Catch Compositional Generalization in Deep Learning: Model, Meaning and Data

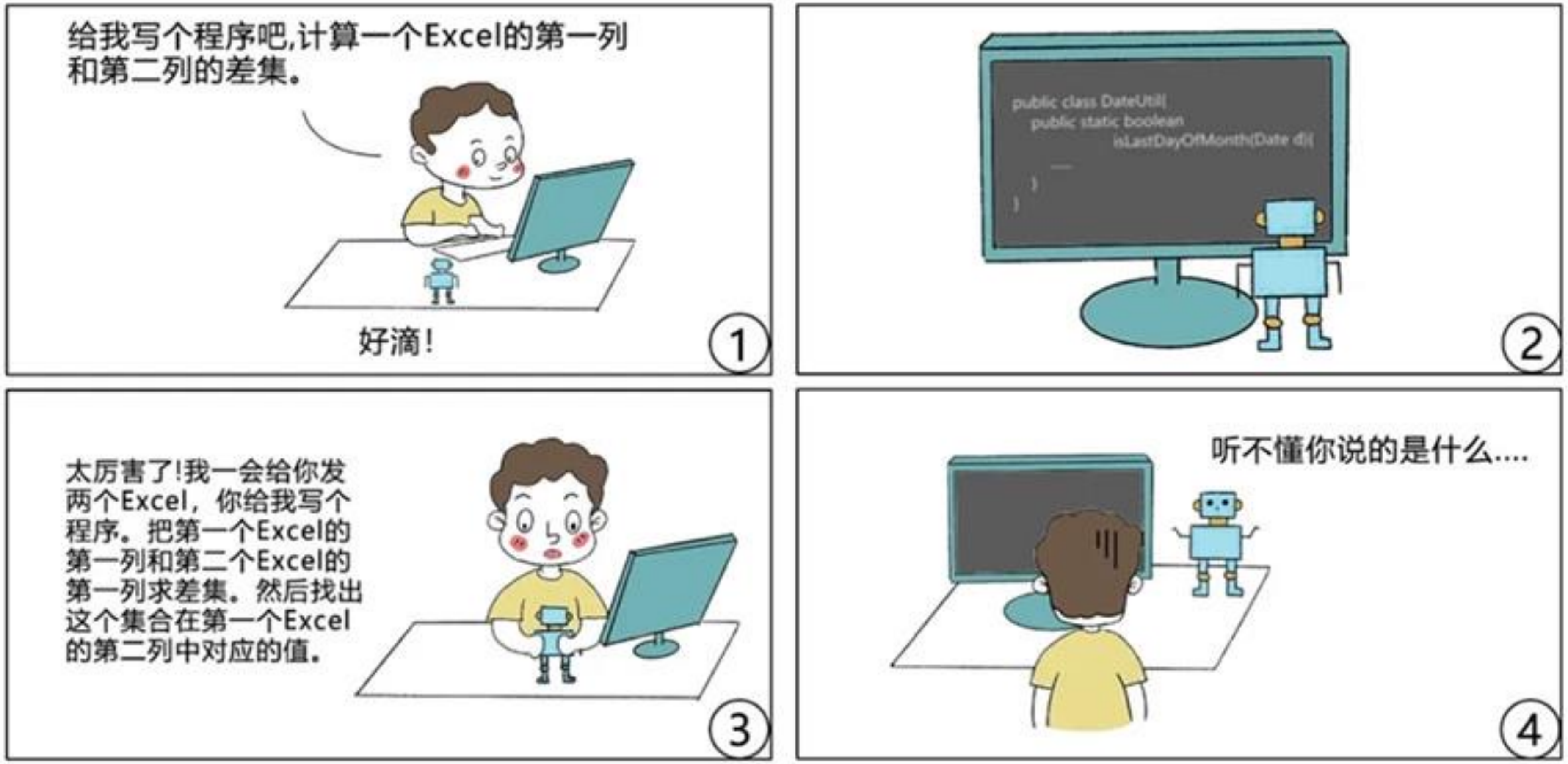
Qian Liu (qian.liu@buaa.edu.cn)

On Behalf of **MSRA DKI Team**

Alibaba Invited Talk @ 2021.09.16

Content

The current state of AI programmers



Compositional Generalization

- The compositionality of programs $\Longrightarrow$ huge search space of programs

- Compositional Generalization: human intelligence exhibits the algebraic capacity to dynamically recombine exis

Infinite use of finite means.

