Pre-trained language models can track some ERP components in language processing

Jiaxuan Li, Richard Futrell
University of California, Irvine

Introduction: N400 and post-N400 positivities (PNP) are sensitive to predictability of the next word and the degree to which sentence context constraints following words [1, 2]. Traditional measures of word predictability and contextual constraint are based on human offline judgment. In contrast, pre-trained language models are optimized for word prediction based solely on language input and might reflect the statistical distributions in language better. We assess the validity of computational neural network measures in predicting N400 and PNP, and compare the predictive power of neural network measures with human offline judgments.

Method: We used the EEG dataset (n=24) from [3]. There were 780 Chinese numeral-classifier-noun and verb-noun phrases. Cloze probabilities were collected from a noun-completion task. In the original experiment, each construction was grouped into five conditions (Table 1) according to *expectancy* (cloze) and *constraint* (max-cloze). We calculated entropy and surprisal of materials from seven pre-trained language models [4, 5], and from human cloze task. Based on analysis in [3], the ERPs on critical nouns from four regions (Anterior, Mid-frontal, Mid-posterior, Parietal) around midline electrodes in the 300-500ms (N400) and 600-1000ms (PNP) time windows were analyzed. We used linear-mixed effect models (see Formula M0) for predictors from pre-trained language models (*gpt2*; *bert-1,2*; *rbt-1,2*; *roberta-1,2,* where models marked with 2 are larger variants) with a Bonferroni correction on p-value, and from human offline completion task (*human*). We contrasted the results with the model (*condition)* used in [3] with pre-defined contrasts (Formula M1).

Result: *Correlation*: There was a significant correlation between cloze and surprisal across all language models ($rs < -.1$, $ps < .001$). Entropy estimated from *gpt2* was correlated with *constraint* ($r = -.16$, $p < .001$) and *human* ($r = .27$, $p < .001$). In contrast, entropy from *rbt-1* was correlated with measures of contextual constraints in an opposite direction (*constraint*: $r = .19$; *human*: $r = -.19$, $ps < .001$). Overall, *gpt2* best tracks predictability and contextual constraint in human offline comprehension. *ERP (Classifier)***:** across various pre-trained language models and cloze-based models, there was a significant surprisal effect on N400, although with different scalp distributions. In the PNP time window, there was no significant main effect of surprisal or entropy for all language models, whereas *condition* predicted a significant cloze effect in anterior-frontal region, and a constraint effect in parietal region (human cloze). *ERP (Verb)***:** For verb construction, only *gpt-2* predicted a significant surprisal effect on N400 as *condition* and *human*. In the PNP time window, surprisal from *bert-2* had a significant effect on N400, though the scalp distribution was different from models with cloze-based predictors. Importantly, *bert-2* and *condition* both predicted a significant constraint effect on the anterior-frontal PNP. The results are summarized in Fig. 1.

Conclusion: We find that surprisal estimated from *gpt2* can predict N400, whereas entropy calculated from *bert-2* are tentatively more promising to capture PNP component. The capacity of different language models to track ERPs might be related to the learning objective of models: GPT is optimized for next-word prediction, which assigns with the predictive nature of N400. BERT is trained to recover the masked token in middle of a sentence, which might make it better for predicting ERP components correlated with conflict resolution or re-analysis. Language models generally find ERPs elicited by verb construction are more difficult to predict than by Classifier constructions, likely due to their difficulties to handle event structure and world knowledge required for verb processing.

Table 1. Experimental stimuli.

	High Constraint			Low Constraint	
	High Cloze	Low Cloze	Anomalous	Low Cloze	Anomalous
Classifier	一扇门 one-CL door	一扇猪肉 one-CL pork	一扇水果 one-CL fruit	一块蛋糕 one-CL cake	一块水 one-CL water
Verb	激化矛盾 Intensify conflict	激化能量 Intensify energy	激化灯 Intensify lamp	影响贸易 Influence trade	影响时间 Influence time

Formula.
M0: Amplitude ~ surprisal + entropy + (1+surprisal + entropy | subject) + (1|item)
M1: Amplitude ~ contrast1 (high cloze v.s. low cloze v.s. anomalous) + contrast2 (high constraint low cloze v.s. low constraint low cloze) + (1 + contrast1 + contrast2 | subject) + (1|item)

