Abstraction via Exemplars? A Representational Case Study on Lexical Category Inference in BERT

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

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

kmisra@utexas.edu najoung@bu.edu

Research Questions

- Are abstraction and exemplar accounts of linguistic generalization necessarily at odds?

  Answer: *Not necessarily! Pre-trained language models can demonstrate generalization to novel linguistic expressions while being compatible with both accounts.*

- **Case Study RQ:** How do pre-trained language models perform lexical category-membership inference (N/J/V/ADV) of novel tokens from exposure to a single observation?

  Answer: *By facilitating movement towards category-specific regions within representational space.*

Behavioral results from replicating K&S

- **Model:** `bert-large-uncased-whole-word-masking` [2]
  - Used the tokens `[unused1]`–`[unused994]` in the model's vocabulary to represent the novel words.
  - Froze the entire model except for the embeddings of the two words being learned from context and trained for 70 epochs
- **Stimuli:**
  - **Source:** Sentences sampled from MNLI [5] – a dataset that the BERT model has not encountered in training.
  - **Train set:** Pairs of single-sentence exemplars.
  - **Validation and Test sets:** 200 sentence-pairs per category, obeying design constraints set by K&S.

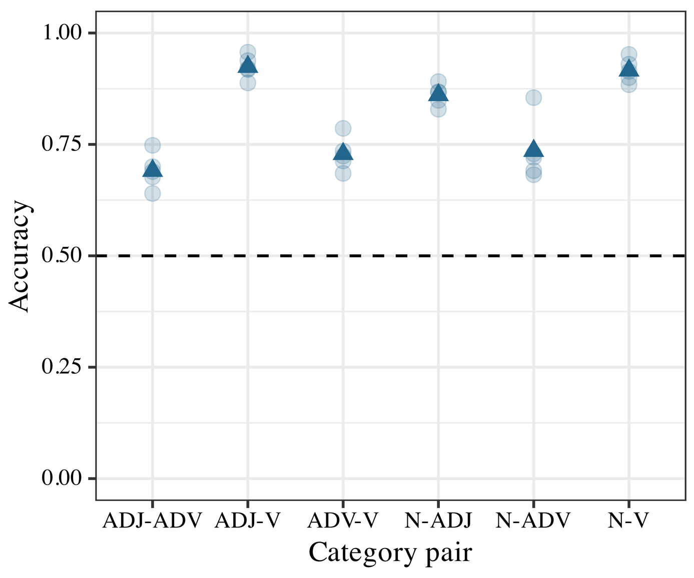


Figure 1. Results from replicating Kim and Smolensky [4]. Triangles represent mean accuracies across five runs (shown as circles), each of which uses different pairs of novel tokens. **Chance performance is 50%.**

Does observing this behavior entail abstractions?

- Abstractions are sufficient but not necessary to give rise to the observed behavior.
- Non-zero chance that the model could simply be analogizing to a single exemplar (I saw a fluffy **wug**. → `wug = cat`)
- **What drives the model's generalization?** We turn to representational analyses to answer this!

Conclusions

- In BERT, there exist parts of the embedding space that license category-conforming predictions near the centroid of known members.
- BERT does **not** explicitly store individual training exemplars (only sophisticated summary representations in the form of type-level embeddings).
- BERT also does **not** explicitly store the particular abstractions that we were testing for; they manifest in the form of regions in the embedding space the aforementioned summary representations live in.
- **Abstraction-consistent generalization behaviors can emerge in learners that do not store abstractions nor individual training exemplars explicitly.**

Aside: Relation to prototype theory?

- There are no explicit prototypes stored for `[noun]`, `[verb]`, etc.—they are emergent! But BERT does have one type of summary representation: its **embeddings**!
- Prototypes of different categories, or at different levels are seem to be (recursively) computed on-the-fly if one level of summary representations are available.
- **Q: What are the right level(s) of granularity that can sufficiently enable generalization?**

The nature of linguistic knowledge and generalization

Pure Abstraction View

Generalizations are facilitated using *stored abstractions*

Radical Exemplars View

No stored abstractions, only *stored exemplars.*

Generalization = on-the-fly analogy across exemplars

Abstraction-via-exemplars?