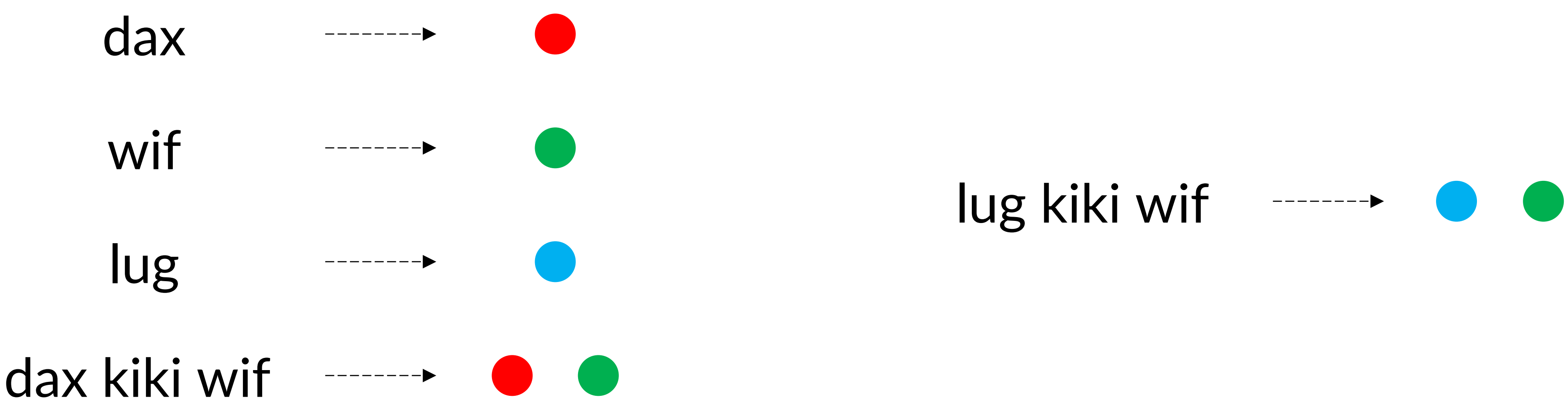
—— Noam Chomsky

Compositional Generalization in Cognition

Compositional generalization is an ability to recombine known parts to understand novel sentences which have never been encountered before.

Observed Examples

A Novel Example



Credit: Lake et al. 2019

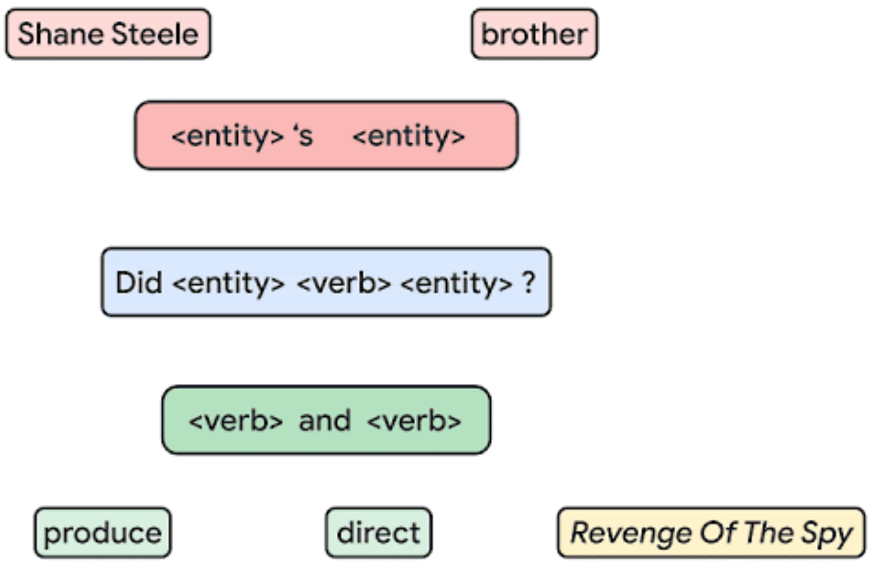
Compositional Generalization in NL2Code

The Simplified version of the CommAI Navigation (SCAN) is a **synthetic** benchmark (Lake & Baroni. 2018) with navigation commands and action sequences.

	自然语言		导航动作序列
训练集	run twice	$\Longrightarrow$	RUN RUN
	jump and walk	$\Longrightarrow$	JUMP WALK
测试集	run and jump twice	$\Longrightarrow$	RUN JUMP JUMP

Compositional Generalization in NL2Code

CFQ (Compositional Freebase Questions) is a **realistic** benchmark (Keysers et al. 2020) that comprehensively measure compositional generalization on KBQA.



