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

Mai Ha Vu[1], So Young Lee[2]
[1]University of Oslo, [2]Miami University

Due to the high performance of recent neural-network based language models (LMs) such as BERT[1], there has been a growing body of research comparing them to human language processing[2–4]**.** A method in answering this question is replicating human sentence processing studies with LMs, but the choice for the test task remains understudied. In this paper we first fail to replicate Shin and Song's (2020) exact results[5], but in the second experiment, where we choose a task that is more faithful to the original human experiment[6], we replicate and strengthen their conclusion that BERT is not affected by illusion effects in Negative Polarity Item (NPI)-licensing.

NPIs such as *ever* must be c-commanded by a licensor, such as *no.* Human processors, however, can mistakenly accept an NPI to be licensed even when the potential licensor only linearly precedes, but does not c-command it[6], as in (1a) (cf. (1b))*.* Shin and Song (2020) found that BERT gave a lower surprisal score in predicting *no* in the matrix clause than either *no* or *the* in the embedded clause, indicating that BERT is not influenced by illusion effects.

We conducted two experiments. In Experiment 1, we replicated Shin and Song (2020) with the same pre-trained BERT LM and dataset[6], and an added condition of plural and singular noun phrases in the potential licensor positions (2-3). Shin and Song (2020) report to only have tested plural NPs. Since BERT gives the same scores in the plural NP condition regardless of whether it also calculated scors for the singular NP condition, the plural condition is an exact replication of Shin and Song's (2020) experiment, which we fail to do (Figure 1). Unlike in Shin and Song (2020), the surprisal score in the plural condition was not different between matrix "*no*" and matrix "*the*" ($p=0.2$), or matrix "*no"* and embedded "*the"* ($p=1$). However, in the singular condition the score for predicting matrix "*the*" matched what was reported in Shin and Song (2020) for "*the*", higher than the score for matrix "*no*" ($p<0.001$). Thus, plurality was a confound, confirmed by the interaction between licensor and plurality ($p<0.001$), and Shin and Song (2020) likely reported singular data for "*the*", but plural data for "*no*".

In Experiment 2, BERT calculated surprisal scores for the NPI instead of the licensor (4-5), a more faithful task to the original human study, which measured ERPs for the NPI[6]. The results showed no interaction between licensor and plurality ($p=0.28$), and consistently high scores for all the conditions where the NPI was unlicensed, including for illusion effect. When matrix "*no*" licensed the NPI, the surprisal score for the NPI was the lowest across both experiments (Figure 2). In other words, changing the task to predict the NPI instead of the licensor provided more unambiguous results.

The results showed that predicting the determiner in licensor position was confounded by the plurality of the NP. However, switching to an LM task that was more similar to the original human experimental task strengthened previous findings that pre-trained BERT learned the structural conditions for NPI licensing and is unaffected by illusion effects[5].

(1) a. *The horses [that no gamblers have bet on] have ever won.
b. No horses [that the gamblers have bet on] have ever won.

Experiment 1:

(2)	a. The horses [that {no/the} gamblers have bet on] have ever won.	Embedded, plural
	b. The horses [that {no/the} gambler has bet on] have ever won.	Embedded, singular
(3)	a. {No/The} horses [that the gamblers have bet on] have ever won.	Matrix, plural
	b. {No/The} horse [that the gamblers have bet on] has ever won.	Matrix, singular

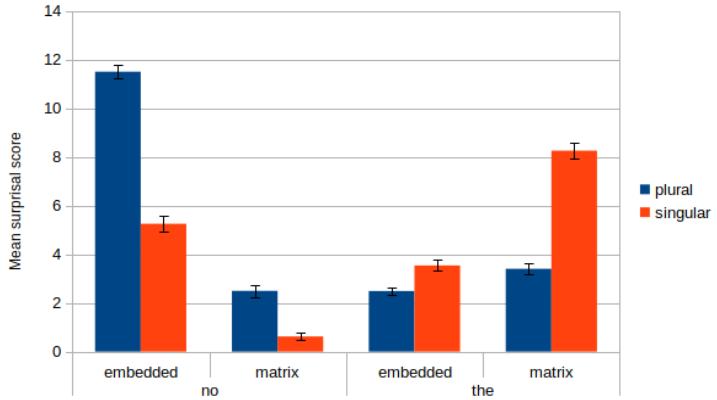

Figure 1. Mean surprisal scores for predicting *the* or *no*. Bars indicate standard error.

Experiment 2:

(4)	a. The horses [that **no** gamblers have bet on] have {ever} won.	Embedded *no*, plural
	b. The horses [that **no** gambler has bet on] have {ever} won.	Embedded *no*, singular
	c. **No** horses [that the gamblers have bet on] have {ever} won.	Matrix *no*, plural
	d. **No** horse [that the gamblers have bet on] has {ever} won.	Matrix *no*, singular
(5)	a. The horses [that **the** gamblers have bet on] have {ever} won.	Embedded *the*, plural
	b. The horses [that **the** gambler has bet on] have {ever} won.	Embedded *the*, singular
	c. **The** horses [that the gamblers have bet on] have {ever} won.	Matrix *the*, plural
	d. **The** horse [that the gamblers have bet on] has {ever} won.	Matrix *the*, singular

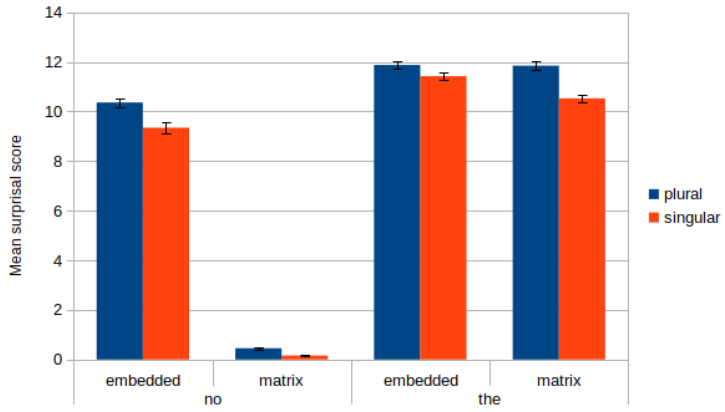

Figure 2. Mean surprisal scores for predicting *ever*. Bars indicate standard error.

Bibliography:

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*
[2] Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). *Proceedings of the Association of Computational Linguistics*
[3] Marvin, R., & Linzen, T. (2018). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*
[4] Warstadt, A., & Bowman, S. R. (2020). *Proceedings of the 42th annual conference of the Cognitive Science Society*
[5] Shin, U., & Song, S. (2020). *CUNY*
[6] Xiang, M., Dillon, B., & Phillips, C. (2009). *Brain and Language*