Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential

Lindborg & Rabovsky, *P of the Ann Meet of the Cog Sci Soc.* 43 (2021)

Pratik Bhandari

Paper discussion (vd-mit)

20/06/2022

Summary of the paper

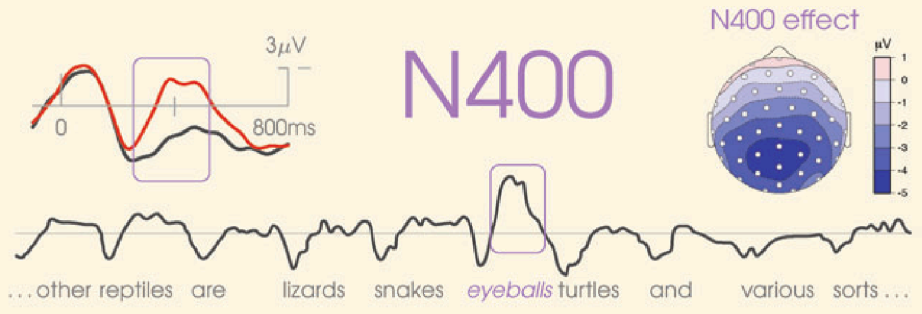
GPT-2 activation updates can predict N400 amplitudes.

Online semantic updates in a deep learning model and the human brain are (partially) similar.

[See also Caucheteux & King (2022) for a similar (and reverse) claim using fMRI and MEG data]

N400

- A **N**egative going Event-related potential
- Appears ∼**400** ms after word onset



- Associated with semantic processing and meaning representation (e.g., Kutas & Federmeier, 2011)

GPT-2

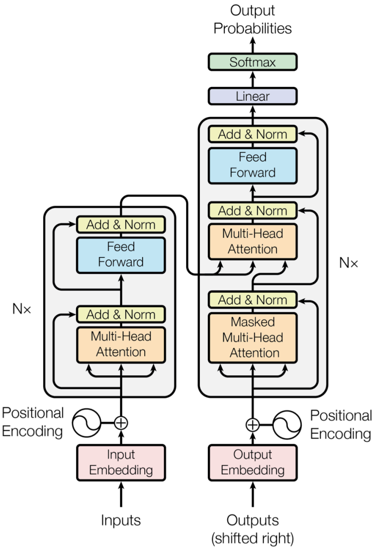


Figure 1: GPT-2 model architecture (Vaswani et al., 2017)

Stack of 36 decoder modules is used

GPT-2

- Trained with a large body of text
- Isn't explicitly modeled to 'represent' meaning of the text
- Has access to both past and future context (compare with a human)

This study

Q: Does the representation of meaning in GPT-2 correspond with N400 amplitude?

where,

representation of meaning in GPT-2 → network updates $u(n)$ calculated at each 36 layer

$$u(n) = \sum_{i=1}^{D} |a(n)_i - a(n-1)_i| = \|a(n) - a(n-1)\|_1$$

Is activation updates $a(n) \approx$ output probabilities at each decoder layer?

Experiments

- Quantitative experiments:
  - Presented stimuli from Frank et al. (2013, 2015) to GPT-2
  - Evaluated if GPT-2 updates predict N400, and if the effects of lexical-semantic variables on GPT-2 updates and N400 are similar
- Qualitative experiments:
  - Tested in 4 experimental paradigms, if GPT-2 can simulate N400 effects

Quantitative experiments

- Participants in the EEG study (Frank et al., 2015): n=24
- Items: 205 sentences from the UCL corpus of reading times
- Other lexical-semantic variables:
  - Log-transformed word frequency (British National Corpus)
  - Sentence position
  - Surprisal (*4*-gram model)

Quantitative experiments

- First set of tests: if activation updates significantly predict N400 amplitudes (at each word)
- Second set of tests: if lexical-semantic variables have same effect on both N400 and u

Quantitative experiments: Analysis 1

Linear mixed effects models to test if activation updates (u) significantly predict N400 amplitudes (at each word)

- For j^{th} decoder layer ranging from 1 to 36
    - `m_j <- lmer(N400 ~ 1 + u_j + N400_baseline + 1|subject + 1|item...)`
- Compare with a base model
    - `m0 <- lmer(N400 ~ 1 + N400_baseline + 1|subject + 1|item...)`

Quantitative experiments: Result 1

- The effect of u was significant in 31 models;
  - largest in the "deep intermediate layers 21-25"; non-significant in the outer layers

