MIAMI UNIVERSITY

Comparing neural-network based language models to human sentence processing: choice of task matters

March 20-24, 2022

Mai Ha Vu & So Young Lee
University of Oslo, Miami University

Human Sentence Processing Conference 2022

Introduction

- **BERT**: A deep transformer-based language model with state-of-the art performance on natural language tasks → *How does it match human performance for sentences where humans usually make mistakes due to processing?*
- **NPI illusion effect**: Reduced P600 in ERP measure for *ever* in illusion and licensed conditions compared to unlicensed condition (Xiang et al., 2009)

  (1) *The horses [that no gamblers have bet on] have ever won. → Illusion (false positive)
  (2) No horses [that the gamblers have bet on] have ever won. → Licensed (true positive)
  (3) *The horses [that the gamblers have bet on] have ever won. → Unlicensed (true negative)

- Shin & Song (2020): BERT shows no NPI illusion effect, but the surprisal scores were calculated for the licensor, not the NPI → differs from Xiang et al. (2009)
- **Research question**: Does BERT show *NPI illusion effects* if the test task is more similar to the experimental task we are replicating (Xiang et al., 2009)?

Methods

- **Stimuli**: 150 sentence stimuli adapted from Xiang et al (2009)
- **Model**: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), specifically pretrained bert-large-uncased
- **Measure**: Surprisal from BERT's softmax layer for specific lexical items in the place of [MASK]
    - Experiment 1: Predict the licensor
      [MASK] horses that the gamblers have bet on have ever won. *no* vs. *the*
    - Experiment 2: Predict the licensee
      No horses that the gamblers have bet on have [MASK] won. *ever*
- Independent variables for both experiments:
    - Licensor: no vs. the
    - Licensor position: matrix vs. embedded
    - Plurality of modified licensor DP: singular vs. plural

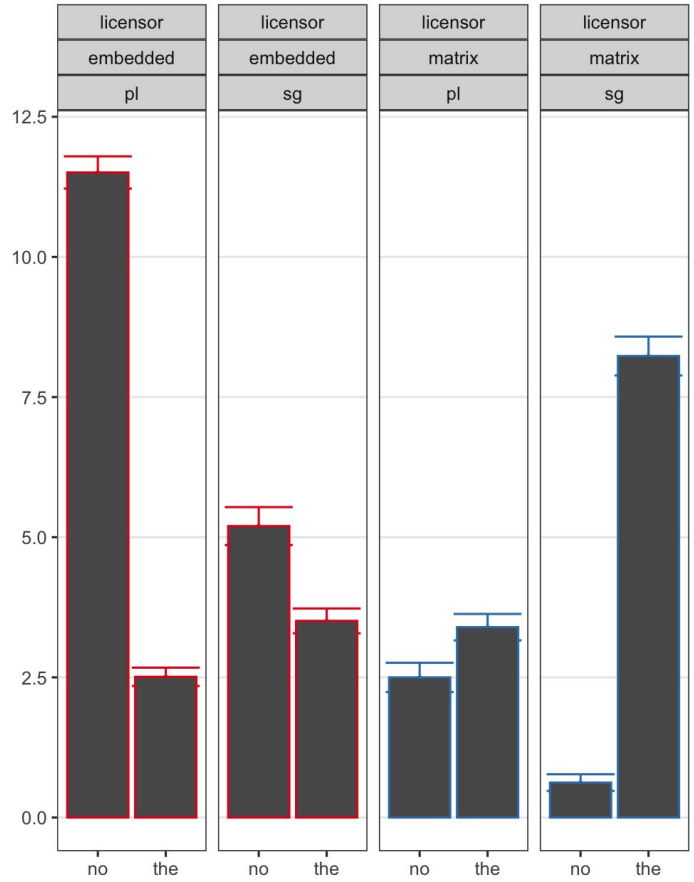
Results: Experiment 1

(4) The horses [that {no/the} gamblers have bet on] have ever won. (Embedded, plural)

(5) The horses [that {no/the} gambler has bet on] have ever won. (Embedded, singular)

(6) {No/The} horses [that the gamblers have bet on] have ever won. (Matrix, plural)

(7) {No/The} horse [that the gamblers have bet on] has ever won. (Matrix, singular)



