COMPOSITIONAL LINGUISTIC GENERALIZATION IN ARTIFICIAL NEURAL NETWORKS

NAJOUNG KIM

(NYU CENTER FOR DATA SCIENCE, BU LINGUISTICS)

02/15/2022

AGENDA

1. Background
2. A test for compositional linguistic generalization
3. Learning biases for compositional linguistic generalization
4. Conclusion & Future work

Compositional Linguistic Generalization

- Producing/comprehending linguistic expressions that have not been encountered before (“generalization”)
- …by composing known constituents

KNOWN:

kicker kicked

```
{The suspect kicked the victim → kick(s, v)
 The suspect kicked the man    → kick(s, m)
 The thief kicked the victim   → kick(t, v)
 The man kicked the crog       → kick(m, c)
}
```

The crog kicked the victim → kick(c, v) Compositional

→ kick(m, c) Any sentence with ‘crog’ has this meaning

→ kick(s, v) Take the most frequently observed kicker

Compositional Linguistic Generalization

- Producing/comprehending linguistic expressions that have not been encountered before (“generalization”)
- …by composing known constituents

KNOWN:

kicker kicked

```
{The suspect kicked the victim → kick(s, v)
 The suspect kicked the man    → kick(s, m)
 The thief kicked the victim   → kick(t, v)
 The man kicked the crog       → kick(m, c)
}
```

The crog kicked the victim → kick(c, v) Compositional according to Theory 1

Theory 1: $NP_1\ V_{tr}\ NP_2 \rightarrow d(V_{tr})(d(NP_1), d(NP_2))$

Compositional Linguistic Generalization

- Producing/comprehending linguistic expressions that have not been encountered before (“generalization”)
- …by composing known constituents

KNOWN:

{The suspect kicked the victim → kick(s, v)
 The suspect kicked the man → kick(s, m)
 The thief kicked the victim → kick(s, t)
 The victim kicked the bucket → die(v)
}

Theory 1: $NP_1\ V_{tr}\ NP_2 \rightarrow d(V_{tr})(d(NP_1), d(NP_2))$

The suspect kicked the bucket → kick(s, b)

Compositional generalization can only derive the literal meaning under Theory 1

Compositional Linguistic Generalization

- Producing/comprehending linguistic expressions that have not been encountered before (“generalization”)
- …by composing known constituents

KNOWN:

{The suspect kicked the victim → kick(s, v)
 The suspect kicked the man → kick(s, m)
 The thief kicked the victim → kick(s, t)
 The victim kicked the bucket → die(v)
}

Theory 2: $NP_1\ V_{tr}\ NP_2 \rightarrow d(V_{tr})(d(NP_1), d(NP_2))$

but if $V_{tr} = \text{kicked}$ and $NP_2 = \text{the bucket} \rightarrow \text{die}(\text{d}(\text{NP}_1))$

The suspect kicked the bucket → die(s)

The suspect kicked the bucket → kick(s, b)

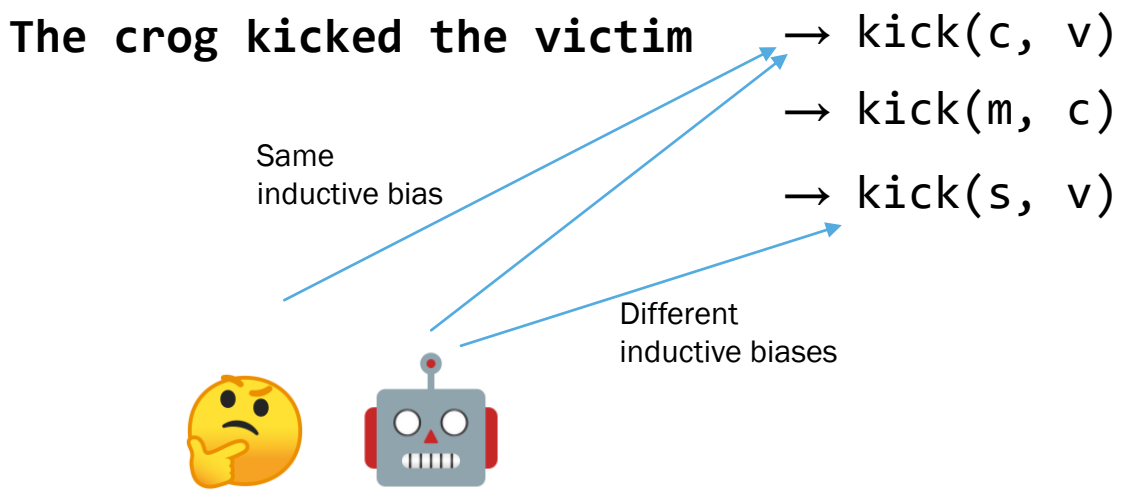
Compositional generalization includes idiomatic meaning under Theory 2

Compositional Linguistic Generalization

- Producing/comprehending linguistic expressions that have not been encountered before (**“generalization”**)
- …by **composing** known constituents
- How composition works is theory dependent! (predicts different generalizations)
- Thus, “compositional linguistic generalization” for a given expression crucially depends on the particular compositional theory of language assumed

Inductive Bias (Learning Bias)

- Determines how a learner generalizes to unseen inputs



Compositional

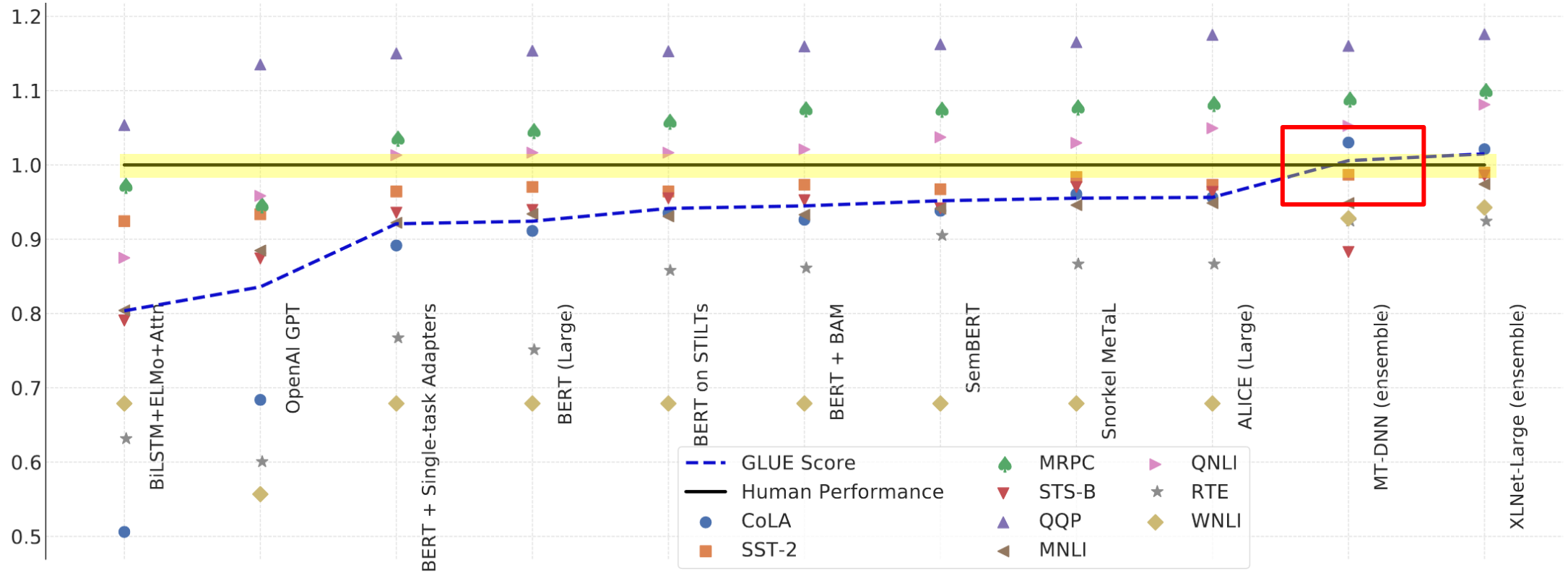
Any sentence with ‘crog’ has this meaning

Take the most frequently observed agent

“Poverty of the stimulus” experimental paradigm (Wilson 2006) teases apart learners’ inductive biases

Artificial Neural Networks (ANNs): The Success Story

- Approaching “human performance (?)” in many tasks that require language understanding



Wang et al. (2019)

Artificial Neural Networks (ANNs): The Success Story

- Approaching “human performance (?)” in many linguistic evaluation tasks

What did John fry?
*What did John fry the potato? (Warstadt et al. 2019)

The farmer that the parents love swims
*The farmer that the parents love swim (Marvin and Linzen 2018)

- Growing interest in analyzing the linguistic capacity of ANNs!