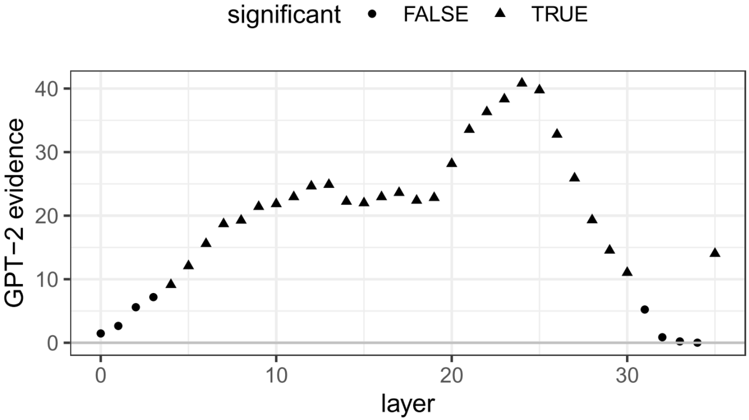


Figure 2: Significant LME by GPT-2 layers

Quantitative experiments: Analysis 2

- Predicting N400 from three lexical-semantic variables
  - `m_f <- lmer(-1*N400 ~ 1 + N400_baseline + frequency + 1|subject + 1|item...)`
  - `m_p <- lmer(-1*N400 ~ 1 + N400_baseline + position + 1|subject + 1|item...)`
  - `m_s <- lmer(-1*N400 ~ 1 + N400_baseline + surprisal + 1|subject + 1|item...)`
- Predicting activation updates at each decoder layer from three lexical-semantic variables
  - `m_j_f <- lmer(u_j ~ 1 + frequency + 1|item...)`
  - `m_j_p <- lmer(u_j ~ 1 + position + 1|item...)`
  - `m_j_s <- lmer(u_j ~ 1 + surprisal + 1|item...)`
- Compared the significance of the effect and direction of lexical-semantic variables on N400 and activation updates

Quantitative experiments: Result 2 (Effects of lexical-semantic variables)

- On N400: Significant effects only of *surprisal* and *sentence position*
- On activation updates: Significant effects also of *lexical frequency*
  - Non-sigificant effect only of surprisal in the *deep layers* 32-36

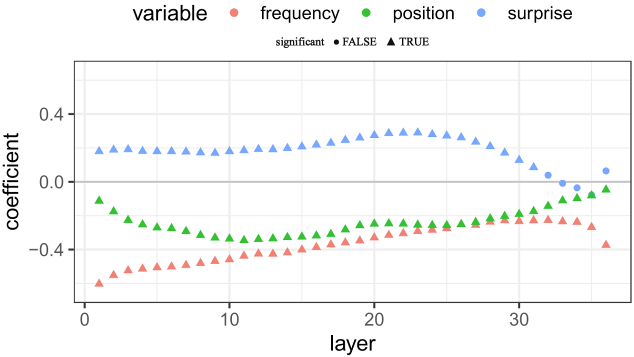


Figure 3: Regression coefficients for lexical-semantic predictors at all GPT-2 layers

What does it mean?

- Surprisal effects not significant in the deep layers
  - Contrast with N400 + P600 effects, i.e., late processing (e.g., Brouwer et al., 2021)

Qualitative experiments

Can GPT-2 activation updates simulate N400 effects in 4 experimental paradigms?

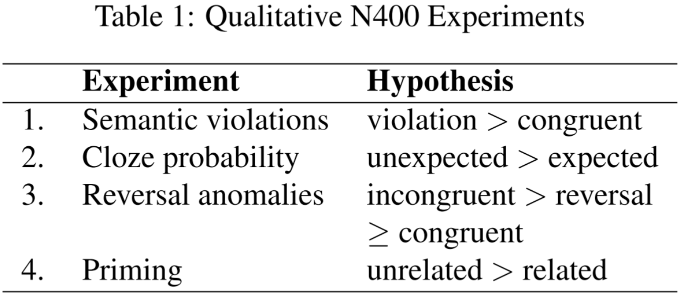


Table 1: Qualitative N400 Experiments

	Experiment	Hypothesis
1.	Semantic violations	violation $>$ congruent
2.	Cloze probability	unexpected $>$ expected
3.	Reversal anomalies	incongruent $>$ reversal $\geq$ congruent
4.	Priming	unrelated $>$ related

Qualitative experiment: Semantic violations

Incongruent > Congruent

- Congruent: *I take my coffee with cream and sugar.*
- Violation (aka Incongruent): *I take my coffee with cream and dog.*