Credit: Keysers et al. 2020

Which Swedish founder of [M0] produced [M2] ?

```
SELECT DISTINCT ?x0 WHERE {
    ?x0 ns:film.producer.film|ns:film.production_company.films [M2] .
    ?x0 ns:organization.organization_founder.organizations_founded [M0] .
    ?x0 ns:people.person.nationality ns:m.0d0vqn
}
```

Measuring Compositional Generalization

The SCAN benchmark is split in **handcraft ways** to form the challenges:

	Add jump	Around Right	Length Generalization
Train	jump walk twice walk around left	turn around left turn opposite right walk around left	look around left look around left twice look around left twice after look
Test	jump around left	turn around right	look around left twice after look around left

No complex command of jump in training

"around right" is held out from the training set

Train: length of the action sequence is shorter than 24 actions; Test: all action sequences longer than or equal to 24 actions.

Measuring Compositional Generalization

The CFQ benchmark is split based on **automatic algorithms** which highlight properties that intuitively correlate with compositional structure:

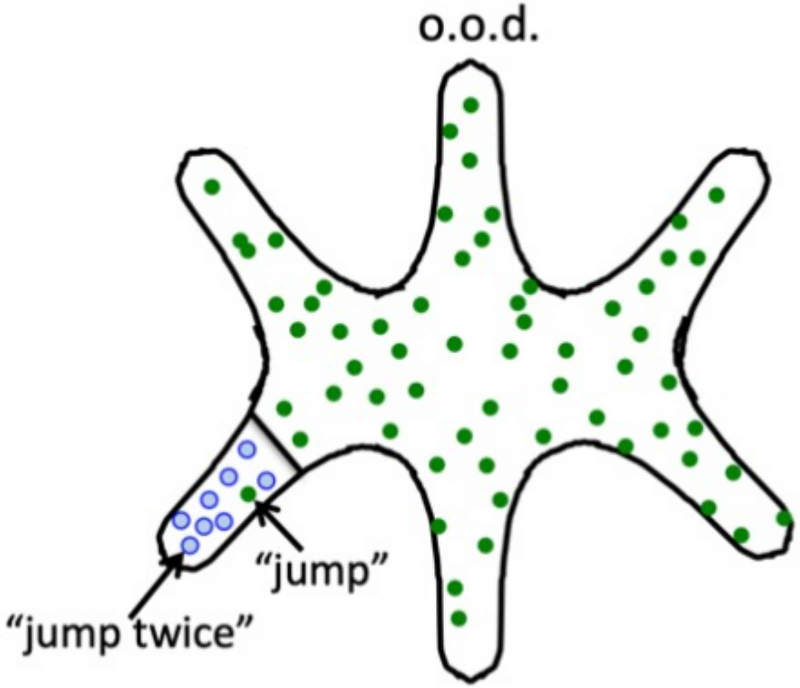
(1) Similar atom distribution: All test atoms occur in train, and Distribution of atoms is similar between train and test.

(2) Different compound distribution: Distribution of compounds is different between train and test.

Train	**Test**
Who **directed** Inception?	Who **produced** Inception?
Did Greta Gerwig **produce** Goldfinger?	Did Greta Gerwig **direct** Goldfinger?

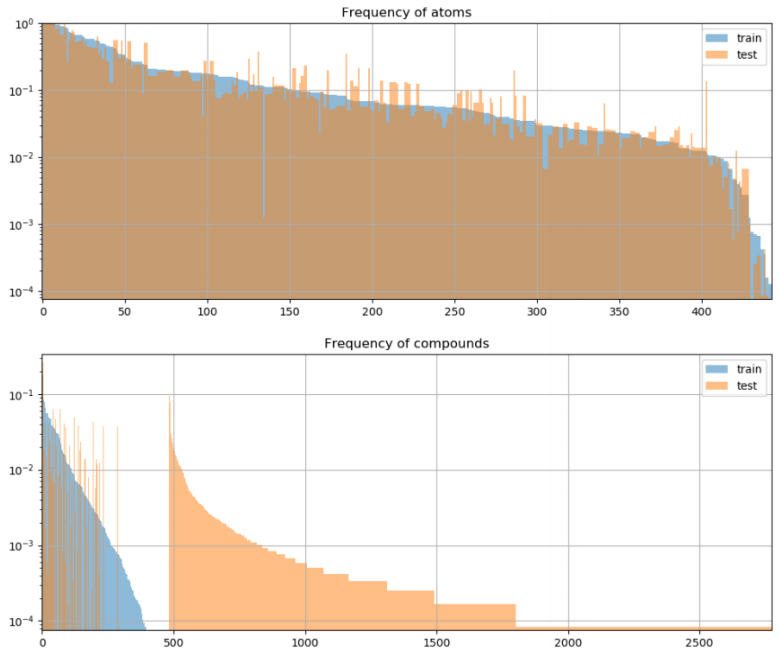
Measuring Compositional Generalization

The SCAN split distribution



Credit: Russin et al. 2020

The CFQ split distribution



Credit: Keysers et al. 2020

A Promising Direction

- Datasets
  - ✓ **SCAN** (Lake & Baroni, ICML'18)
  - ✓ **CFQ** (Keysers et al, ICLR'20)
  - ✓ **COGS** (Kim & Linzen, EMNLP'20)
  - ✓ **Grounded SCAN** (Ruis et al, NeurIPS'20)