Compositional Linguistic Generalization in ANN Learners

- The consensus seems to be that (vanilla, general-purpose) ANN learners do not demonstrate robust compositional generalization that humans are capable of

- But existing tests for compositional generalization are limited (in terms of measuring linguistic generalization):

    - e.g., SCAN (Lake and Baroni 2018)

        - Given {*run* -> RUN, *jump* -> *JUMP*, *jump twice* -> JUMP JUMP}, what is *run twice*?
        - Great task, but limited expressivity (no way to express predicate-argument structure)

- One contribution: proposing a task that measures compositional linguistic generalization (that humans are able to make)

- Why do we care?

This Work

- This work: comparing the inductive biases of ANN and human learners on assigning meaning representations to novel complex expressions

- Outcome 1: Their inductive biases align 🤔 🟰 🤖
  - Useful as a demonstration of how compositional capacity can be implemented in a distributed system
  - Useful as a model that has the same inductive bias as humans that we can study with more freedom in possible experimental manipulations
  - Implications for Artificial Intelligence

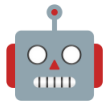
This Work

- This work: comparing the inductive biases of ANN and human learners on assigning meaning representations to novel complex expressions

- Outcome 2: Their inductive biases do not align 🤔 ≠ 🤖
  - New question: what factors contribute to changing the inductive bias of the models to more closely match that of humans?
  - The answer(s) to this new question may inform us about the ways in which compositional generalization arises in an intelligent system
  - + AI implications

This Work

- This work: comparing the inductive biases of ANN and human learners on assigning meaning representations to novel complex expressions

- Primarily a study of machine cognition 
- …using the findings/tools from (human) cognitive science
- …that will help understand better and improve AI
- …that may inspire future human subject studies (“animal models” analogy from McCloskey 1991)

AGENDA

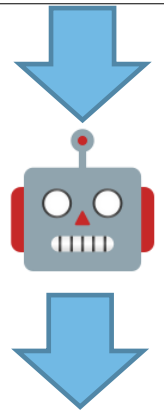
1. Background
2. A test for compositional linguistic generalization
3. Learning biases for compositional linguistic generalization
4. Conclusion & Future work

A Test for Compositional Linguistic Generalization

- Meaning representation assignment (translation): assigning a logical form to a sequence of words
- Test format: sequence mapping (sentences to logical forms)

The compositional theory adopted is made explicit by defining syntax, semantics, and the mapping rule

INPUT: A cat danced.



OUTPUT: $\text{cat}(x_1)\ AND\ \text{dance.agent}(x_2, x_1)$

S -> NP VP
NP -> Det N
…

d(NP VP) = d(NP) AND
d(V).agent(x_j, x_i)
d(a cat) = cat(x_i)
…..

A Test for Compositional Linguistic Generalization

- The training set contains various **systematic gaps**: certain patterns are **withheld** during training
- The "out-of-distribution" generalization set contains the **withheld** examples that can be correctly mapped onto their meaning representations by composing parts available in the training data
- For example…

```
TRAINING:  {A hedgehog ate the cake,
            Alex danced,
            The butterfly saw a flower}

GENERALIZATION: Alex saw a hedgehog
```

"a hedgehog" never seen as grammatical object during training

Examples like "A hedgehog ate the cake" → "Exposure examples" (examples with restricted distribution of primitives)

A Test for Compositional Linguistic Generalization

- The training set contains various **systematic gaps**: certain patterns are **withheld** during training
- There is also an “in-distribution” generalization set that consists of novel sentences that **don’t** pertain to the withheld patterns
- For example…

```
TRAINING:  {A hedgehog ate the cake,
            Alex danced,
            The butterfly saw a flower}

GENERALIZATION: Alex saw the cake
```

“Alex”, “saw”, “the cake” all seen in the same syntactic position during training, although the combination is novel

A Test for Compositional Linguistic Generalization

- 21 different generalization cases—ask me about them later!
- Generalization cases are inspired by generalizations human learners can make, according to the theoretical and developmental linguistics literature
- e.g., Children as young as 20 months old display subject -> object generalization (Tomasello and Olguin 1993)

- 24K training examples, 3K “in-distribution” generalization examples, 21K “out-of-distribution” generalization examples

Types of Generalization

- Can be broadly divided into two categories: **lexical and structural**
- **Lexical generalization:** assigning meaning representations to novel combinations of familiar primitives and familiar syntactic configurations

```
TRAINING:   {A hedgehog ate the cake,
             Alex danced,
             The butterfly saw a flower}

GENERALIZATION: Alex saw a hedgehog
```

1 exposure example / lexical generalization case

- "hedgehog" is familiar, transitive VPs with NP subject/objects are familiar
- But "hedgehog" in the object NP position is novel

Types of Generalization

- Can be broadly divided into two categories: lexical and structural
- Structural generalization: assigning meaning representations to known primitives in novel syntactic configurations

```
TRAINING:  {Luke danced,
            The cat danced,
            Noah knew that the cat danced,
            Emma said that Noah knew that the cat danced}

GENERALIZATION: Luke said that Noah knew that Emma saw that the cat danced.
```

- Generalization example is structurally novel because a depth 3 embedded CP is not in the training set

Types of Generalization

- Can be broadly divided into two categories: lexical and structural
- Structural generalization: assigning meaning representations to known primitives in novel syntactic configurations

```
TRAINING:  {The cat saw the rat on the mat,
            The girl liked the cat on the table}

GENERALIZATION: The cat on the table saw the rat
```

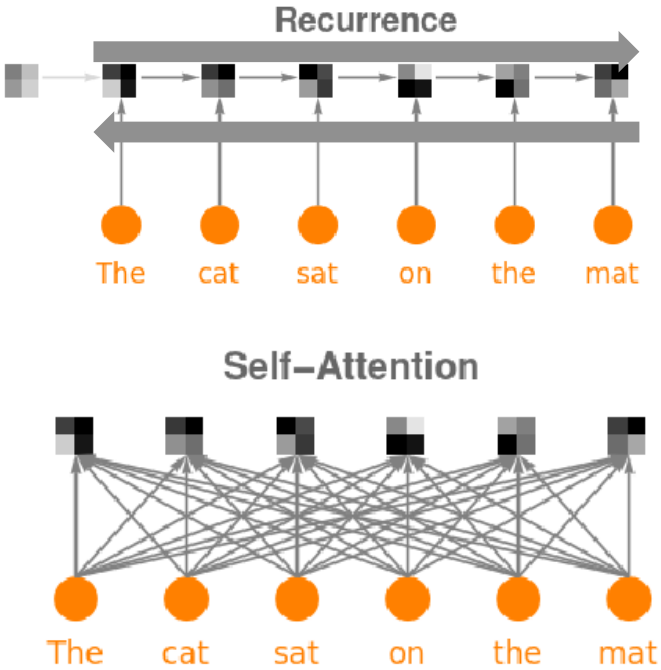
- Generalization example is structurally novel because PP modification in the subject position is not in the training set

Models & Training

- Models: Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997),

  bidirectional LSTM,

  Transformer (Vaswani et al. 2017)

