BERT, a deep-learning language model, learns NPI licensing but does not suffer from NPI illusion

Unsub Shin and Sanghoun Song (Korea University)

Recent development in computational models made it easier and more accurate to simulate human behavior in sentence processing (Wilcox et al., 2020; Merkx & Frank, 2020). Present study investigated whether a deep Transformer model BERT (Devlin et al., 2019) processes long-distance dependency and grammatical illusion in the same way as human language processors do. More specifically, we examined how BERT processes NPI licensing and NPI illusion.

Negative Polarity Items (NPIs) such as *ever* constitute a grammatical sentence only when it is c-commanded by a word or licensor that provides a negative context such as *no* or *few* e.g., *No*/**Some*/**The prisoner has ever talked to the priest* (Ladusaw, 1980). Research showed human processors are good at detecting whether an NPI and its licensor make a legitimate structural relationship. It was also shown they may mistakenly accept a potential licensor not occurring in a c-commanding position, for example in an embedded clause such as **The man [that no woman liked] has ever been to the party*. This phenomenon is called NPI illusion (Vasishth et al. 2008). Recent studies suggested BERT can capture some semantic features and structural information (Hewitt & Manning, 2019) but it is not fully resolved whether BERT can also learn the linguistic mechanism underlying NPI processing (Warstadt & Bowman 2020).

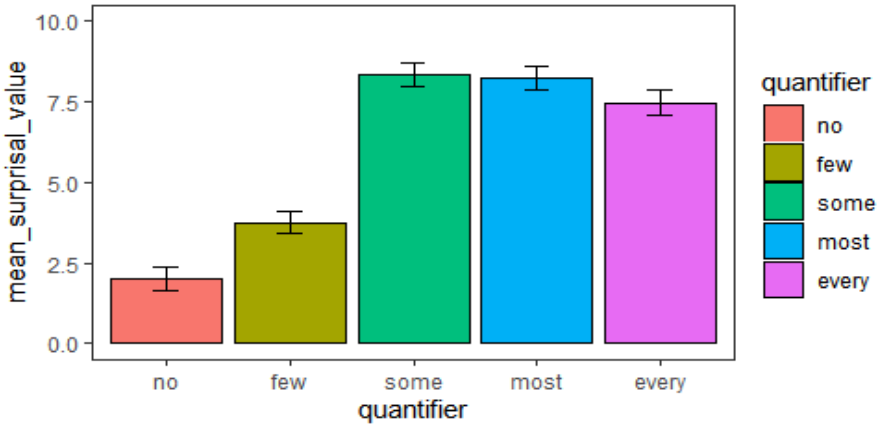
We conducted two experiments with BERT: We first investigated whether BERT can discern between semantically licit (negative) and illicit (positive) licensors of the NPI *ever* by testing five different quantifiers, *no, few, some, most* and *every* (Experiment 1). We used 150 sentence stimuli adapted from Xiang et al.'s (2009) (Table 1). Second, we examined whether BERT is susceptible to NPI illusion like humans by varying syntactic positions of the potential licensors (Experiment 2). We tested another set of 150 sentence stimuli in which a potential licensor *no* occurs in an embedded clause, violating the c-commanding condition of NPI licensing, and compare it with the condition where *no* occurs in the legitimate matrix clause. We also tested stimuli with no negative word as a control, i.e. *the*. In both experiments, we evaluated model performance by computing lexical surprisal values (Smith & Levy, 2013) from the output softmax layer, i.e. higher surprisal as a sign of increased processing difficulty.

