Memory-Based Predictors From GPT-2 Attention Predict Reading Times Over Surprisal

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

Much work in broad-coverage sentence processing has focused on studying the role of expectation operationalized in the form of surprisal [3, 8] using language models (LMs) to define a conditional probability distribution of a word given its context [15, 2]. However, expectation-based accounts have empirical shortcomings, such as being unable to fully account for garden-path effects [16] or predict the timing of delays in certain constructions [9]. For this reason, some research has begun to focus on the effects of memory and attention using predictors calculated from language model representations. For example, [12] recently drew connections between the self-attention patterns of Transformers [17] and cue-based retrieval models of sentence comprehension [e.g. 10]. Their *attention entropy*, which quantifies the diffuseness of the attention weights over previous tokens, showed patterns that are consistent with similarity-based interference observed during the processing of subject-verb agreement. However, these results relied on identifying one attention head specialized for the *nsubj* dependency, and an aggregated version of this predictor was not very strong in predicting naturalistic reading times in the presence of a surprisal predictor [13].

This work therefore defines and evaluates several entropy- and distance-based predictors derived from the self-attention patterns of the Transformer-based GPT-2 language model [11] on two naturalistic datasets, in the presence of a strong GPT-2 surprisal baseline. First, normalized attention entropy (NAE) expands upon attention entropy [12] by re-normalizing the attention weights and controlling for the number of tokens in the previous context. Additionally, three distance-based predictors that quantify the shift in attention patterns across consecutive timesteps are presented, based on the idea that the reallocation of attentional focus entails processing difficulty. These predictors are ΔNAE, which quantifies the change in diffuseness across timesteps, Manhattan distance (MD), which directly measures the magnitude of change in attention weights over all tokens, and Earth Mover's Distance (EMD), which is the minimum amount of 'work' necessary to transform the allocation of attention over previous tokens from one timestep to the next. Moreover, motivated by work on interpreting large language models that question the connection between attention weights and model predictions [e.g. 4], a norm-based analysis of transformed vectors [6, 7] is applied to GPT-2 to define novel formulations of attention weights (Figure 1, in blue).

In order to evaluate the contribution of these predictors, continuous-time deconvolutional regression models [14] containing commonly used baseline predictors, unigram and GPT-2 surprisal, and one predictor of interest each were fitted to self-paced reading times [1] and eye-gaze durations [5] collected during naturalistic reading of English text. The baseline predictors include word length measured in characters and index of word position within each sentence (both Natural Stories and Dundee), as well as saccade length and whether or not the previous word was fixated (Dundee only). The memory-based predictors were calculated from the attention patterns of heads on the topmost layer of GPT-2 Small. The results in Figure 2 show that across both corpora, most of the predictors make a notable contribution to regression model fit under all attention formulations on held-out data, with Attn-N+NAE and AttnRL-N+MD being the most predictive among the entropy- and distance-based predictors respectively. The fact that the baseline model contains robust predictors such as unigram surprisal and GPT-2 surprisal supports Attn-N+NAE and AttnRL-N+MD as predictors of comprehension difficulty. In terms of magnitude, these two predictors showed large effect sizes on the Natural Stories Corpus, which were more than twice that of GPT-2 surprisal. On the Dundee Corpus, however, the effect size of AttnRL-N+MD was much smaller (Table 1).

These predictors showed moderate correlation to unigram surprisal at around 0.5 on both corpora, and weak correlation to GPT-2 surprisal at around 0.3 on Natural Stories, and around 0.4 on Dundee. Together with the regression results, this further suggests that the proposed predictors capture a measure of attention focus that is distinct from word frequency or predictability.

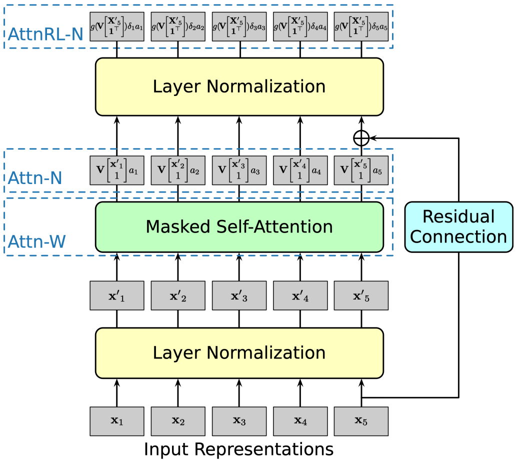

Figure 1: Computations performed within the self-attention block of one head of the GPT-2 language model at a given timestep ($i = 5$). The linear nature of the subsequent computations allows this aggregation to be deferred to after the residual connection and layer normalization, thereby allowing updated representations to define novel formulations of attention weights (i.e. Attn-N, AttnRL-N).