Abstraction vs. radical exemplars could be a false dichotomy (see Ambridge [1] and responses)

Compressed encoding of exemplars could lead to **emergent abstraction-like structures and behaviors!**

Case in point: Neural Network Language Models!

Our work: Contributes further evidence for the Abstraction-via-exemplars view by presenting a case-study on category membership inference for novel words!

Measuring movement behavior in representation space

What is the behavior of the novel token representations as they are updated on the single-exposure contexts?

Analysis:

- Track the movement of the embeddings in two-dimensional space (obtained using Principal Component Analysis) as they are updated during training.
- **Results:** Final states of the embeddings of the novel tokens **move closer in two-dimensional space to centroids of regions occupied by known, unambiguous category exemplars** (N=500 per category).

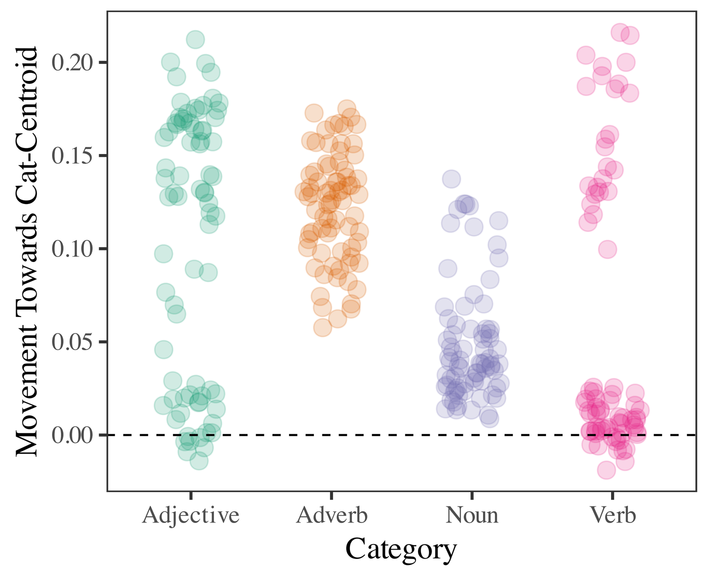


Figure 2. Relative movement of the novel token representations with respect to known category exemplars for each category after training on the K&S experiments. 0.0 indicates no movement.

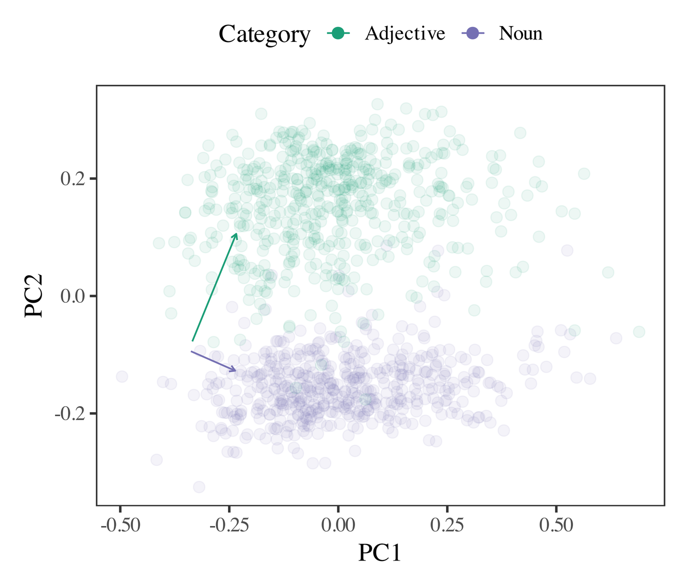


Figure 3. Average movement (indicated by arrows) of a novel token's two-dimensional representation from its initial state in the ADJ–NOUN experiment. Points indicate known, unambiguous adjectives, and verbs.

Future work

How can this paradigm and analysis toolkit be used to answer theoretically significant questions in human language processing and acquisition?