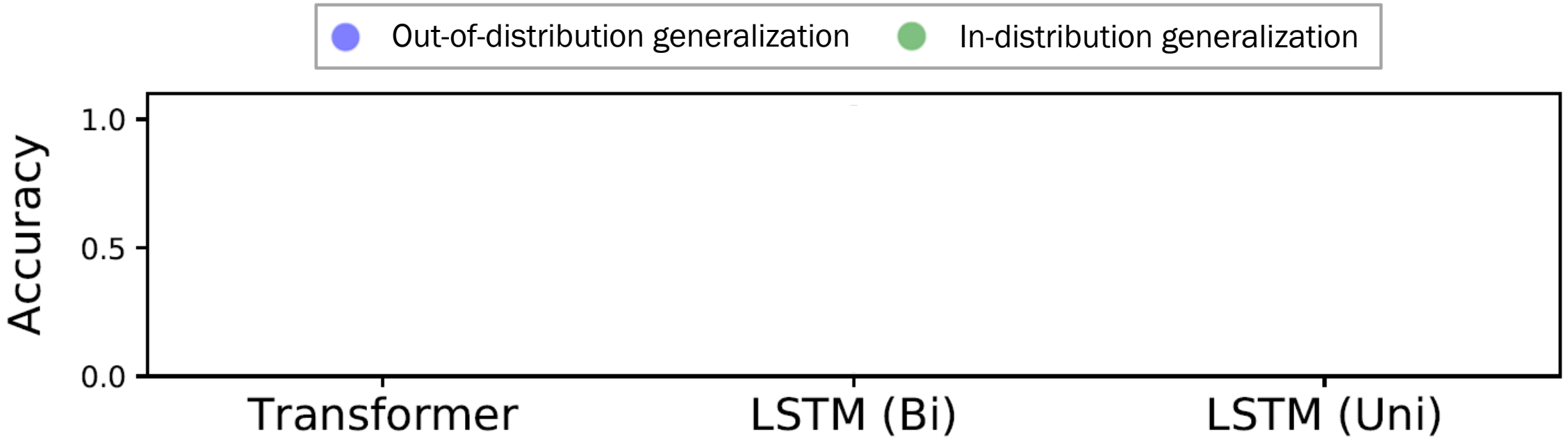


Models & Training

- Models: LSTM (Hochreiter and Schmidhuber 1997), bidirectional LSTM, Transformer (Vaswani et al. 2017)
- Comparable # of trainable parameters in each model:
  Transformer (9.5M), biLSTM (10M), LSTM (11M)

- Trained from scratch on the training set only
- Each model trained five times with different random seeds (determines initial weights / order of training examples)
- Metric: exact string match accuracy

Results



- Overall low out-of-distribution generalization accuracy, despite high in-distribution accuracy
- High variation over random seeds

Error Analysis: Single Lexical Retrieval Errors

- Prominent pattern: “Single lexical retrieval errors”
  - Correct output structure, but misretrieved a denotation for a single lexical item in the **input**

INPUT: A frog burned Sophia.

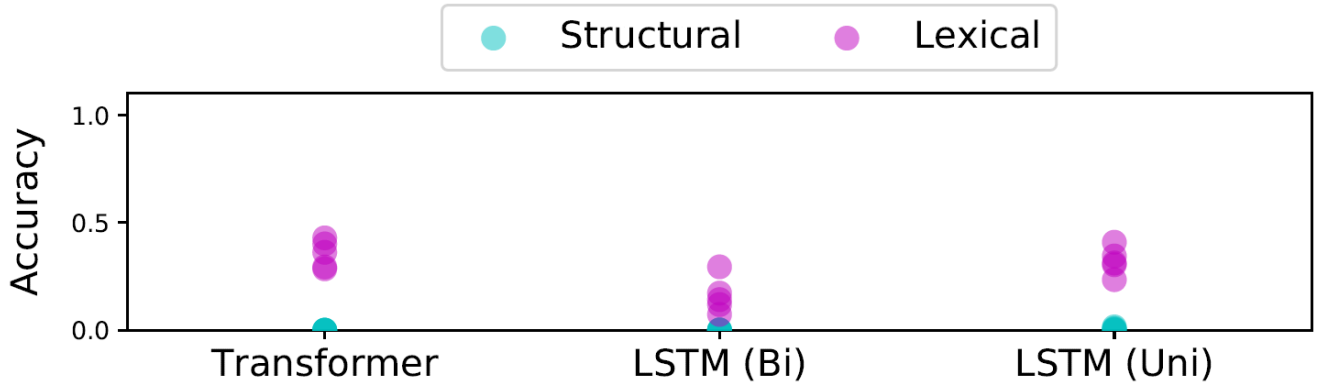
TARGET: frog(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

ERROR: director(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

- Single lexical retrieval errors account for 17% / 43% / 56% of total Transformer / LSTM / BiLSTM predictions

(Almost) Complete Failure on Structural Generalization

- All models were unsuccessful in translating novel structures



- Unidirectional LSTMs achieved very marginal success (1% structural generalization accuracy)
  - LSTMs also produced errors closer to target outputs (token edit distance to target outputs: 11 and 14 for bidirectional / unidirectional LSTM, respectively, and 42 for Transformer)

Intermediate Takeaway

- The learning biases of the models tested (note, not a general claim about the architecture) are different from human learners—they show different generalization patterns
- Translating novel structures is especially challenging
- For lexical generalization, the erroneous outputs are often “single lexical retrieval errors”

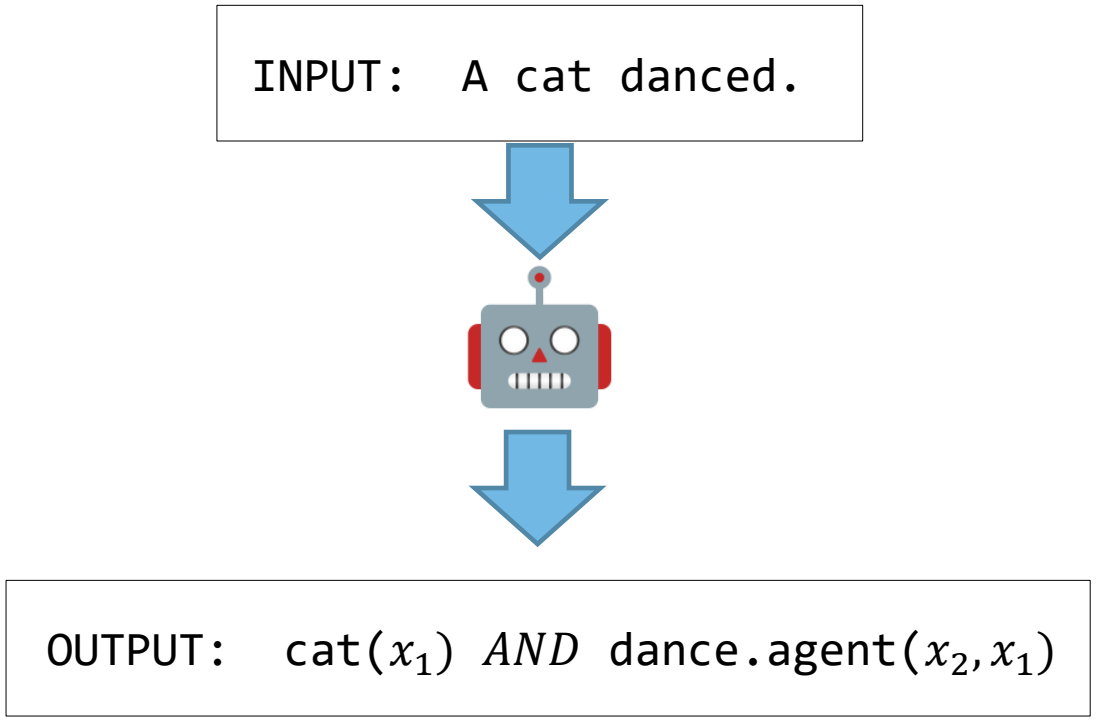
What can we do?

AGENDA

1. Background
2. A test for compositional linguistic generalization
3. Learning biases for compositional linguistic generalization
4. Conclusion & Future work

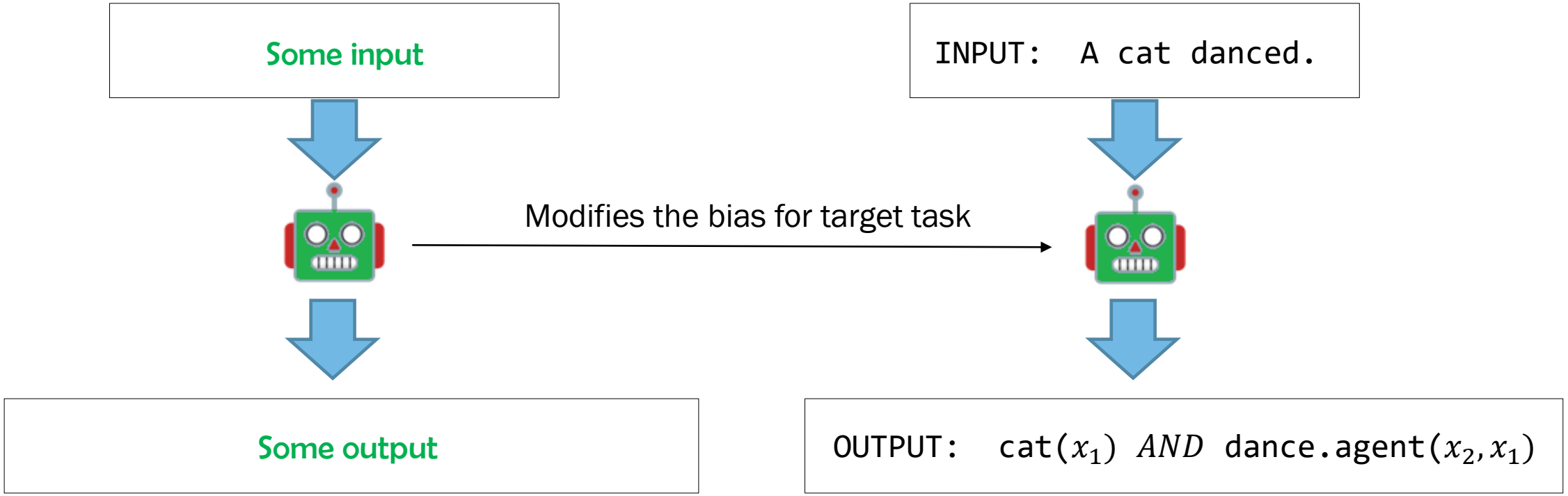
What Can We Do to Modify the Learning Biases of the Models?