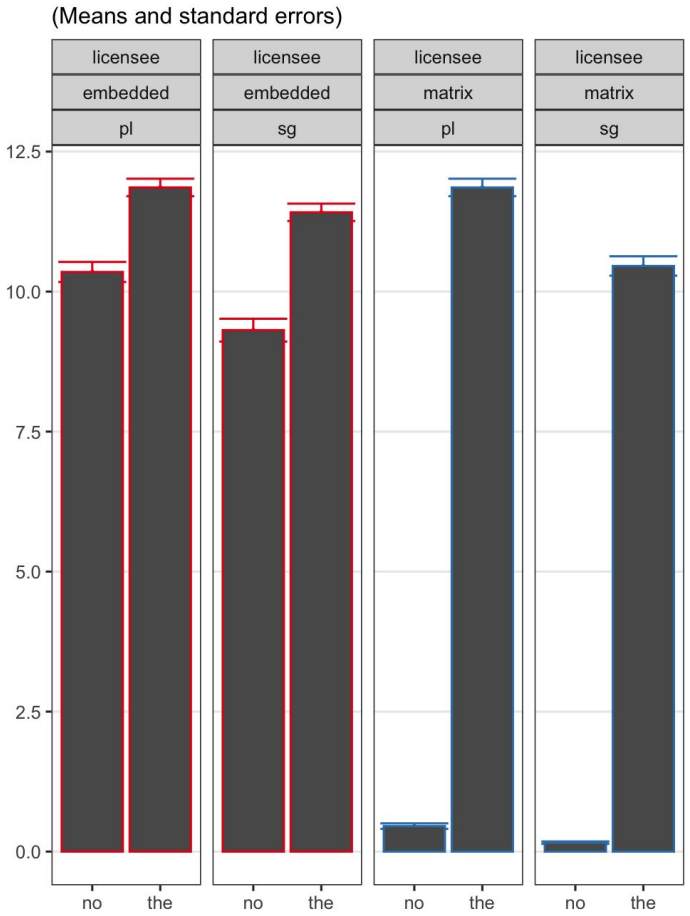
Results: Experiment 2

(8) a. The horses [that no gamblers have bet on] have {ever} won. (Embedded *no*, plural)
b. The horses [that the gamblers have bet on] have {ever} won. (Embedded *the,* plural)

(9) a. The horses [that no gambler has bet on] have {ever} won. (Embedded *no*, singular)
b. The horses [that the gambler has bet on] have {ever} won. (Embedded *the*, singular)

(10) a. No horses [that the gamblers have bet on] have {ever} won. (Matrix *no*, plural)
b. The horses [that the gamblers have bet on] have {ever} won. (Matrix *the*, plural)

(11) a. No horse [that the gamblers have bet on] has {ever} won. (Matrix *no*, singular)
b. The horse [that the gamblers have bet on] has {ever} won. (Matrix *the*, singular)



Discussion

- **Main findings**:
  - We could not fully replicate Shin and Song's (2020) findings - results slightly match if we consider plural condition for "no" conditions and singular condition for "the" condition
  - BERT showed illusion effect when it had to calculate surprisal scores for the NPI (Anova, TukeyHSD $p < .001$) (as opposed to for the licensor)
- **Limitations**:
  - BERT is bi-directional, so experiment did **not** replicate human *online* processing, which is what Xiang et al. (2009) were studying
- **Future work**:
  - Force BERT to give unidirectional judgments
  - Examine other measures of LM performance, e.g., the ones listed in Warstadt et al. (2019)

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Shin, Unsub, and Sanghoun Song. 2020. “BERT, a Deep-Learning Language Model, Learns NPI Licensing but Does Not Suffer from NPI Illusion.” https://www.cuny2021.io/wp-content/uploads/2021/02/CUNY_2021_abstract_243.pdf.

Warstadt, Alex, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, et al. 2019. “Investigating BERT’s Knowledge of Language: Five Analysis Methods with NPIs.” *ArXiv:1909.02597 [Cs]*, September. http://arxiv.org/abs/1909.02597.

Xiang, Ming, Brian Dillon, and Colin Phillips. 2009. “Illusory Licensing Effects across Dependency Types: ERP Evidence.” *Brain and Language* 108 (1): 40–55. https://doi.org/10.1016/j.bandl.2008.10.002.