- Methods
  - ✓ CGPS (Li et al, EMNLP'19)
  - ✓ Meta Seq2Seq (Brenden M. Lake, NeurIPS'19)
  - ✓ Permutation Equivariant Seq2Seq (Gordon et al, ICLR'20)
  - ✓ GECA (Jacob Andreas, ACL'20)

  …

Far From Compositional Generalization

No model can successfully solve compositional challenges on SCAN!

Model	Add Jump	Around Right	Length
Seq2Seq	1.2	2.5	13.8
CNN	69.2	56.7	0.0
Syntactic Attention (Russin et al. 2019)	91.0	28.9	15.2
CGPS (Li et al. 2019)	98.8	83.2	20.3
GECA (Jacob Andreas. 2020)	86.0	82.0	-
Meta Seq2Seq (Brenden M. Lake. 2019)	99.9	99.9	16.6
Equivariant Seq2Seq (Gordon et al. 2020)	99.1	92.0	15.9

*green models trained w/o extra resources

*blue models trained with extra resource

Opportunities: from the perspective of ML

- **Model:** Cooperative Modules

  Compositional Generalization by Learning Analytical Expressions [NeurIPS'20]

- **Meaning:** Semantic Structure in Code

  Hierarchical Poset Decoding for Compositional Generalization in Language [NeurIPS'20]

- **Data:** Potential of Monolingual Data

  Revisiting Iterative Back-Translation from the Perspective of Compositional Generalization [AAAI'20]

Content

1 What is Compositional Generalization

2 Model: Cooperative Modules

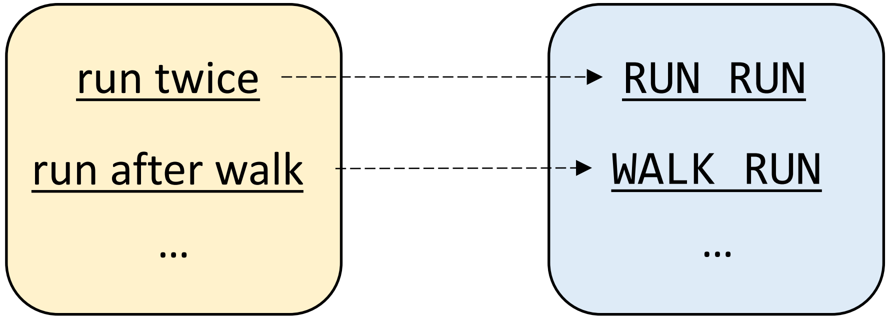
3 Meaning: Semantic Structure in Code

4 Data: Potential of Monolingual Data

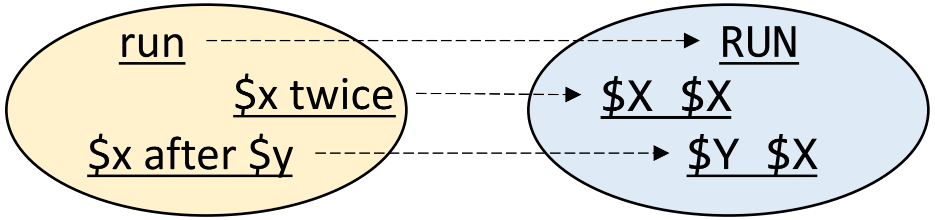
Model on Compositionality

The compositionality of language constitutes an algebraic system, of the sort that can be captured by symbolic functions with variable slots (M. Baroni, 2019).

Current Neural based Models



Our Model



Learn Analytical Expressions

The understanding of “run opposite left after walk twice” can be regarded as a hierarchical application of symbolic functions.