- Transfer learning from auxiliary tasks is a promising approach (Caruana 1997, Zhang et al. 2014, Meyerson and Miikkulainen 2018, Peters et al. 2018, Devlin et al. 2019, A LOT of current NLP!)



What Can We Do to Modify the Learning Biases of the Models?

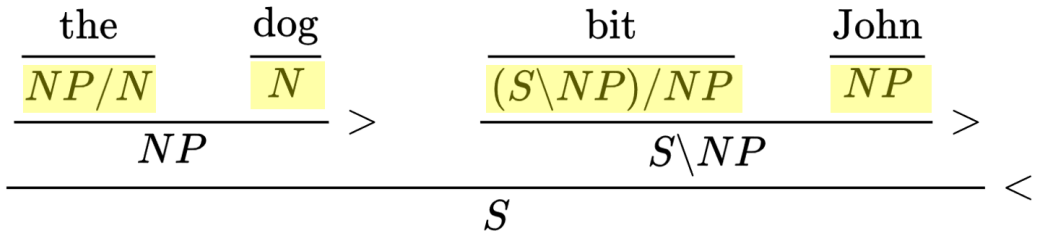
- Transfer learning from auxiliary tasks is a promising approach (Caruana 1997, Zhang et al. 2014, Meyerson and Miikkulainen 2018, Peters et al. 2018, Devlin et al. 2019, A LOT of current NLP!)



Experiment 1: Transfer Learning from Auxiliary Tasks

- Compared three different auxiliary tasks

Auxiliary Task 1: CCG Supertagging



- CCG supertagging (Bangalore and Joshi 1998): tagging task that is informative of combinatorial, phrase-structural constraints
- The tags are informative of how adjacent constituents combine with each other
- The constituency structure of a given well-formed expression can be deduced from the tags

Auxiliary Task 1: CCG Supertagging

```
INPUT:    The   dog     bit       John
TARGET:  NP/N    N   (S\NP)/NP    NP
```

- Hypothesis: the structural information from the tagging task will help structural generalization
- CCG supertagging as an auxiliary task is helpful for tasks requiring sensitivity to linguistic structure, such as end-of-sentence detection
(Kim et al. 2019)

Auxiliary Task 2: Glossing (Word-by-word Translation)

- We saw in Part 1 that all models frequently produced single lexical retrieval errors

Single Lexical Retrieval Errors as Violations of Faithfulness Constraints

INPUT: A frog burned Sophia.

TARGET: $\texttt{frog}(x_1)\ AND\ \texttt{burn.agent}(x_2, x_1)\ AND\ \texttt{burn.theme}(x_2, \texttt{Sophia})$

ERROR: $\texttt{director}(x_1)\ AND\ \texttt{burn.agent}(x_2, x_1)\ AND\ \texttt{burn.theme}(x_2, \texttt{Sophia})$

- Outputs with single lexical retrieval errors violate two faithfulness constraints (in the terminology of Optimality Theory; Prince & Smolensky 1993/2002)
  - MAX ("No deletion"): Input `frog` should have a corresponding element in the output
  - DEP ("No insertion"): Output `director` should have a corresponding element in the input

- …most likely in favor of satisfying a conflicting constraint like "be probable!"
  - e.g., Subsequence $\texttt{director}(x_1)$ more probable than $\texttt{frog}(x_1)$ given the training data

Single Lexical Retrieval Errors as Violations of Faithfulness Constraints

INPUT: A frog burned Sophia.

TARGET: frog(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

ERROR: director(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

- The Optimality Theoretic constraints mentioned here serve as **descriptions of the output** rather than a statement about how the models operate

Auxiliary Task 2: Glossing (Word-by-word Translation)

- All models lacked bias for faithful outputs—outputs often contained single lexical retrieval errors that violate several faithfulness constraints
- Glossing task is a task that requires **maximally faithful input-output mappings**

```
INPUT:   The cat danced
TARGET:  The' cat' danced'
```

- **MAX (no deletion) and DEP (no insertion),** the constraints often violated by model outputs in Part 1, are fully satisfied
- Hypothesis: this task will help improve single lexical retrieval errors by promoting output faithfulness

Auxiliary Task 3: Word Prediction in Context ("Language Modeling")

Language Modeling

```
P(The cat sat on the mat) ?
~ P(mat|The cat sat on the) ?
```

Denoising variant of "word prediction in context"

```
INPUT:  The cat <x> on the <y>
TARGET: <x> sat <y> mat
```

- Elman (1991): Task of predicting the next word can help models discover linguistic regularities
- Has been shown to be almost universally helpful across various language tasks (identifying entailment, question-answering, discourse relation identification…)
- Has been shown to be effective for capturing complex syntactic phenomena like long distance dependencies (Gulordava et al. 2018, Goldberg 2019)
- Has been claimed to be effective for compositional generalization (Furrer et al. 2020, Tay et al. 2021)

Auxiliary Task 3: Word Prediction in Context (“Language Modeling”)

Language Modeling	`INPUT: The cat sat on the ___` `TARGET: mat`
Denoising variant of “word prediction in context”	`INPUT: The cat <x> on the <y>` `TARGET: <x> sat <y> mat`

- Main benefit might be an improvement in the models’ ability to detect substitutability of words or phrases
- No particular hypothesis on what kinds of errors this task will improve—mostly empirically motivated

Multiple Auxiliary Tasks: Glossing + CCG Supertagging

- Hypotheses on what glossing and CCG supertagging will be helpful for are complementary
    - Glossing: Unfaithful single lexical errors
    - CCG Supertagging: Structural generalization errors
- Maybe using both as auxiliary tasks would have a compound benefit

Comparison of Auxiliary Task Formats

CCG Supertagging

```
INPUT:    Kim    ate       cookies
TARGET:   NP   (S\NP)/NP    NP
```

Glossing

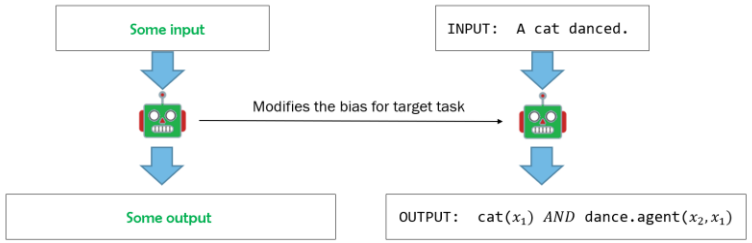
```
INPUT:    Kim    ate    cookies
TARGET:   Kim'   ate'   cookies'
```

Language Modeling

```
INPUT:    Kim     <x>     cookies
TARGET:   <x>     ate
```

