COGS: A Compositional Generalization Challenge Based on Semantic Interpretation

Najoung Kim[1] Tal Linzen[2]
[1]Department of Cognitive Science, Johns Hopkins University [2]Department of Linguistics, New York University

Motivation

- Humans can interpret expressions that they have never encountered before, by **composing** the the meanings of the parts that they already know (Montague, 1974)
- Current NLP models fall short on this ability---they suffer from poor **out-of-domain** generalization performance
- Existing benchmarks for compositional generalization have limited linguistic expressivity (e.g., Lake and Baroni 2018, Keysers et al. 2020)

Research questions

- Can popular neural models generalize compositionally like humans do on a semantic interpretation task? (A: Not really)
- Is there a difference between between lexical and structural generalization? (A: Yes, models find structural generalization more difficult overall)

Testing Generalization with Systematic Gaps

Task

Sequence-to-sequence semantic parsing
(English sentence → logical form)

INPUT: *John ate the cookie*
OUTPUT: $*\text{cookie}(x_3)$; $\text{eat.agent}(x_1, \text{John}) \wedge \text{eat.theme}(x_1, x_3)$

Examples of systematic gaps (inspired by humans)

- Subject → Object: NPs seen as both subject and object during training, but noun 'hedgehog' only seen as subject. e.g.,

Train: {*A hedgehog saw John, John liked the cake, The baby ran*}
Dev/Test: *The hedgehog saw the cake*
Gen.: *The baby liked the hedgehog*

- Deeper recursion: only depth n sentential complement embedding seen during training. e.g.,

Train: {*Emma said that Noah knew that the cat danced, The queen walked*}
Dev/Test: *The queen knew that the cat said that Noah walked*
Gen.: *Emma said that Noah said that the cat knew that the queen walked*

Experiment Setup

Dataset

- 30K unique sentences sampled from a PCFG and split into 80:10:10 (train:dev:test)
- Sentences with controlled distribution (e.g., `hedgehog' only as subject) added to the training set
- Out-of-distribution generalization set separately sampled from different PCFGs (21K sentences for 21 generalization types)
- Metric: exact match accuracy

Model

- LSTM, biLSTM, Transformer
- Comparable # of parameters in each model: Transformer (9.5M), BiLSTM (10M), LSTM (11M)
- Not pretrained; trained from scratch on COGS only