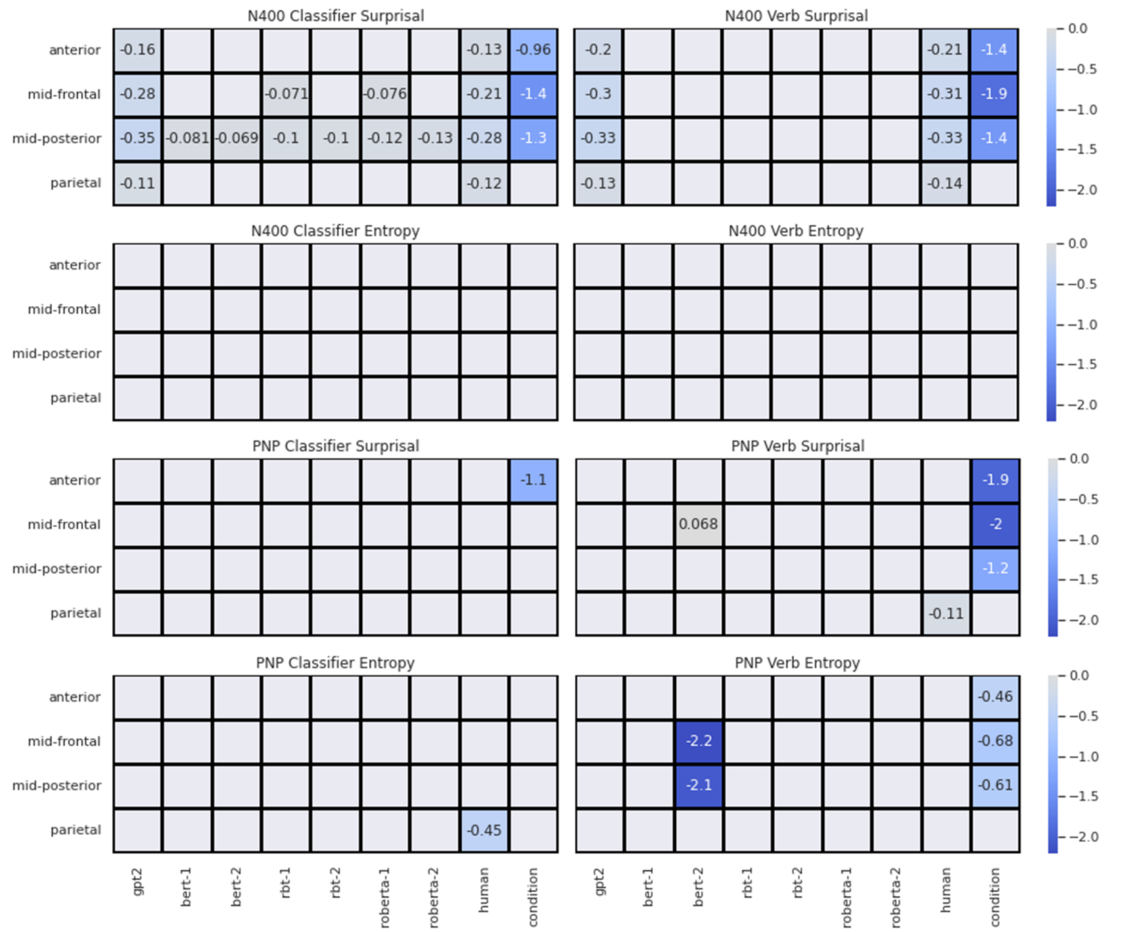

Fig. 1. Statistical analysis with model M0 and M1 (*condition*). Entropy and surprisal are estimated from pre-trained language models or human completion tasks. The colored cells represent significant effects, and the numbers are estimated beta values.

References. [1] Kutas & Hillyard (1984). *Nature*, *307*(5947), 161-163. [2] DeLong & Kutas (2020). *Language, Cognition and Neuroscience*, *35*(8), 1044-1063. [3] Li, Ou & Xiang (2021). 34th CUNY Conference. [4] Zhao et al. (2019). *EMNLP-IJCNLP-2019,* 241. [5] Cui et al. (2020). *Association for Computational Linguistics,* 657-668.