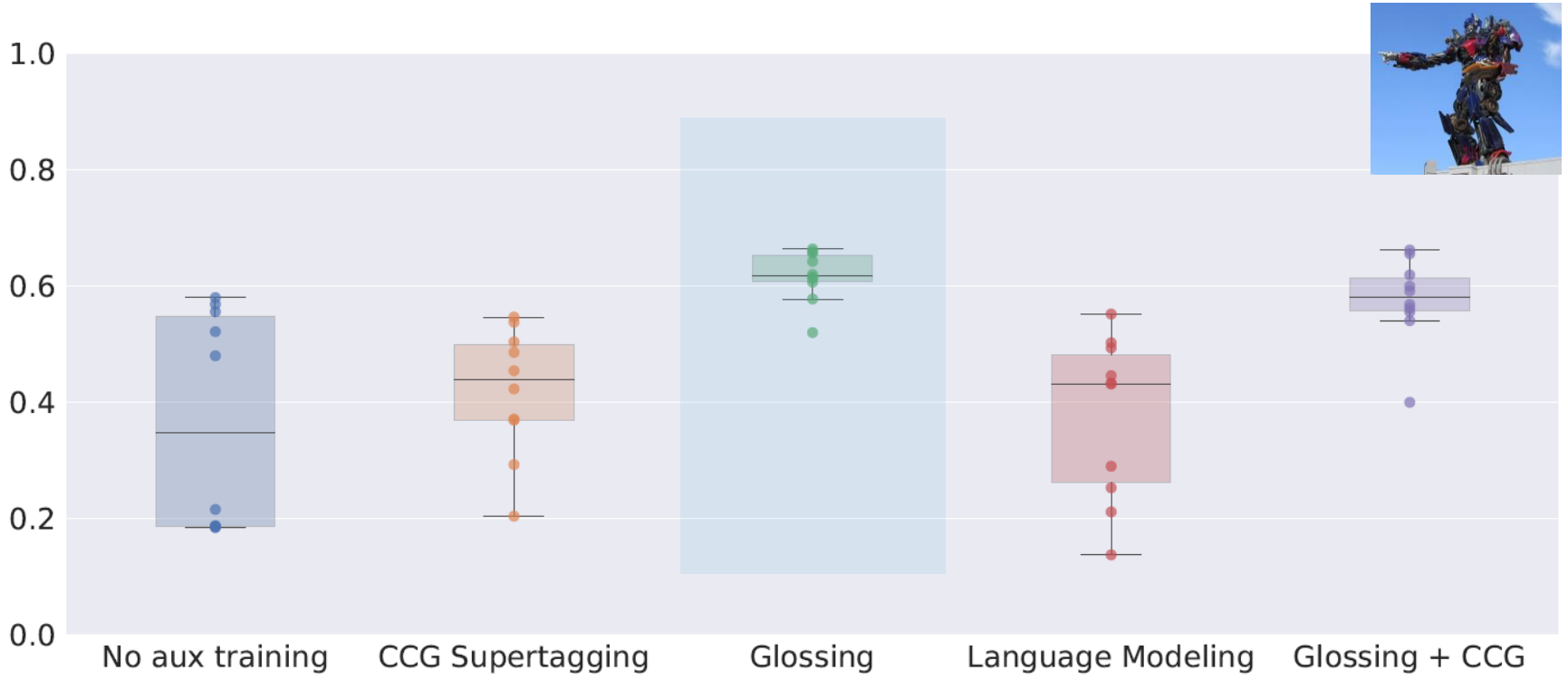
Used the same text data (WSJ corpus) to generate these datasets

Model & Training



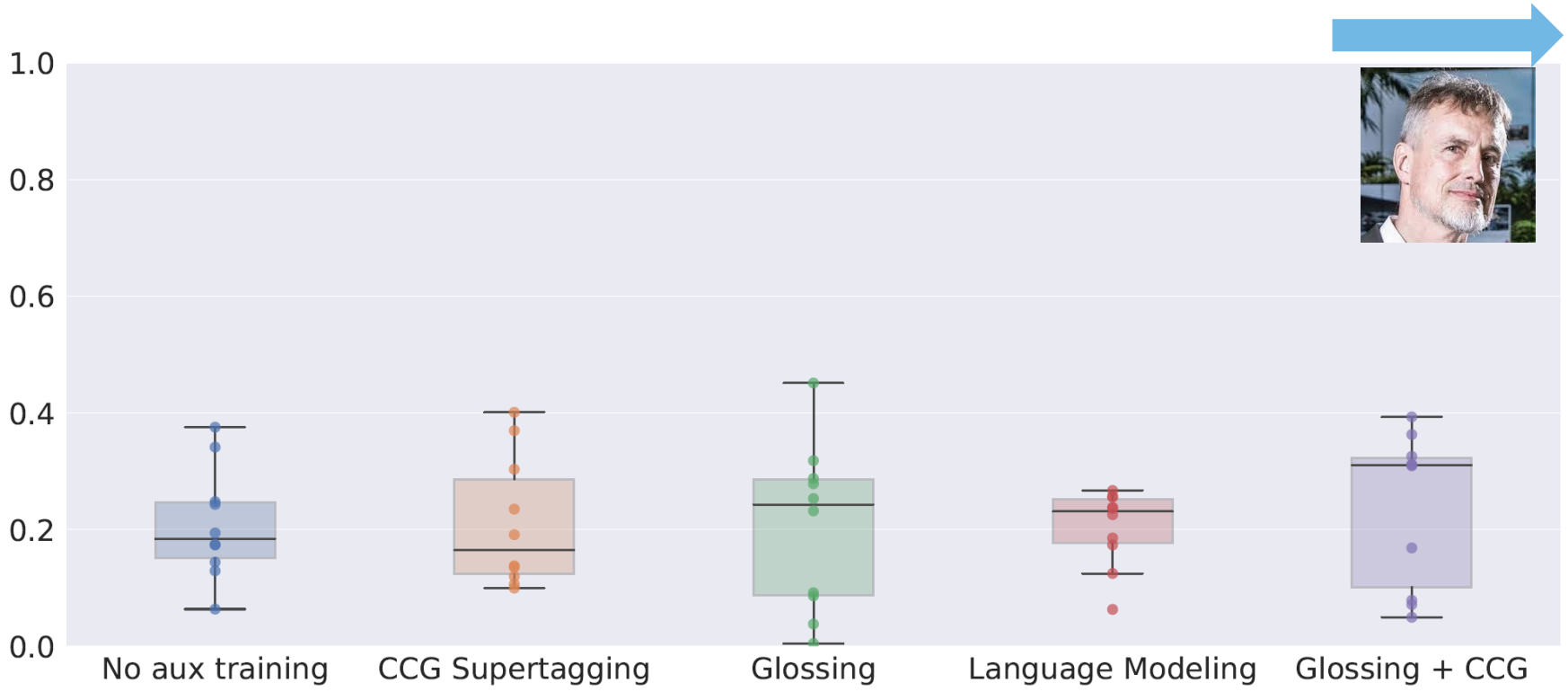
- Same set of models tested before (Transformer, LSTM, BiLSTM)
- Added an auxiliary task training step before training on the dataset for the compositional generalization test (probably not the most effective way!)
- Trained a new set of baseline models (i.e., models w/o auxiliary training) due to vocabulary size increase from introducing additional data
  - The only difference from previous models is the size of the vocabulary
- Trained each model 10 times with different random seeds

Results (Transformer)



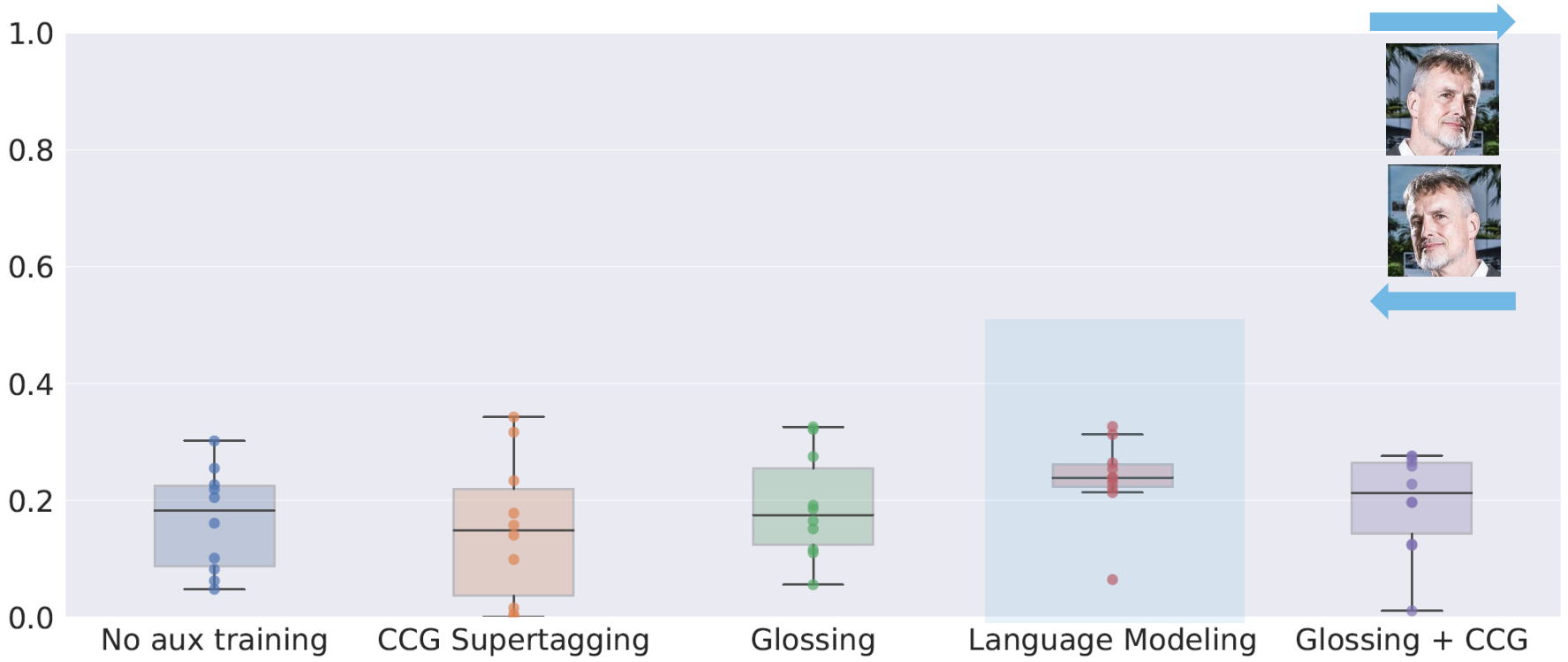
Substantial increase in generalization accuracy / reduction in variation across random restarts with glossing task as auxiliary objective!

