Word-Level Estimation of Shannon and Rényi Entropy Improves Reading Time Predictions

Christian Clark (clark.3664@osu.edu),[1] Byung-Doh Oh,[2] William Schuler[1]
[1]Ohio State, [2]Nanyang Technological University

Background. Recent studies of human sentence processing have reported evidence of psycholinguistic effects from the contextual entropy of the current word being processed [8, 11, 4]. The most widely studied variant of contextual entropy is Shannon entropy, defined as the expected surprisal of the current word W_i: $\mathsf{H}(W_i \mid w_{1..i-1}) = -\sum_{w \in V} \mathsf{P}(w \mid w_{1..i-1}) \log_2 \mathsf{P}(w \mid w_{1..i-1})$, where V is the vocabulary. One study [8] also considers effects from Rényi entropy, a generalization of Shannon entropy that captures a wider range of possible processing strategies. The Rényi entropy of order α is defined as $\mathsf{H}_\alpha(W_i \mid w_{1..i-1}) = \lim_{\beta \to \alpha} \frac{1}{1-\beta} \log_2 \sum_{w \in V} (\mathsf{P}(w \mid w_{1..i-1}))^\beta$; setting $\alpha = 1$ recovers Shannon entropy. Shannon and Rényi entropy measure the predictive processing difficulty before each new word is encountered, in contrast to raw surprisal, whose effects can be understood as integration costs for an already observed word [1, 8].

Previous work estimates contextual entropy using a language model (LM) like GPT2 [9]. However, because words can span multiple subword tokens in an LM's vocabulary—and therefore are intractable to sum probabilities over—entropy is typically calculated over each word's first token instead. This practice results in a systematic underprediction of true word entropy [8], which is magnified in contexts in which multi-token words are probable. To address this issue, we calculate LM-based entropy estimates using a Monte Carlo (MC; [7]) technique that randomly samples token sequences to explore, and thus allows words to span multiple tokens. We then evaluate the fit of the MC estimates to naturalistic reading times.

Methods. Mixed-effects regression experiments were conducted on five English reading time corpora containing self-paced reading times [3, 10] or first-pass (FP) and go-past (GP) durations from eye tracking [5, 6, 2]. Baseline predictors in the regression models included word length, word index, unigram surprisal, LM surprisal of the current and previous word (SPR, FP, and GP), and whether the previous word was fixated (FP and GP only). Per-subject random slopes were initially included for all predictors, but some were removed to ensure convergence. For each corpus and response type, 10-fold cross-validation was used to find an average difference in log likelihood (ΔLL) between a regression model containing only baseline predictors, and a regression model additionally containing an entropy predictor (either first-token entropy or MC word entropy). We tested Shannon entropy and Rényi entropy with $\alpha = 1/2$, the latter of which was shown by [8] to make strong reading time predictions. GPT2-small was the LM used to calculate entropy and surprisal predictors. MC estimates of word entropy were based on 512 next-word samples. Paired permutation tests over the per-fold ΔLL values were conducted to determine significance.

Results. When using Shannon entropy (Table 1a), replacing first-token estimates with MC estimates improves ΔLL scores on the Natural Stories self-paced reading corpus. However, results are less consistent across other corpora, with two cases in which first-token entropy significantly outperforms MC word entropy. Nonetheless, the improvement from MC word entropy on Natural Stories is strong enough to lead to a significant improvement in the aggregated permutation test.

The results using Rényi entropy (Table 1b) show a more robust improvement in reading time predictions from MC word entropy. Higher ΔLL values are observed across all corpora, with significant improvements from MC word entropy on several corpora and in the aggregated evaluation.

On the whole, the results point to concrete differences between LM-based entropy predictors that operate at the word level compared to token-level predictors. Especially in the case of Rényi entropy, the word-level predictors tend to provide a closer match to human reading times. The gap between word- and token-level predictors warrants caution against using first-token entropy in psycholinguistic modeling.

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$
NS SPR	1.26×10^{-4}	3.05×10^{-4} ***
Brown SPR	2.10×10^{-6}	-1.75×10^{-5}
Dundee FP	1.28×10^{-5}	-4.09×10^{-6}
Dundee GP	9.21×10^{-6}	-4.60×10^{-6}
Provo FP	-7.54×10^{-6}	1.18×10^{-5}
Provo GP	2.13×10^{-4} **	6.48×10^{-5}
GECO FP	8.11×10^{-5} ***	1.17×10^{-5}
GECO GP	-1.38×10^{-6}	-3.45×10^{-6}
Combined	7.09×10^{-5}	1.17×10^{-4} *

(a) Shannon Entropy

Corpus	$\Delta\text{LL}_{\text{FT}}$	$\Delta\text{LL}_{\text{MC}}$
NS SPR	2.61×10^{-4}	1.43×10^{-3} ***
Brown SPR	-1.05×10^{-5}	1.99×10^{-4}
Dundee FP	1.02×10^{-6}	1.10×10^{-4} **
Dundee GP	-4.09×10^{-6}	1.90×10^{-4} **
Provo FP	-1.29×10^{-5}	-8.95×10^{-6}
Provo GP	4.51×10^{-5}	6.57×10^{-5}
GECO FP	5.86×10^{-6}	2.76×10^{-5}
GECO GP	0.00	1.14×10^{-5}
Combined	9.85×10^{-5}	5.78×10^{-4} ***

(b) Rényi Entropy

Table 1: Increases in per-datapoint log likelihood (measured in nats) from adding a target entropy predictor to a baseline regression model for predicting self-paced reading (SPR) time, first-pass (FP) duration, or go-past (GP) duration. $\Delta\text{LL}_{\text{FT}}$ and $\Delta\text{LL}_{\text{MC}}$ respectively refer to log likelihood improvements from first-token and MC entropy approximations. NS means Natural Stories. Significance levels are * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References

[1] B. Cevoli et al. Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science*, 2022.

[2] U. Cop et al. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 2017.

[3] R. Futrell et al. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*, 2021.

[4] M. Giulianelli et al. Generalized measures of anticipation and responsivity in online language processing. In *EMNLP Findings*, 2024.

[5] A. Kennedy et al. The Dundee corpus. In *Proc. ECEM*, 2003.

[6] S. G. Luke and K. Christianson. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 2018.

[7] N. Metropolis and S. Ulam. The Monte Carlo method. *JASA*, 1949.

[8] T. Pimentel et al. On the effect of anticipation on reading times. *TACL*, 2023.

[9] A. Radford et al. Language models are unsupervised multitask learners. *ArXiv*, 2019.

[10] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 2013.

[11] E. G. Wilcox et al. Testing the predictions of surprisal theory in 11 languages. *TACL*, 2023.