Benchmarking Neural Networks as Models of Language Processing

Ethan Wilcox[1], Jon Gauthier[2], Jennifer Hu[2], Peng Qian[2] and Roger Levy[2]
Contact: wilcoxeg@g.harvard.edu

Neural network language models (LMs) have achieved state-of-the-art results on many Natural Language Processing tasks, and have been proposed as models of human language processing since as far back as Elman (1990). However, evaluation metrics for language models as models of human sentence processing remain under-developed. Perplexity is the traditional broad-coverage metric for neural network evaluation (see computational supplement); however, it does not capture language models' knowledge of fine-grained linguistic phenomena and is not directly comparable to human behavior. We use two psycholinguistic benchmarking methods to evaluate neural networks vis-a-vis human behavior directly: The first is a targeted assessment of syntactic generalization in neural LMs by treating them like subjects in a psycholinguistic experiment. The second probes the link between model-derived surprisal and human reading times (Demberg and Keller, 2008), asking how well a language model predicts gaze duration in online reading. For each method, we train between 21-26 individual models across four architectures (see computational supplement) on four datasets of varying sizes derived from the BLLIP corpus (Charniak et. al, 2000). We also evaluate a series of off-the-shelf models, pre-trained on larger amounts of data. Our results indicate that model architecture contributes more than training data size to humanlike performance. However, different architectures fare differently on each benchmark, suggesting that different modeling approaches may be required to capture different aspects of human language processing.

Psycholinguistic Assessment of Grammar: We adopt the psycholinguistic approach to neural network evaluation (e.g. Marvin & Linzen 2018; Futrell et. al., 2018), which treats models like subjects in a psycholinguistic experiment, asking whether models assign higher probability to critical regions of grammatical minimal-pair sentence variants. For example, if models have learned proper subject/verb number agreement, given the prefix "*The keys to the cabinet…"* they should assign a higher probability to "*are"* than to "*is"*. We divide assessment into 34 test suites, each of which comprises between 20-30 items and probes one particular grammatical phenomena. Test suites are further grouped into 6 circuits which probe related phenomena. Model performance is reported as a Syntactic Generalization (SG) score which is the proportion of times the model prefers the grammatical variant. The results from our controlled experiments can be seen in Figure 1. We find no simple relationship between SG score and perplexity. In Figure 2, we report the SG score by model architecture and by training data size. These results indicate that model architecture, not training data size, has the largest impact on SG score, with the structurally supervised models performing best.

Predictive Power of Reading Times: There is a tight linear relationship between the time it takes to read a word and its surprisal, or negative log probability (Smith and Levy, 2013) derived from an LM. But which models are the best predictors of human online processing behavior? We derive a model's predictive power by computing its delta log likelihood (ΔLogLik; see supplementary materials for description) on three corpora of human reading times, two self-paced reading and one eye-tracking. The results from these experiments can be seen in Figure 3. We find a monotonic relationship between ΔLogLik and perplexity, which had previously been established for simpler language models (Goodkind and Bicknell, 2018). We also find an effect of architecture, but it is not the same as the one found in our psycholinguistic grammar assessment. Here, n-gram models overperform based on their perplexity, and even out-perform some neural network models on the eye-tracking dataset.

[1] Department of Linguistics, Harvard
[2] Department of Brain and Cognitive Science, MIT