Results (LSTM)



No noticeable improvements with any auxiliary task

Results (Bidirectional LSTM)



Minor benefit of denoising (“language modeling”)

Transformer Error Analysis: Were the Target Errors Fixed?

Hypothesis about the benefit of glossing task:

It will reduce single lexical errors by promoting faithfulness!

INPUT: A frog burned Sophia.

TARGET: frog(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

ERROR: director(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

Average single lexical error rate for Transformer models **without** glossing task: 25.1%

Average single lexical error rate for Transformer models **with** glossing task: 5.7%

Outputs ARE more faithful!

Structural Generalization: No Success!

- No combination of model and auxiliary task yielded substantial improvements in structural generalization
- The best performance (2.4% accuracy) was achieved by a bidirectional LSTM model trained on both Glossing and CCG supertagging tasks

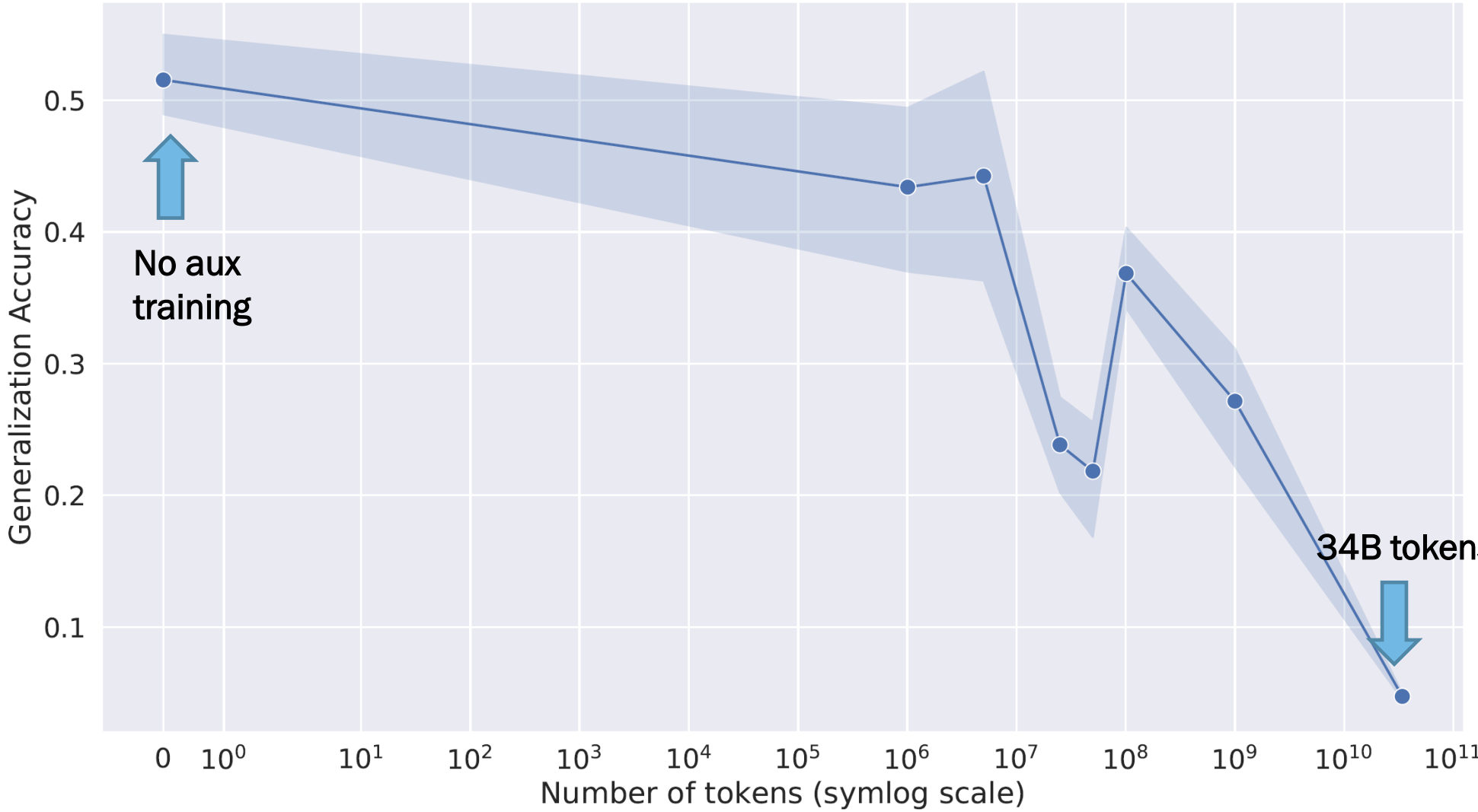
Experiment 2: Deep Dive into Predictive Pretraining as Auxiliary Task

- Scale is usually considered important for the effectiveness of predictive pretraining as an auxiliary task (Kaplan et al. 2020)
- The amount of data that we used in Experiment 1 is very small (1.2M tokens)
- Small amount even for human learners: a lower-end estimate is ~3M words/year (Hart and Risley 1995, re-cited from Linzen 2020)
- Investigating the effect of data size on compositional linguistic generalization is important

Setup

- Same sequential auxiliary task training setup with predictive pretraining (denoising)
- Different Transformer-based model (T5-small of Raffel et al. 2020) to allow investigation of much greater amount of training than our resources allow (34B tokens)
- Varied the amount of training (0, 1M, 5M, 25M, 50M, 100M, 1B, 34B tokens)
  - 34B-token model is a publicly available model

Results



More training data → lower generalization accuracy!

Error Analysis: Homogeneity of Errors in Models with More Training

- An error analysis revealed that the model with the largest amount of training data (34B tokens) had a very homogeneous pattern of error (i.e., single lexical retrieval errors)

INPUT: A frog burned Sophia.

TARGET: frog(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)
ERROR: director(x_1) AND burn.agent(x_2,x_1) AND burn.theme(x_2, Sophia)

