Abstraction via exemplars? A representational case study on lexical category inference in BERT

Kanishka Misra[τ,π] and Najoung Kim[β]

[τ]The University of Texas at Austin, [π]Purdue University, [β]Boston University

Linguistic abstractions are often posited as the mechanism by which language learners generalize to novel expressions. However, Ambridge [1] argues that such abilities arise not through linguistic abstraction but instead via stored exemplars encountered during learning. One counterargument [2], among many [3], takes inspiration from the unprecedented success of neural network language models (LMs) such as BERT [4]. Specifically, LMs deviate from both the radical exemplar account and the pure abstraction-based account—their parameters store a summarized form of their training exemplars as guided by their training objective (word-prediction in context), rather than storing individual exemplars or encoding abstractions explicitly. Empirically, this has shown to give rise to generalization behaviors aligning with abstract linguistic structures [5]. Abstraction could therefore be viewed as arising *via* the (compressed) encoding of exemplars.

In this work, we provide further empirical evidence for the *abstraction-via-exemplars* account, by presenting a case study on lexical category inference in BERT, an LM that is trained to predict missing words in fill-in-the-blank (FITB) inputs (e.g., *I ___ you.* → *"like"*). We adapt the experiments of Kim and Smolensky [6, hereafter K&S], who tested BERT's generalization of the usage of novel tokens belonging to either NOUN/VERB/ADJ/ADVERB, from exposure to a single observation Their tests are inspired by developmental linguistic studies that probe for lexical category abstraction in children [7]. Briefly, K&S used pairs (see Figure 1) of FITB stimuli that disambiguate the category of the novel token and trained the model's embeddings of the novel tokens only (initialized with random-values), keeping all other parameters constant. They then tested BERT's accuracy for predicting each of the novel tokens on a set of FITB inputs (N=100 per category) that are lexically disjoint from the training examples, and found overall above-chance performance indicating desirable generalization patterns. Using representations learned by this method, we perform two analyses that relate the representational change in the novel tokens' embeddings with BERT's generalization patterns on the K&S experiment. First, we track the movement of the embeddings in two-dimensional space (obtained using Principal Component Analysis) as they are updated during training. Figure 2 suggests that the final states of the embeddings of the novel tokens move closer in two-dimensional space to that of known, unambiguous category exemplars (N=500 per category). That is, regardless of initialization, *wug* (Figure 1) becomes more noun-like, while *dax* becomes more adjective-like. To further analyze this behavior, we take multiple novel tokens (N=20 per category) and assign them values randomly sampled from regions in two-dimensional space populated by category exemplars. We then project them into BERT's embedding space, yielding novel token representations with values informed by two-dimensional encoding of category regions. Evaluating BERT to predict these novel tokens on the K&S testing set—without any additional training—yields substantially above-chance performance across multiple category pairs (Table 1).

Together, these analyses suggest that representational movement towards regions of known category exemplars can result in behavior that is consistent with successful generalization to novel expressions. Overall, this lends empirical support to the claim that abstraction and exemplar based accounts need not be at odds, and that abstraction-like behavior can emerge in learners that encode training exemplars.

(498 words)

1. **Training**
   a. I saw a *wug* run (N-context)
   b. that book was very *dax*! (Adj-context)

2. **Testing**
   a. they can find the ___. (N-context)
   b. it is quite ___ of you. (Adj-context)

Evaluation

$$P_{\text{BERT}}(wug \mid \text{N-context}) > P_{\text{BERT}}(dax \mid \text{N-context})$$
$$\&$$
$$P_{\text{BERT}}(dax \mid \text{Adj-context}) > P_{\text{BERT}}(wug \mid \text{Adj-context})$$

Figure 1: Experimental setting proposed by Kim and Smolensky [6], illustrated with NOUN vs. ADJ.
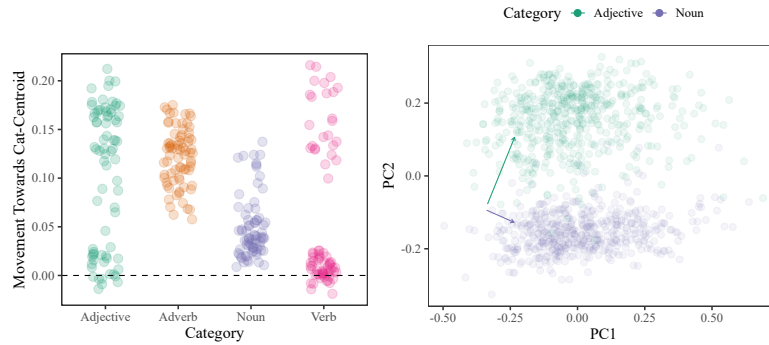

Figure 2: **Left:** Relative movement of the novel token representations with respect to known category exemplars for each category after training on the K&S experiments. 0.0 indicates no movement. **Right:** Movement (indicated by arrow) of a novel token's two-dimensional representation from its initial state in the ADJ–NOUN experiment. Points indicate known, unambiguous adjectives and nouns.

Category Pair	Accuracy
ADJ–ADVERB	$0.93_{\pm 0.03}$
ADJ–VERB	$0.70_{\pm 0.06}$
ADVERB–VERB	$0.87_{\pm 0.05}$
NOUN–ADJ	$0.80_{\pm 0.08}$
NOUN–ADVERB	$0.89_{\pm 0.04}$
NOUN–VERB	$0.81_{\pm 0.08}$

Table 1: Accuracies (with 95% CI) on the test set of K&S [6] obtained by randomly sampling values from two-dimensional regions of category-exemplars (e.g., see Fig 1-Right) which are projected to serve as BERT embeddings for novel, unseen tokens (N=20 each). **Chance performance is 0.50**.

References

[1] Ben Ambridge. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559, 2020.

[2] Kyle Mahowald, George Kachergis, and Michael C Frank. What counts as an exemplar model, anyway? A commentary on Ambridge (2020). *First Language*, 40(5-6):608–611, 2020.

[3] Ben Ambridge. Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language*, 40(5-6):640–659, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *NAACL 2019*, pages 4171–4186, 2019.

[5] Yoav Goldberg. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

[6] Najoung Kim and Paul Smolensky. Testing for grammatical category abstraction in neural language models. *Proceedings of the Society for Computation in Linguistics*, 4(1):467–470, 2021.

[7] Barbara Höhle, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, 5(3):341–353, 2004.