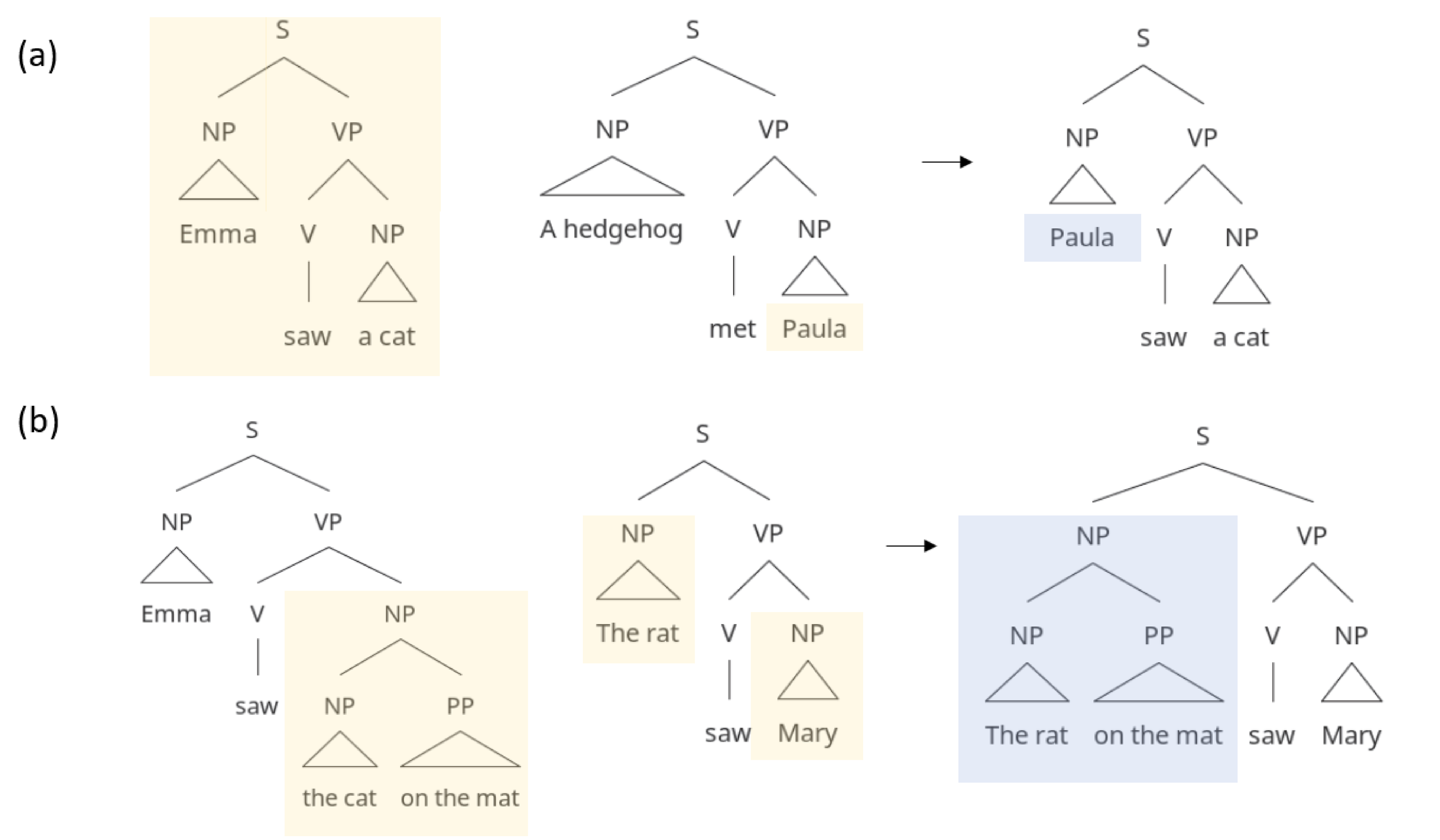
Lexical and Structural Generalization

(a) Lexical generalization:

A novel composition of a known structure and a known lexical item

(b) Structural generalization:

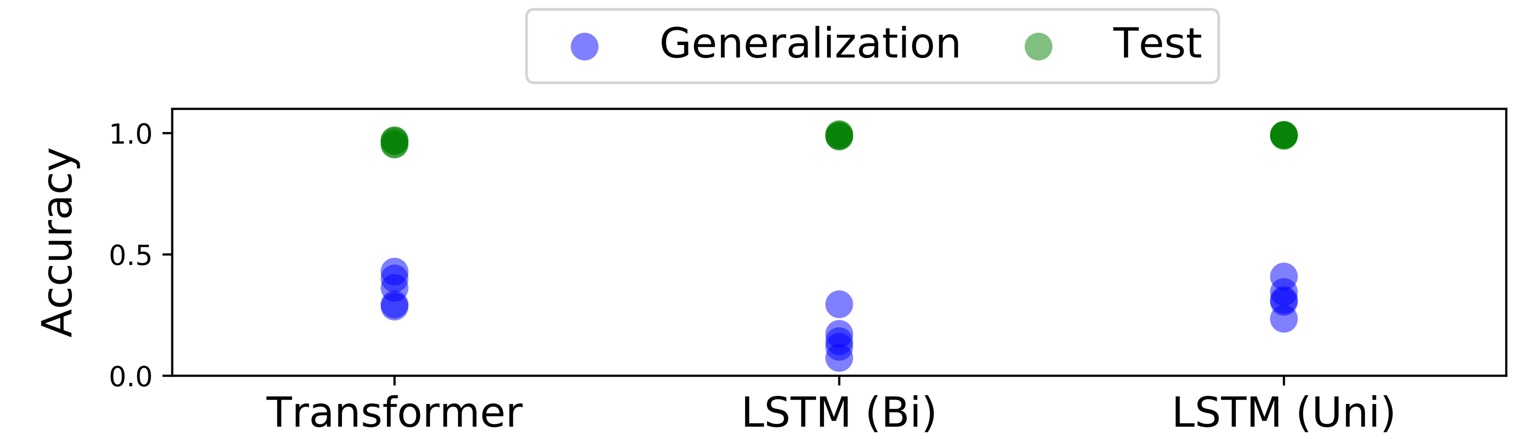
A novel composition of two known structures