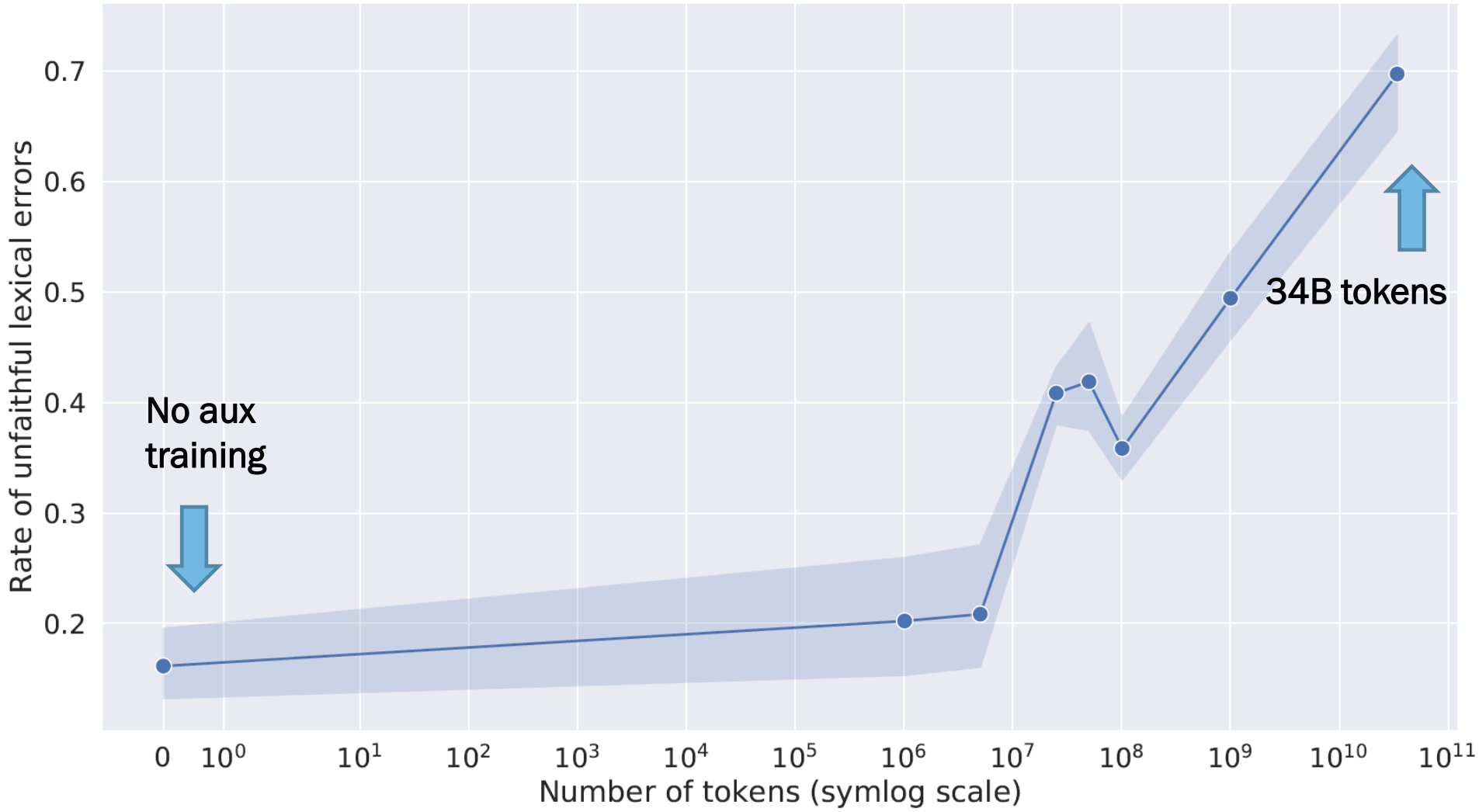
Error Analysis: Homogeneity of Errors in Models with More Training

- …whereas the model without auxiliary training showed relatively diverse patterns

INPUT: Lina drew Natalie.

TARGET: draw.agent(x_2, Lina) AND draw.theme(x_1, Natalie)
ERROR: draw.agent(x_2, Lina)

Q: Does the rate of single lexical retrieval errors scale with data size?



More training data → More unfaithful lexical errors

Again a case of Faithfulness/observed frequency competition? Predictive pretraining promotes the latter

Summary of Experiments

- Experiment 1: Comparing different auxiliary objectives
    - The glossing task helped add a faithfulness bias to the Transformer model, but not to others
    - CCG supertagging task did not help structural generalization
    - Denoising marginally benefited bidirectional LSTMs

- Experiment 2: Investigating the effect of training data size in denoising (predictive pretraining)
    - Surprisingly, more training data led to poorer compositional generalization
    - Increased rate of single lexical errors (unfaithful errors)

AGENDA

1. Background
2. A test for compositional linguistic generalization
3. Learning biases for compositional linguistic generalization
4. Conclusion & Future work

Takeaways

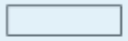
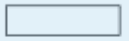
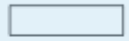
- Faithfulness is important for compositional linguistic generalization, and ANN learners tested often made generalizations that are unfaithful to the input (lack faithfulness bias)
    - Inductive bias for faithful generalizations could be injected through an auxiliary task that consists of maximally faithful input-output mappings
- Larger amounts of auxiliary training on denoising (predictive pretraining) led to **worse** compositional generalization
    - More pretraining data -> more unfaithful generalizations
- But almost complete failure, regardless of auxiliary training, on generalizing to novel structures (still somewhat true in follow-up studies: can talk more)

Concurrent findings

- Several papers have achieved improvements beyond what's discussed today…
  - Akyürek & Andreas (2021): **83%** acc. with LSTM + lexical translation + lexicon learning rule
  - Csordás, Irie & Schmidhuber (2021): **81%** acc. with modifications to the baseline setup
    - No early stopping, relative positional embeddings, disabling label smoothing
  - Bergen, O'Donnell & Bahdanau (2021): 87.4% with Edge Transformer
  - Tay et al. (2021): **77.5%** (pretrained Transformer) / **76.9%** (pretrained Conv seq2seq)
    ⚠ IMO confounded results—paper coming soon!
- Accuracy is more or less ~80%, which is about the ratio of lexical generalization cases in the dataset (~85%)

Concurrent findings

- Not all papers report breakdowns, but…

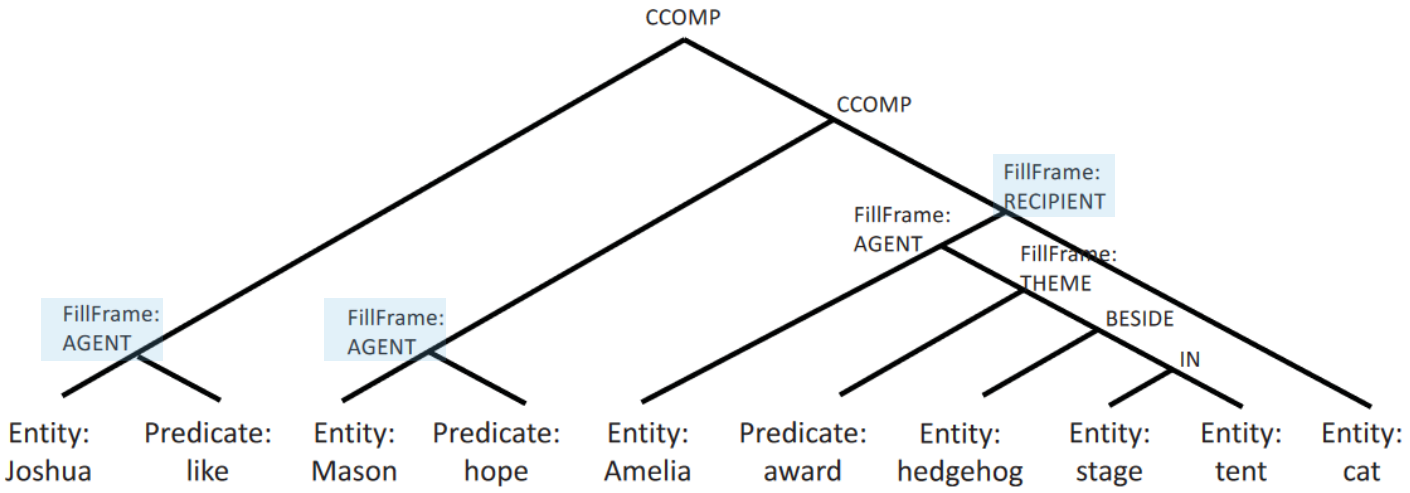
Categories	LSTM	+ copy	+ simple
primitive → {subj, obj, inf}			
active → passive			
obj PP → subj PP			
passive → active			
recursion			
unacc → transitive			
obj → subj proper			
subj → obj common			
PP dative ↔ obj dative			
all			

Akyürek & Andreas (2021)

Structural generalization is hard!

Solutions that work

- Liu et al. (2021): Tree-LSTM with latent grammar learning for both syntactic/semantic algebra with homomorphism assumption between the two
  - ~97% accuracy
  - Very strong priors *specific to the dataset*—permitted semantic operations are pre-defined



(b) An example of generated results in COGS benchmark with the input "Joshua liked that Mason hoped that Amelia awarded the hedgehog beside the stage in the tent to a cat".

Structural generalization is hard (without **very** strong priors)

Solutions that work

- Qiu et al. (2021): Induce a latent Quasi-Synchronous Context Free Grammar (QCFG), sample from the induced grammar and use the samples as additional training data
  - 98.9% accuracy!

