Comparison of Structural and Neural Language Models as Surprisal Estimators

Byung-Doh Oh   Christian Clark   William Schuler

The Ohio State University

March 4, 2021
34th Annual CUNY Conference

Introduction

Evaluation of surprisal estimates from large neural language models (NLMs) (Goodkind & Bicknell, 2018; Hao et al., 2020; Prasad et al., 2019)

Very little work (e.g. Hale et al., 2018) comparing their predictive power to that of surprisal from structural parser-based processing models

This work: Comparison of predictive power of surprisal estimates from different models on three different datasets (SPR, ET, fMRI)

Extension to our Structural Processing Model

NP

$horse_1$

horses

Word generation probability, P(horses | $horse_1$ NP)

- To a lemma x_t, apply a morphological rule r_t to generate word w_t
- Lemma x_t: result of applying GCG lemmatization rules (e.g. `horse`)
- Morphological rule r_t: inverse of GCG lemmatization rules (e.g. `attach-s`)

$$\mathsf{P}(w_t \mid p_t) = \sum_{x_t, r_t} \mathsf{P}(x_t \mid p_t) \cdot \mathsf{P}(r_t \mid p_t\ x_t) \cdot \mathsf{P}(w_t \mid p_t\ x_t\ r_t)$$

- Two character-based RNN sub-models for estimating $\mathsf{P}(x_t \mid p_t)$ and $\mathsf{P}(r_t \mid p_t\ x_t)$

Psycholinguistic Evaluation

Comparison of our surprisal estimates against those from widely-used pretrained language models

- GLSTM (Gulordava et al., 2018)
- JLSTM (Jozefowicz et al., 2016)
- RNNG (Hale et al., 2018)
- GPT2 (Radford et al., 2019)

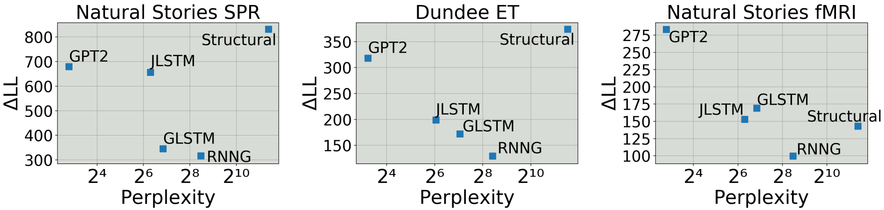
Evaluation metric: Δlog-likelihood (Goodkind & Bicknell, 2018; Hao et al., 2020)

- Improvement in log-likelihood due to including a surprisal predictor

Evaluation on

- Natural Stories self-paced reading (Futrell et al., 2018)
- Dundee eye-tracking (Kennedy et al., 2003)
- Natural Stories fMRI (Shain et al., 2019)

Results



(a) Baseline LL: -17485.2 (b) Baseline LL: -60807.5 (c) Baseline LL: -269825.1

- Our structural model may provide a more human-like account of processing difficulty
- May suggest a larger role of morphology, phonotactics, and orthographic complexity
- Latency-based measures and blood oxygenation levels may capture different aspects of processing difficulty

Conclusion

- An incremental parser that incorporates information about propositional content and syntactic categories into a probability model
- Independent contribution of propositional content and syntactic category information in predicting reading times

- A character-based model that can be used to estimate word generation probabilities in a parser-based model
- Substantially better fits to self-paced reading and eye-tracking data compared to surprisal from widely-used NLMs

Thank you for listening!

Source code:
`https://github.com/modelblocks/modelblocks-release`

References I

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The Natural Stories Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 76–82.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 10–18.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205.

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2727–2736.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *Proceedings of the 10th Workshop on Cognitive Modeling and Computational Linguistics*, 75–86.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv*.

References II

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. *Proceedings of the 12th European conference on eye movement.*

Prasad, G., van Schijndel, M., & Linzen, T. (2019). Using priming to uncover the organization of syntactic representations in neural language models. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 66–76.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *ArXiv.*

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