Corpus	Predictor	Effect Size (p-value)
Natural Stories	Attn-N+NAE	6.87 ms ($p < 0.001$)
	GPT2Surp	2.56 ms
	AttnRL-N+MD	6.59 ms ($p < 0.001$)
	GPT2Surp	2.82 ms
Dundee	Attn-N+NAE	N/A (n.s.)
	GPT2Surp	4.22 ms
	AttnRL-N+MD	1.05 ms ($p < 0.001$)
	GPT2Surp	3.81 ms

Table 1: The per standard deviation effect sizes of the predictors on the held-out partition of the Natural Stories Corpus and the Dundee Corpus. Statistical significance was determined by a paired permutation test of the difference in by-item squared error between the baseline regression model and the respective full regression model containing the predictor of interest. The effect sizes of GPT-2 surprisal from the same regression models are presented for comparison.

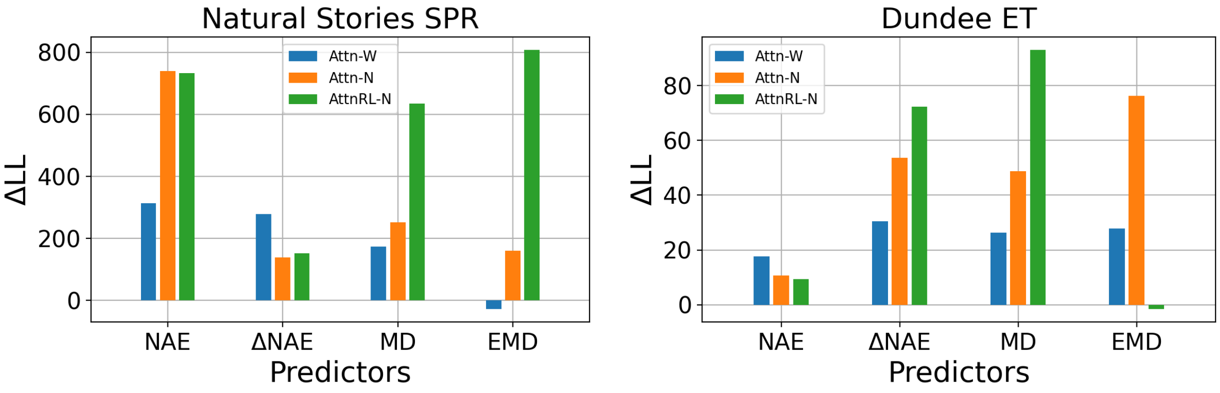

Figure 2: Improvements in CDR model log-likelihood from including each predictor on the exploratory partition of Natural Stories self-paced reading data (left) and Dundee eye-tracking data (right).

References

[1] R. Futrell et al. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*, 55:63–77, 2021.

[2] 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.

[3] J. Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proc. NAACL*, pages 1–8, 2001.

[4] S. Jain and B. C. Wallace. Attention is not explanation. In *Proc. NAACL*, pages 3543–3556, 2019.

[5] A. Kennedy et al. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*, 2003.

[6] G. Kobayashi et al. Attention is not only a weight: Analyzing Transformers with vector norms. In *Proc. EMNLP*, pages 7057–7075, 2020.

[7] G. Kobayashi et al. Incorporating residual and normalization layers into analysis of masked language models. In *Proc. EMNLP*, pages 4547–4568, 2021.

[8] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

[9] R. Levy et al. The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, 69:461–495, 2013.

[10] R. L. Lewis et al. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10(10):447–454, 2006.

[11] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[12] S. H. Ryu and R. L. Lewis. Accounting for agreement phenomena in sentence comprehension with Transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proc. CMCL*, pages 61–71, 2021.

[13] S. H. Ryu and R. L. Lewis. Using Transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th Annual Conference on Human Sentence Processing*, 2022.

[14] C. Shain and W. Schuler. Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215, 2021.

[15] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.

[16] M. van Schijndel and T. Linzen. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6), 2021.

[17] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.