Qualitative experiment: Cloze probability

Unexpected > Expected

- Expected: *The children went outside to play.*
- Unexpected: *The children went outside to talk.*

Qualitative experiment: Reversal anomalies

Incongruent $>$ Reversal $\geq$ Congruent

- Congruent: *For breakfast, the boys would only eat ...*
- Reversal anomaly: *For breakfast, the eggs would only eat ...*
- Inongruent: *For breakfast, the boys would only plant ...*

Qualitative experiment: Priming

Unrelated $>$ Related

- Related: *school-university*
- Unrelated: *school-lip*

Qualitative experiments: Results

- Deep intermediate layers activation updates approximate N400 effects for *semantic violation* (experiment 1)
    - Fewer layers for *cloze probability* (experiment 2)

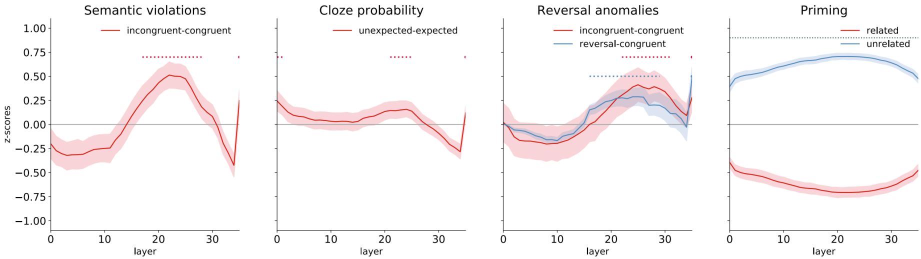


Figure 4: Normalized scores (activation updates) across experimental conditions at all GPT-2 layers. Dotted lines show layers where significant effects were found.

What does it mean?

- Similar to "*activation predicts N400*" in *surprisal* (Quantitative experiment)
- Gradual built-up of meaning starting from the outer layers, although, no 'graduality' in the architecture
    - Meaning representation significant only in the intermediate deep layers?
        - Again, absent in the deep layers 30+

Qualitative experiments: Results

- Deep layers for *reversal anomalies* (experiment 3), but incongruent > reversal effect not significant

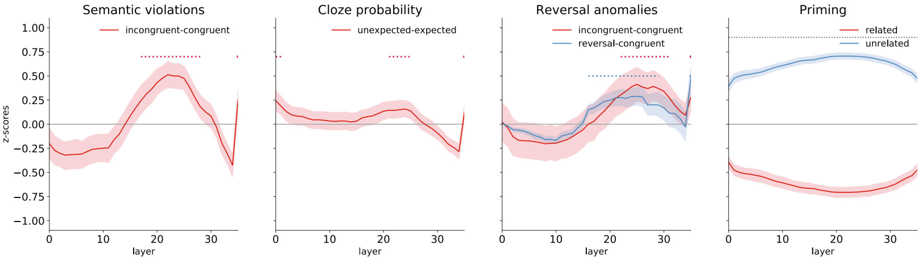


Figure 5: Normalized scores (activation updates) across experimental conditions at all GPT-2 layers. Dotted lines show layers where significant effects were found.

Qualitative experiments: Results

- Unrelated > Related effect (experiment 4) significant at *all* layers

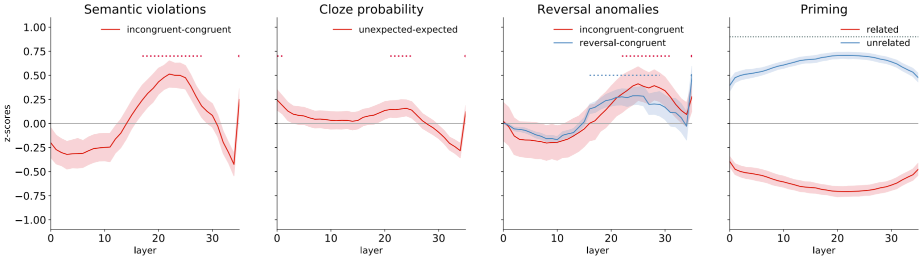


Figure 6: Normalized scores (activation updates) across experimental conditions at all GPT-2 layers. Dotted lines show layers where significant effects were found.

What does it mean?

- Priming effect: too low activation for unrelated words across all layers although GPT-2 is trained on words(?)
    - Contrast with Expt 1 and 2 in which separate activation is not displayed for con/incong; unexp/exp.

Discussion

- Relatively modest claim offered by the authors
- Today's behavioral and neural measures are still coarse grained
- Large language models' can model human language processing
  but it is a statistical similarity rather than the analogy of human lang. proc.