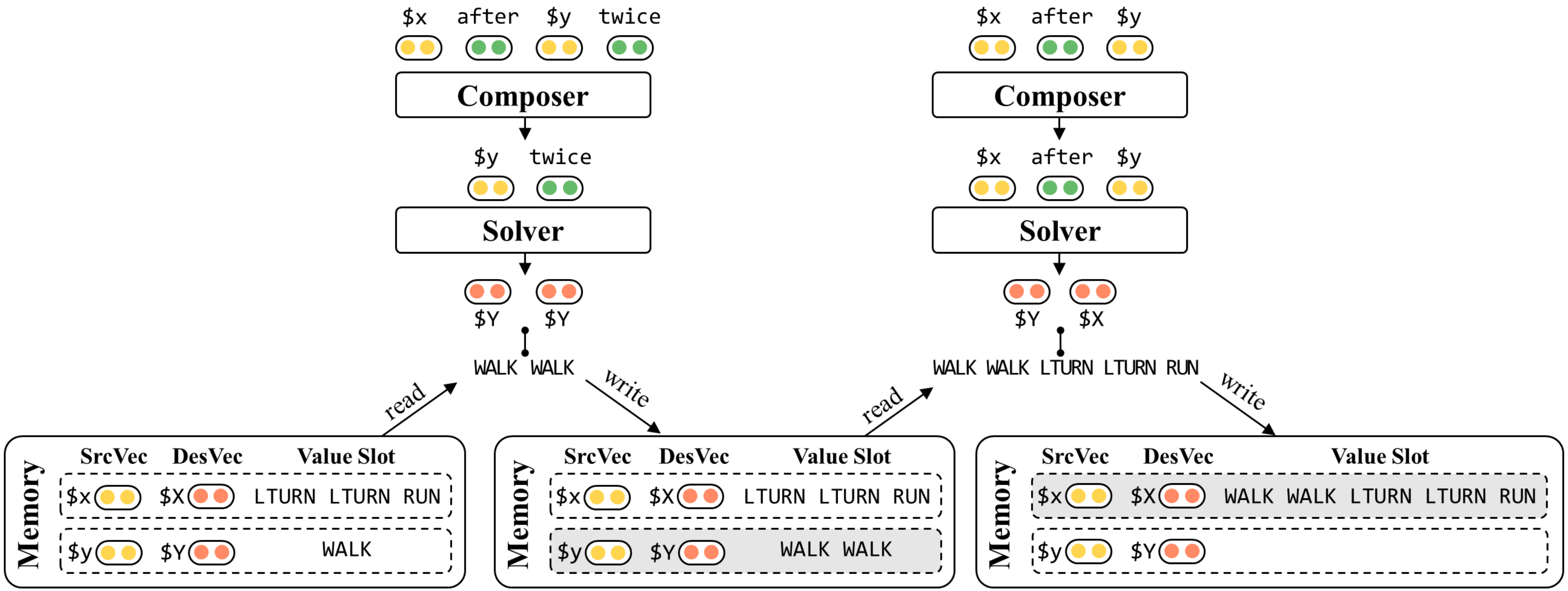
6	$x after $y
5	$x after $y twice
4	$x after walk twice
3	$x opposite $yafter walk twice
2	$x opposite left after walk twice
1	run opposite left after walk twice

Symbolic Function

$x after $y	⟶	$Y $X
$y twice	⟶	$Y $Y
walk	⟶	WALK
$x opposite $y	⟶	$Y $Y $X
left	⟶	LTURN
run	⟶	RUN