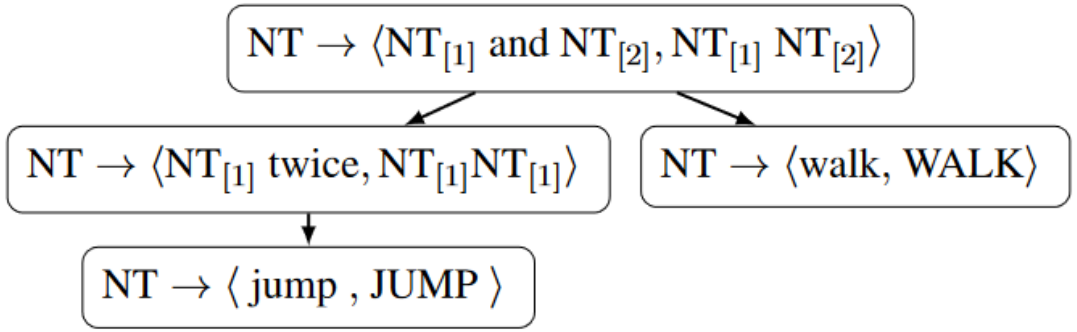


Figure 3: The example derivation of Figure 2 using QCFG notation.

Slightly more general approach (but arguably SCFG is still a quite specific prior)

Inspiration for a Follow-up Human Subject Study

- Modifier generalization to unseen grammatical positions:

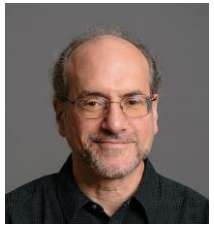
```
TRAINING:  {The cat saw the rat on the mat,
            The girl liked the cat on the table}

GENERALIZATION: The cat on the table saw the rat
```

- Predicted by context-free modification rules but not empirically attested

Thank you!!

Advisors & Collaborators & Funding



Paul Smolensky
(JHU/Microsoft)

Kyle Rawlins
(JHU)

Tal Linzen
(NYU/Google)

Dr. Cookie
(cat)

NSF DDRIG:
BCS-2041221

Case	Training	Generalization
S.4.3.1. Novel Combination of Familiar Primitives and Grammatical Roles		
Subject → Object (common noun)	A **hedgehog** ate the cake.	The baby liked the **hedgehog**.
Subject → Object (proper noun)	**Lina** gave the cake to Olivia.	A hero shortened **Lina**.
Object → Subject (common noun)	Henry liked a **cockroach**.	The **cockroach** ate the bat.
Object → Subject (proper noun)	The creature grew **Charlie**.	**Charlie** worshipped the cake.
Primitive noun → Subject (common noun)	**shark**	A **shark** examined the child.
Primitive noun → Subject (proper noun)	**Paula**	**Paula** sketched William.
Primitive noun → Object (common noun)	**shark**	A chief heard the **shark**.
Primitive noun → Object (proper noun)	**Paula**	The child helped **Paula**.
Primitive verb → Infinitival argument	**crawl**	A baby planned to **crawl**.
S.4.3.2. Novel Combination Modified Phrases and Grammatical Roles		
Object modification → Subject modification	Noah ate **the cake on the plate.**	**The cake on the table** burned.
S.4.3.3. Deeper Recursion		
Depth generalization: Sentential complements	Emma said **that** Noah knew **that** the cat danced.	Emma said **that** Noah knew **that** Lucas saw **that** the cat danced.
Depth generalization: PP modifiers	Ava saw the ball **in the bottle on the table**.	Ava saw the ball **in the bottle on the table on the floor**.
S.4.3.4. Verb Argument Structure Alternation		
Active → Passive	The crocodile **blessed** William.	A muffin **was blessed**.
Passive → Active	The book **was squeezed**.	The girl **squeezed** the strawberry.
Object-omitted transitive → Transitive	Emily **baked**.	The giraffe **baked a cake**.
Unaccusative → Transitive	The glass **shattered**.	Liam **shatterd** the jigsaw.
Double object dative → PP dative	The girl **teleported** Liam the cookie.	Benjamin **teleported** the cake **to** Isabella.
PP dative → Double Object Dative	Jane shipped the cake to John.	Jane shipped John the cake.
S.4.3.5. Verb Class		
Agent NP → Unaccusative subject	The **cobra** helped a dog.	The cobra **froze**.
Theme NP → Object-omitted transitive subject	The hippo **decomposed**.	The hippo **painted**.
Theme NP → Unergative subject	The hippo **decomposed**.	The hippo **giggled**.

REFERENCES

Akyürek, Ekin, and Jacob Andreas. "Lexicon Learning for Few Shot Sequence Modeling." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4934-4946. 2021.

Bangalore, Srinivas, and Aravind Joshi. "Supertagging: An approach to almost parsing." Computational linguistics 25, no. 2 (1999): 237-265.

Bergen, Leon, Timothy O'Donnell, and Dzmitry Bahdanau. "Systematic Generalization with Edge Transformers." Advances in Neural Information Processing Systems 34 (2021).

Caruana, Rich. "Multitask learning." Machine Learning 28, no. 1 (1997): 41-75.

Csordás, Róbert, Kazuki Irie, and Juergen Schmidhuber. "The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 619-634. 2021.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. 2019.

REFERENCES

Elman, Jeffrey L. "Distributed representations, simple recurrent networks, and grammatical structure." Machine Learning 7, no. 2 (1991): 195-225.

Furrer, Daniel, Marc van Zee, Nathan Scales, and Nathanael Schärli. "Compositional generalization in semantic parsing: Pre-training vs. specialized architectures." arXiv:2007.08970 (2020).

Goldberg, Yoav. "Assessing BERT's syntactic abilities." arXiv:1901.05287 (2019).

Gulordava, Kristina, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. "Colorless Green Recurrent Networks Dream Hierarchically." In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1195-1205. 2018.

Hart, Betty, and Todd R. Risley. Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing, 1995.

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural Computation 9, no. 8 (1997): 1735-1780.

REFERENCES

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models.“ arXiv:2001.08361 (2020).

Kim, Najoung, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney et al. "Probing What Different NLP Tasks Teach Machines about Function Word Comprehension." In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), pp. 235-249. 2019.

Lake, Brenden, and Marco Baroni. "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks." In International conference on machine learning, pp. 2873-2882. PMLR, 2018.

Linzen, Tal. "How Can We Accelerate Progress Towards Human-like Linguistic Generalization?." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5210-5217. 2020.

Liu, Chenyao, Shengnan An, Zeqi Lin, Qian Liu, Bei Chen, Jian-Guang Lou, Lijie Wen, Nanning Zheng, and Dongmei Zhang. "Learning Algebraic Recombination for Compositional Generalization." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1129-1144. 2021.

REFERENCES

Marvin, Rebecca, and Tal Linzen. "Targeted Syntactic Evaluation of Language Models." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1192-1202. 2018.

McCloskey, Michael. "Networks and theories: The place of connectionism in cognitive science." Psychological Science 2, no. 6 (1991): 387-395.

Meyerson, Elliot, and Risto Miikkulainen. "Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back." In International Conference on Machine Learning, pp. 3511-3520. PMLR, 2018.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations." In Proceedings of NAACL-HLT, pp. 2227-2237. 2018.

Prince, Alan, and Paul Smolensky. Optimality Theory: Constraint interaction in generative grammar. Optimality Theory in Phonology 3 (1993/2002), Wiley Online Library.

Qiu, Linlu, Peter Shaw, Panupong Pasupat, Paweł Krzysztof Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. "Improving Compositional Generalization with Latent Structure and Data Augmentation." arXiv preprint arXiv:2112.07610 (2021).

REFERENCES

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research 21 (2020): 1-67.

Tay, Yi, Mostafa Dehghani, Jai Gupta, Dara Bahri, Vamsi Aribandi, Zhen Qin, and Donald Metzler. "Are Pre-trained Convolutions Better than Pre-trained Transformers?." arXiv:2105.03322 (2021).

Tomasello, Michael, and Raquel Olguin. "Twenty-three-month-old children have a grammatical category of noun." Cognitive Development 8, no. 4 (1993): 451-464.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in Neural Information Processing Systems, pp. 5998-6008. 2017.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. "SuperGLUE: a stickier benchmark for general-purpose language understanding systems." In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 3266-3280. 2019.

REFERENCES

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. "Neural network acceptability judgments." Transactions of the Association for Computational Linguistics 7 (2019): 625-641.

Wilson, Colin. "Learning phonology with substantive bias: An experimental and computational study of velar palatalization." Cognitive Science 30, no. 5 (2006): 945-982

Zhang, Qilin, Gang Hua, Wei Liu, Zicheng Liu, and Zhengyou Zhang. "Can visual recognition benefit from auxiliary information in training?." In Asian Conference on Computer Vision, pp. 65-80. Springer, Cham, 2014.