SG Score vs. Perplexity

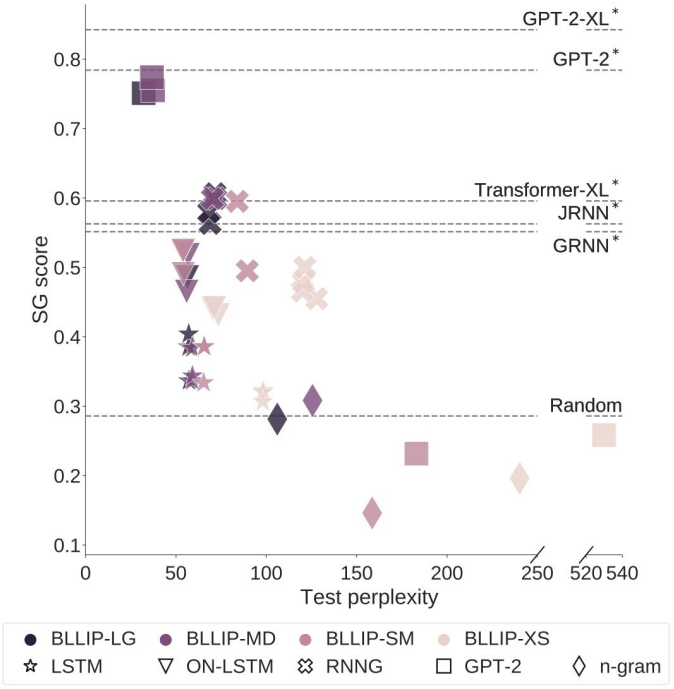

Fig 1: Syntactic Generalization Score vs. Perplexity. Perplexity is not given for off-the-shelf models, as they are evaluated on different test corpora.

SG Score as a function of Model Type and Training-Data size.

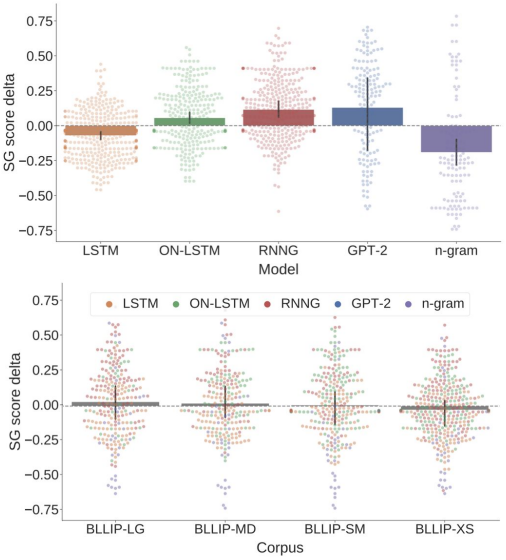

Fig 2: SG score broken down by model architecture (top row) and training corpus size (bottom row).

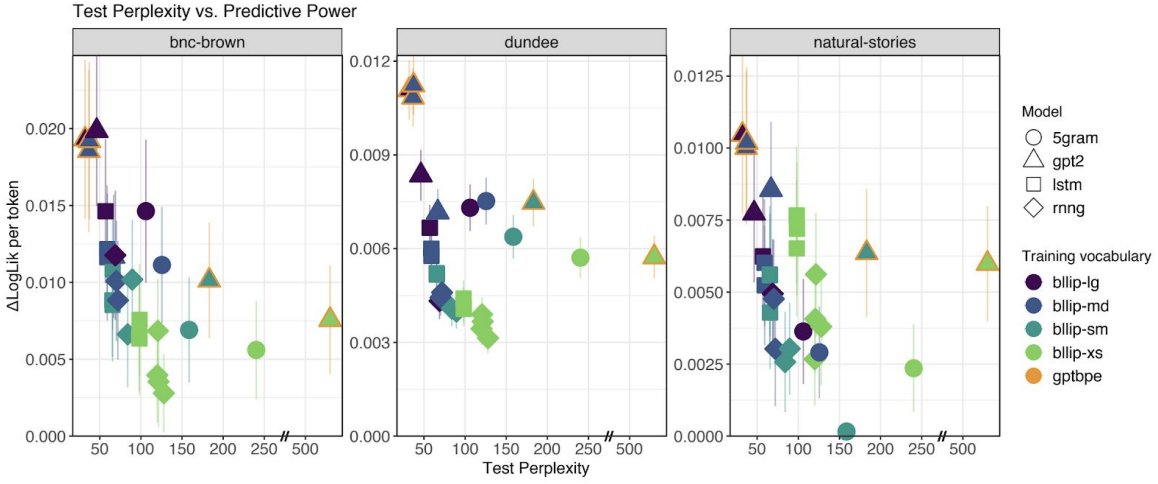

Fig 3: Predictive Power (ΔLogLik) vs. Perplexity. Higher ΔLogLik is a better fit to human data.