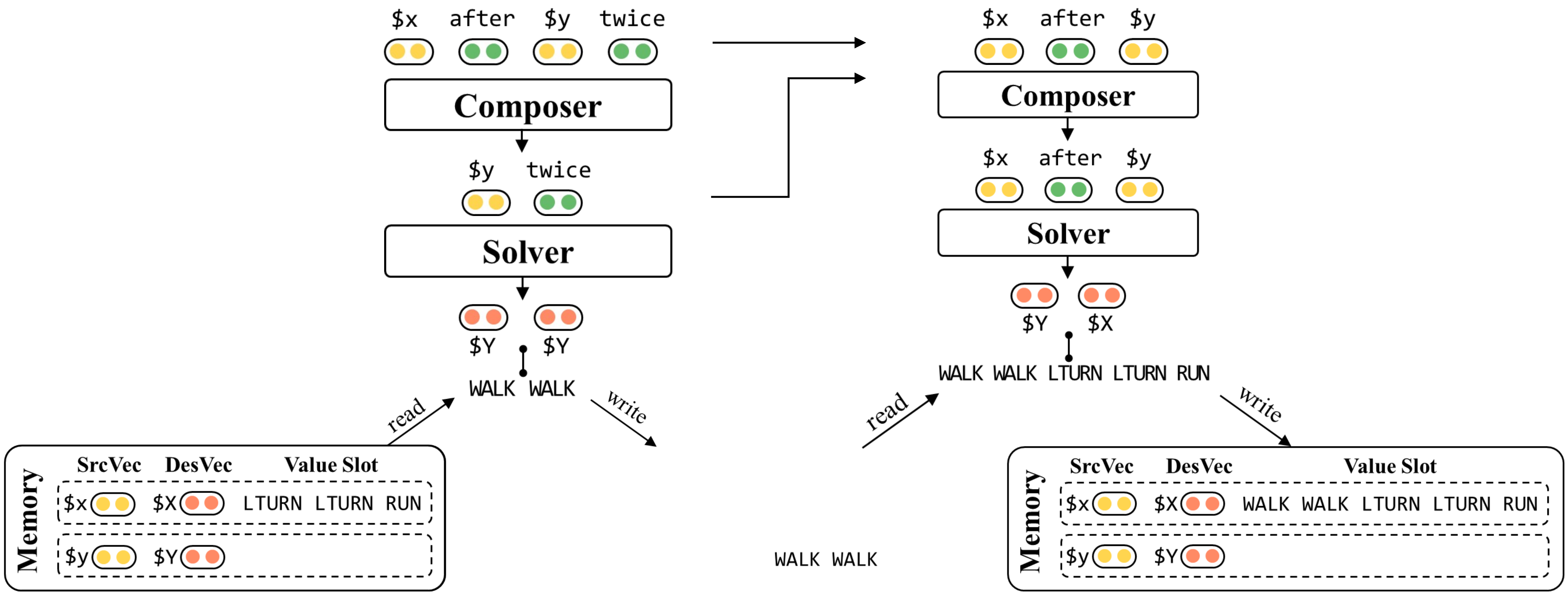
LANE: Memory-Augmented Model

We propose a memory-augmented neural model to achieve compositional generalization by automatically learning the above analytical expressions.



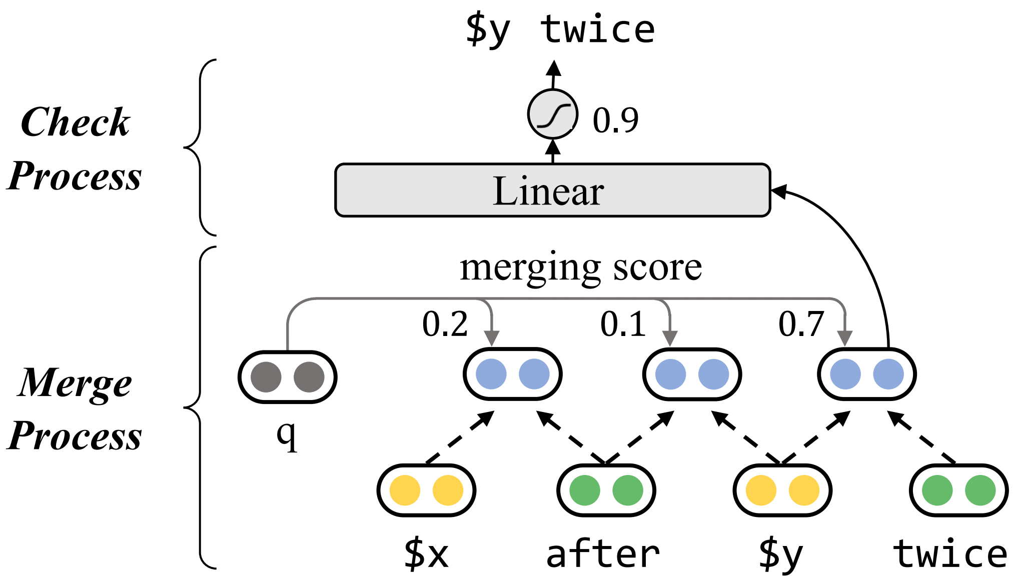
LANE: Memory-Augmented Model

Our model understands via interaction between Composer, Solver and Memory.



Composer: Find Expressions by Merging

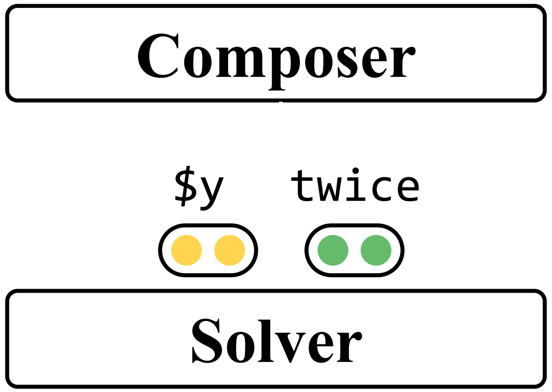
Composer gradually merges elements of the input until a recognizable source analytical expression appears, just as building a binary tree from bottom to top



The Training is Challenging!