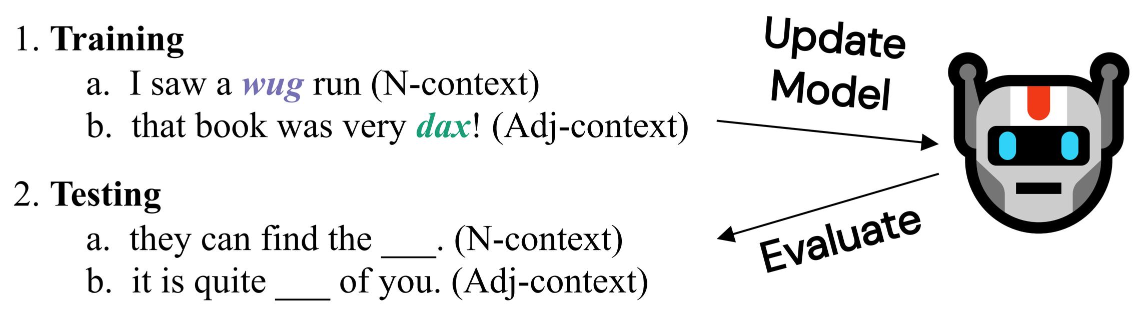
- **New human (and LM) studies:** E.g., an adaptation of the behavioral method we used for LMs to test finer-grained categories in adults (animacy of a noun, verbs prone to dative alternation, etc.)
- **Addressing "what is in the data" questions:** training a model on a developmentally plausible data to test the extent to which there is sufficient information to support the emergence of the target abstractions.

Case Study: Kim and Smolensky (2021)

Target task: Inferring lexical categories (in particular, part-of-speech) of **novel words** from context and making generalizations about them in novel contexts, motivated by an existing infant study involving the head-turn preference paradigm [3].

Method:

- Expose a pre-trained LM to **single contexts containing novel words, where the lexical category of the novel words is unambiguous.**
- Only update the embeddings of novel words, keeping rest of the model frozen.
- Test on unseen test contexts with no lexical overlap with training set, where target words appear in different linear positions.
- **Can novel words be placed in a space that elicits behavior consistent with abstraction over lexical categories?**



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 4. Experimental setting proposed by Kim and Smolensky [4], illustrated with NOUN vs. ADJ.

Investigating latent category-specific regions

How well do category-specific regions lead to abstraction-consistent behavior?

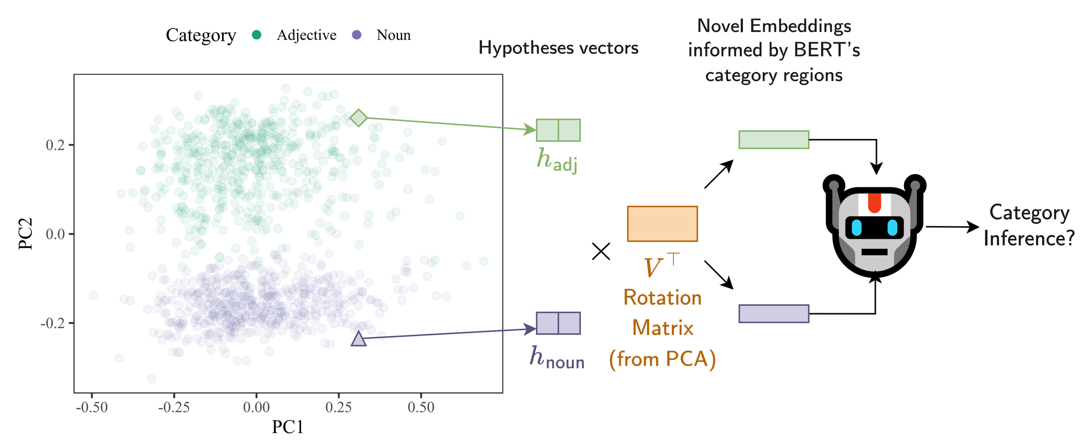


Figure 5. Overview of our method to analyze category-specific regions in BERT. We sample hypotheses vectors from gaussian distributions centered around 2D category regions, project them into BERT's embedding space, and then evaluate on the K&S test set.

Results: Substantially above-chance performance across all category-pairs, obtained without any additional training of the category-informed novel token representations!

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 Kim and Smolensky [4] obtained by randomly sampling values from two-dimensional regions of category-exemplars which are projected to serve as BERT embeddings for novel, unseen tokens (N=20 each). **Chance performance is 0.50.**

There are continuous regions that license category-conforming predictions!

References

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

[2] 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.

[3] 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.

[4] Najoung Kim and Paul Smolensky. Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 467–470, February 2021.

[5] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL 2018*, pages 1112–1122, New Orleans, Louisiana, 2018.