References:

Charniak et. al. BLLIP, 2000 • Demberg and Keller, Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity 2008 • Dyer et. al. Recurrent Neural Network Grammars, 2016. • Elman, Finding Structure in Time, 1990. • Futrell et. al., Neural language models as psycholinguistic subjects: Representations of syntactic state, (2019). • Goodkind, & Bicknell, Predictive power of word surprisal for reading times is a linear function of language model quality, 2018.• Marvin, R., & Linzen, T, Targeted syntactic evaluation of language models, 2018. • Smith, N. J., & Levy, R., The effect of word predictability on reading time is logarithmic, 2013.

Supplemental page for Computational Modeling

Perplexity: Perplexity (PPL) is the inverse geometric mean of the joint probability of a set of words $w_1...w_n$ of a heldout corpus C. Formally, $PPL(C) = p(w_1...w_n)^{-1/N}$ Perplexity measures how well, on average, a language model predicts the upcoming token.

Delta Log Likelihood: ΔLogLik is used to measure models' predictive power. It is the difference in log likelihood between two linear regression models, which have human reading times in milliseconds as their response variables. The first model is a baseline model; it contains only context-insensitive aspects of a word as predictors. The second is a context-sensitive model and includes its surprisal as a predictor. The difference between the log likelihood indicates the amount of reading time variance that can be explained by the LM-derived surprisals. The equation used to derive the ΔLogLik is given in (Eq. 2), below. For the eye-tracking corpus, we included the previous word information as predictor variables; for self-paced reading corpora we included the three previous words. We use 10-fold cross-validation to derive ΔLogLik.

*(**Eq. 2)** Log_Liklihood(lm(psychometric ~ surprisal + word_length + word_frequency) - lm(psychometric ~ word_length + word_frequency)).*

Models Tested

• *N-gram:* We use a 5-gram model with modified Kneser-Ney smoothing.

• *Long Short Term Memory RNN (LSTM):* We use a vanilla long short-term memory network (Hochreiter and Schmidhuber, 1997) based on a boilerplate Pytorch implementation.

• *Ordered Neurons LSTM:* We consider the Ordered-Neurons LSTM architecture (Shen et al., 2019) which encodes an explicit bias towards modeling hierarchical structure by imposing an ordering restriction on the hidden state of an LSTM-RNN.

• *Transformer* (GPT2): Transformers are deep neural networks which use self-attention. We train the GPT2 transformer architecture (Radford et al., 2019) from scratch. Some GPT2 models use *Byte-Pair Encoding (Sennrich et al., 2015)*, which leverages sub-word statistics from a much larger corpus (~8 billion words).

• *Recurrent Neural Network Grammar* (RNNG): RNNGs (Dyer et al., 2016) model the joint probability of a sentence and syntactic structure and are trained on labeled trees that contain complete constituency parses, which we produce for BLLIP sentences with an off-the-shelf constituency parser (Kitaev and Klein, 2018). To compute surprisals from RNNG, we use word-synchronous beam search to approximate the conditional probability of the current word given the context.

Test Suite Circuits: Below is a list of the six circuits with example test items from one of the test suites in the circuit. */# indicate non-grammatical or less felicitous variants. Underlines indicate critical regions where models' by-word probabilities were measured.

Center Embedding	The diamond that the thief {<u>stole glittered</u> / *<u>glittered stole</u>}
Agreement	The keys to the cabinet {<u>are</u> / *<u>is</u>}
Garden Path Effects	The player {#tossed / thrown} the frisbee <u>tripped</u>.
Gross Syntactic State	{While / *∅ } the pianist practiced, <u>the conductor stretched.</u>
Licensing	{*The / No} senator has <u>ever</u> lied.
Long-Distance Dependencies	I know {who / *that} the mayor invited __ <u>to the party.</u>

Additional References: Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. • Sennrich, R., Haddow, B., & Birch, A. (2015). *Neural machine translation of rare words with subword units*