rows respectively), and sum of squared errors of regression models that include each surprisal estimate on the top five subsets of Natural Stories and Dundee. The top and bottom subplots of each row represent values from underpredicted and overpredicted data points respectively.

Conclusion

Larger LMs achieve poorer fit to human reading times because they assign lower surprisal to open-class words, which may be accurately predicted by extensive domain knowledge gleaned from large sets of training examples