Challenges

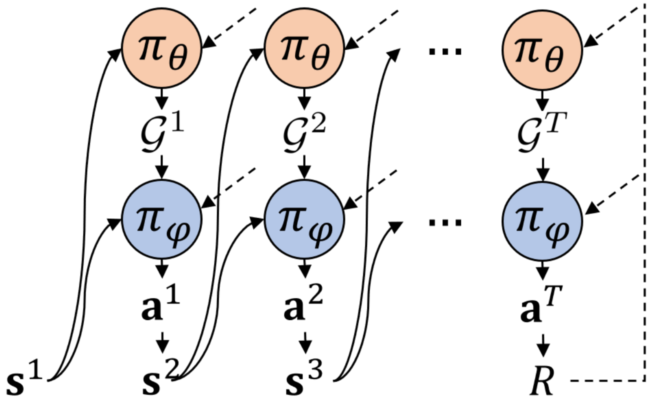
(i) Discrete Action, Non-differentiable.



(ii) Sparse Reward, Hard to Train.

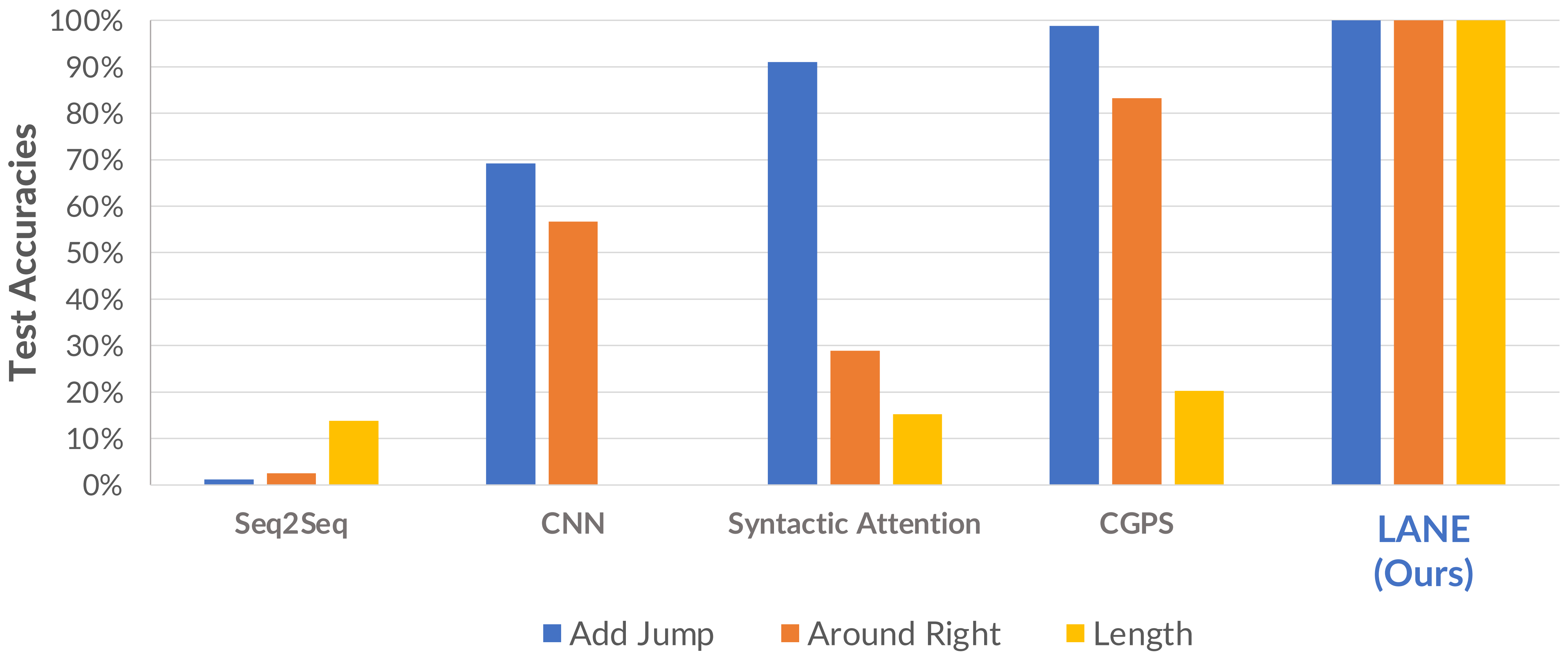
Solutions

(i) Hierarchical Reinforcement Learning.