The results of Experiment 1 (Figure 1) using Dunn's pairwise comparison shows that BERT captures the difference between strong NPI licensor *no* and weak NPI licensor *few* ($z = 3.45$, $p < .006$) and between negative quantifier *no* and non-negative quantifier *most* ($z = 11.74$, $p < .001$). The results of Experiment 2 (Figure 2) reveal that BERT discriminates between the licit and illicit position of NPI licensors ($z = 14.82$, $p < .001$). A much higher surprisal score for the embedded position indicates that the model successfully detects a structural violation. The fact that it is slightly higher than the surprisal for the no-licensor condition ($z = 2.07$, $p < .115$) further supports that *no* in the embedded clause is never considered a licensor for *ever* in the matrix clause. Overall, the results show that BERT successfully encoded the semantic feature of NPI licensors and structural c-command constraints while it was hardly led into NPI illusion as opposed to human language processors. We conducted post hoc analyses using sequential LSTM-RNN (Jozefowicz et al. 2016), which will be discussed in the paper as well.

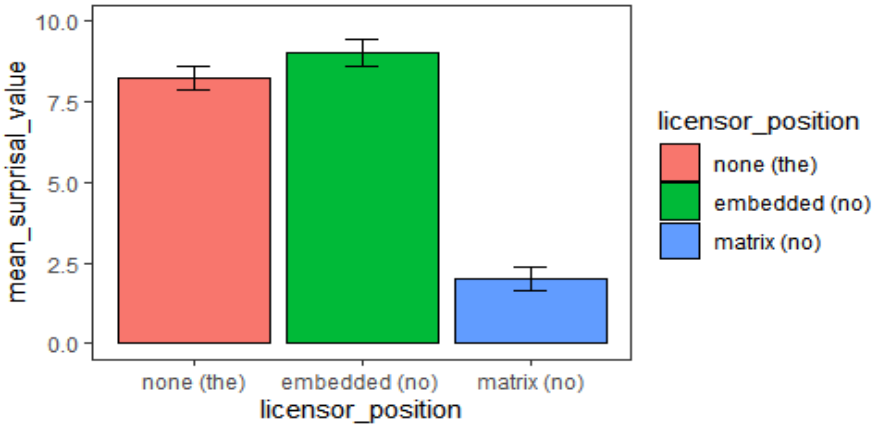
The results of this study suggests that a deep learning language model BERT is fully capable of extracting semantic and syntactic features or constraints required for processing long-distance dependencies such as NPI licensing. However, the fact that BERT is immune to NPI illusion may also suggest that the mechanisms or algorithms BERT relies on in language processing may fundamentally differ from those which humans rely on, e.g. cue-based retrieval, feature-matching, similarity-based analogical reasoning, etc. The current results do not exclude the possibility BERT depends on some surface-related naïve heuristics as well (McCoy et al. 2019). This is, to our knowledge, the first study that investigated BERT's capability in NPI processing and compared its performance between its legitimate licensing and the illusion phenomenon.

Table 1. Example sentence stimuli

Licensor Position	Sentence examples for experiments
Matrix clause	{*no*/*few*/*some*/*most*/*every*} bears [that the competent trainers have treated kindly at all times] have <u>ever</u> gotten out of control.
Embedded clause	The bears [that {*no*/*the*} competent trainers have treated kindly at all times] have <u>ever</u> gotten out of control.



<Figure 1> Experiment 1: Surprisal of five potential licensors in the matrix clause



<Figure 1> Experiment 2: Licensing interactions of the negative quantifier *no* and targeted NPI *ever*.

Reference

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). ACL vol.1, 4171–4186
Hewitt, J. & Manning, C. D. (2019). ACL, vol.1, 4129–4138
Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). arXiv:1602.02410.
McCoy, R. T., Pavlick, E., & Linzen, T. (2019). ACL, vol.1, 3428–3448
Merkx, D. & Frank, S. L. (2020). arXiv:2005.09471.
Smith, N. J. & Levy, R. (2013). *Cognition* 128(3), 302–319
Vasishth, S., Brüssow, S., Lewis, R. & Drenhaus, H. (2008). *Cognitive Science* 32, 685-712.
Warstadt, A. & Bowman, S. (2020). Proceedings of the 42nd Annual Conference of the CSS.
Wilcox, E. Gauthier, J. Hu, J. Qian, P. & Levy, R. (2020) Proceedings of the 42nd Annual CSS.
Xiang, M., Dillon, B., & Phillips, C. (2009). *Brain and Language* 108(1), 40-55.