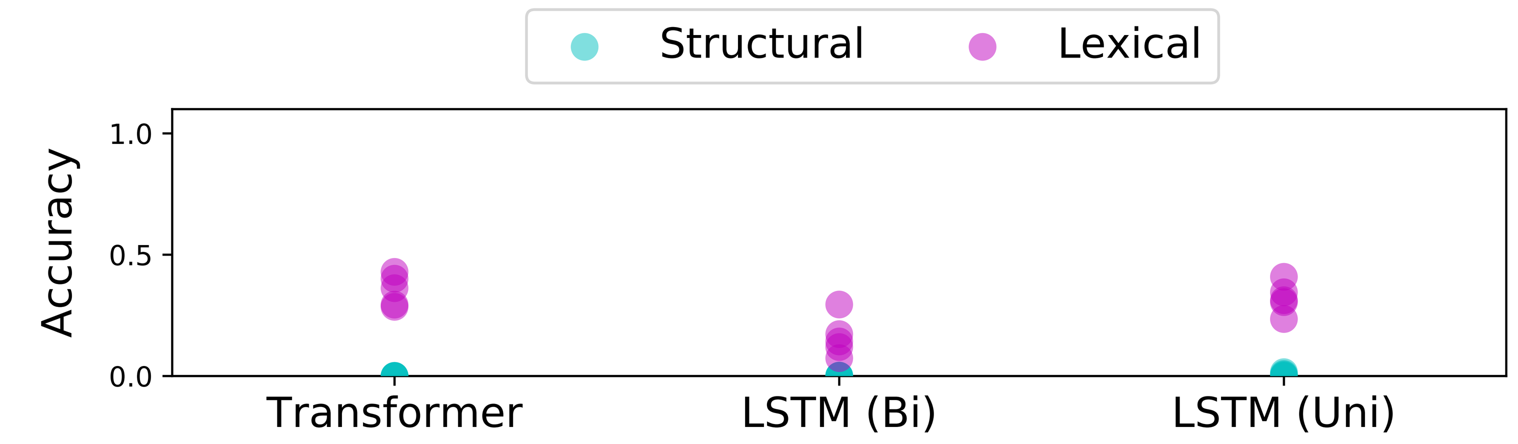
Discussion

1. What would be needed to solve COGS?
2. How to test for *constraints* on generalization? (e.g., not all dative verbs alternate)
3. Empirical support for structural generalization in humans?

Overall Results

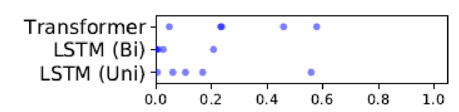
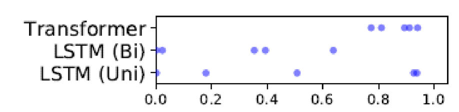
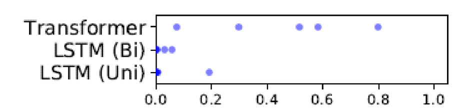
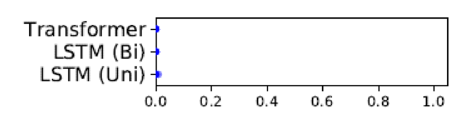
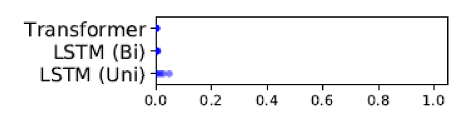
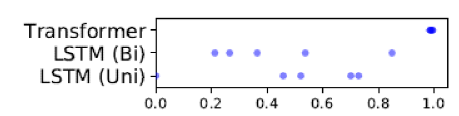


All models performed almost perfectly on the in-domain test set. On the other hand, they performed poorly on the generalization set with systematic gaps.



Performances on cases of structural generalization were especially poor.

Selected Generalization Cases and Performance

Case	Training	Generalization	Accuracy Distribution
Subject → Object (common noun)	*Subject* A **hedgehog** ate the cake.	*Object* The baby liked the **hedgehog**.	
Object → Subject (common noun)	*Object* Henry liked a **cockroach**.	*Subject* The **cockroach** ate the bat.	
Object → Subject (proper noun)	*Object* Mary saw **Charlie**.	*Subject* **Charlie** ate a donut.	
Primitive → Object (proper noun)	*Primitive* **Paula**	*Object* The child helped **Paula**.	
Depth generalization: PP modifiers	*Depth 2* Ava saw the ball **in the bottle on the table**.	*Depth 3* Ava saw the ball **in the bottle on the table on the floor**.	
Active → Passive	*Active* Emma **blessed** William.	*Passive* A child **was blessed**.	

References

Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., Tsarkov, D., Wang, X., van Zee, M., and Bousquet, O. (2020). Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Lake, B. and Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2879--2888.

Montague, R. (1974). English as a formal language. In *Formal Philosophy. Selected papers by Richard Montague*.

Image: cogs by Gregor Cresnar from the Noun Project