Do Artificial Language Models Learn Syntax–Semantics Mappings?

Xiaohan (Hannah) Guo, Bryor Snefjella, Idan A. Blank, (UCLA)

Background. Artificial neural networks (ANNs) have recently emerged as successful models of language processing [1,2]. Specifically, these models implicitly learn a surprising amount of syntactic knowledge [e.g., 3,4]. However, ANNs have been criticized for having no semantic knowledge [e.g., 5,6]. Such criticisms often conflate several issues: grounding linguistic meaning in non-linguistic experience; having common sense / world knowledge; and representing semantic relations, such as event structure. Here, we focus on the latter and test whether ANNs implicitly represent "who did what to whom". Because event semantics might be internally represented even if not evident in ANNs' output (e.g., next word prediction; cf. [7]), we study the hidden representations of these networks.

Materials. We borrow our design from fMRI studies in [8,9]. We use a set of "base" items, and edit each item to create several distinct versions ("conditions"). In Experiment 1 (**Table 1**), base items are simple transitive sentences, and are edited to create 4 conditions, differing from the base in: (A) only lexical items (using synonyms), but not syntax or global meaning; (B) only syntax (active vs. passive), but not words or global meaning; (C) only global meaning (switching agent and patient), but not syntax or words (the critical condition); and (D) all 3 aspects (control). In Experiment 2 (**Table 2**), conditions differ from the base in: (A) one synonymous word, not affecting global meaning; (B) one non-synonymous word, changing global meaning; (C) syntax (active vs. passive / direct- vs. prepositional-object) but not meaning; (D) both syntax and meaning (switching agent and patient); or (E) all aspects (control).

Procedure. We evaluated two representative state-of-the-art transformer architectures, BERT [10] and GPT2 [11]. For each sentence, we extracted unit activations from the last hidden (non-embedding) layer (results hold in other layers); the last sentence token was used (BERT: [SEP]; GPT2: '.'; results hold for all-token averages). For each item, we computed cosine similarities between activations for the "base" sentence and each other version (condition). Similarities were Fisher-transformed to improve normality. We compared conditions in terms of similarities to the "base" via a non-parametric, repeated-measures ANOVA based on restricted permutation of residuals [12-13] (results hold under two other permutation regimes). Specifically, pairs of conditions were compared via Tukey tests within this ANOVA model.

Results and discussion. See **Figure 1.** In Experiment 1, two sentences with the same words and syntax but different meaning (switching agent and patient; "base" vs. condition C) were *more* similar to each other than pairs that had the same meaning but differed in either words ("base" vs. A) or syntax ("base" vs. B). Thus, ANNs represent sentences with different event structures as more similar than sentences with the same event structure. In Experiment 2, sentence pairs that differed in both syntax and meaning ("base" vs. D) were *no less* similar than pairs that differed only in syntax but not meaning ("base" vs. C). Thus, a difference in syntax influenced ANNs' representations to a similar extent regardless of whether it led to a change in meaning or not (in contrast, a word changing to a non-synonym had a larger influence than it changing to a synonym, as expected). Overall, the ANNs we studied might be severely limited as models of human language processing: at least in terms of the overall, distributed pattern of activations across hidden units, ANNs fail to represent sentence semantics, even in a test of "bare" event structure divorced from world knowledge or grounding.

Table 1. Experiment 1 sample materials (94 sets for BERT; 92 sets for GPT2)

Base	Different words	Different syntax	Different meaning	Different all
The teacher praised the thinker	(A) The educator lauded the theorist	(B) The thinker was praised by the teacher	(C) The thinker praised the teacher	(D) The educator was lauded by the theorist

Table 2. Experiment 2 sample materials (113 sets for BERT; 106 sets for GPT2)

Base	Different word		Different syntax		Different all
	Mean same	Mean different	Mean same	Mean different	
Anna invited the composer	(A) Anna invited the songwriter	(B) Anna invited the translator	(C) The composer was invited by Anna	(D) Anna was invited by the composer	(E) Anna was invited by the translator

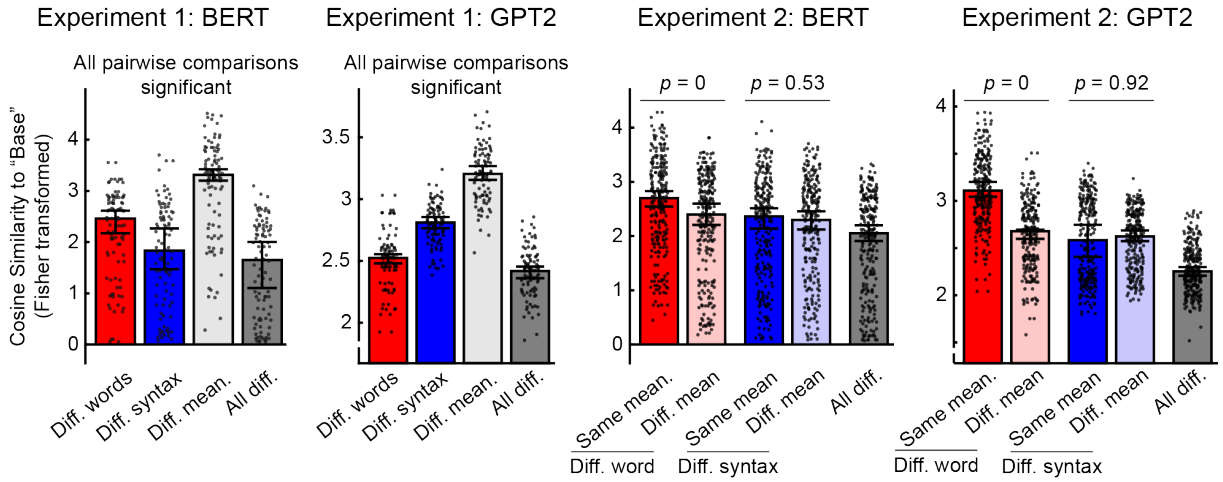

Figure 1. Similarities between sentence pairs. In Experiment 1, note that sentence pairs with *different* meanings (light gray) are more similar to each other, compared to sentence pairs with the *same* meaning but different words (synonyms; red) or syntax (blue). In Experiment 2, note that pairs with different syntax and *different* meanings (light blue) are no less similar to each other compared to pairs with different syntax but the *same* meaning (dark blue). Bars show medians. Error bars show 95% confidence intervals. Dots show individual items.

References. [1] Schrimpf et al. (2020). *bioRxiv preprint*. **[2]** Hu et al. (2020). *ACL*. **[3]** Rogers et al. (2020). *ACL.* **[4]** Manning et al. (2020). *PNAS*. **[5]** Bender & Koller (2020). *ACL*. **[6]** Marcus (2020). *ArXiv preprint*. **[7]** Ettinger (2020). *TACL*. **[8]** Fedorenko et al. (2020). *Cognition*. **[9]** Dapretto & Bookheimer (1999). *Neuron*. **[10]** Devlin et al. (2018). *ArXiv preprint*. **[11]** Radford et al. (2019). OpenAI blog. **[12]** Kherad-Pajouh & Renaud (2015). *Statistical Papers*. **[13]** Anderson & Braak (2003). *JSCS*.