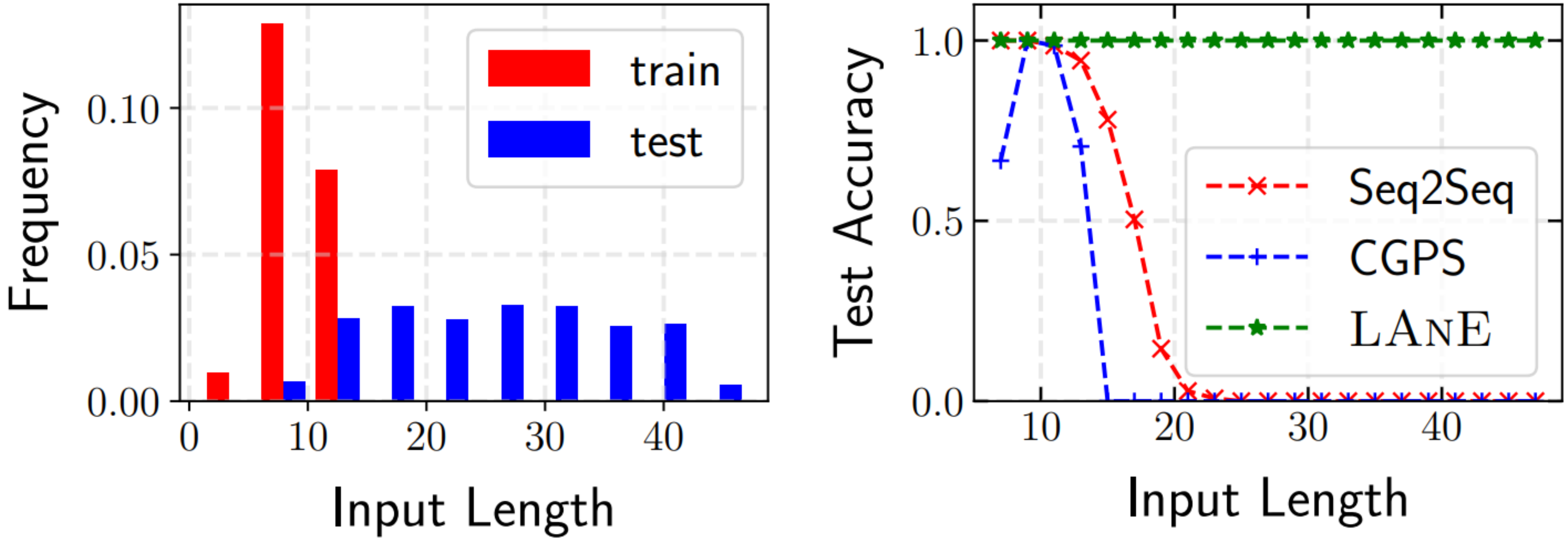
(ii) Curriculum Learning.

Evaluate on SCAN



Evaluate on Longer Inputs

Languages license a theoretically infinite set of sentences due to compositionality, and our model maintains a perfect trend as the input length increases.



Extend to More Realistic Scenarios

SCAN aims to raise the attention on compositional generalization, which simplifies the compositional generalization issue under real scenarios.

CFQ (NL-to-SPARQL)	X Did a male film director edit and direct Star Wars? Y SELECT count (*) WHERE { ?x0 ns:film.director.film m_06mmr . ?x0 ns:film.editor.film m_06mmr . ?x0 ns:people.person.gender m_05zppz }
COGS (NL-to-Logic)	X Charlotte was given the cake on a table. Y cake(x_4) ; give.recipient (x_2, Charlotte) AND give.theme(x_2,x_4) AND cake.nmod.on(x_4, x_7) AND table(x_7)
GEOQuery (NL-to-SQL)	X What state has the largest area? Y SELECT state.name FROM state WHERE state.area = (SELECT MAX(state.area) FROM state)

Part I. Model

Compositional Generalization by Learning Analytical Expressions [NeurIPS'20]

- The key for compositionality is to regard language as an algebraic system, which be captured by analytical expressions.
- Learning analytical expressions can be modeled as the joint optimization of three cooperative modules.
- Latent discrete actions between modules can be tackled by the combination of hierarchical reinforcement learning and curriculum learning.

Content

1. What is Compositional Generalization
2. Model: Cooperative Modules
3. Meaning: Semantic Structure in Code
4. Data: Potential of Monolingual Data

Partial Permutation Invariance

- Semantics is usually invariant to permute some components in code.
- There are many equivalent meaning representations, but current deep learning decoders just select one certain order as the target for optimization.

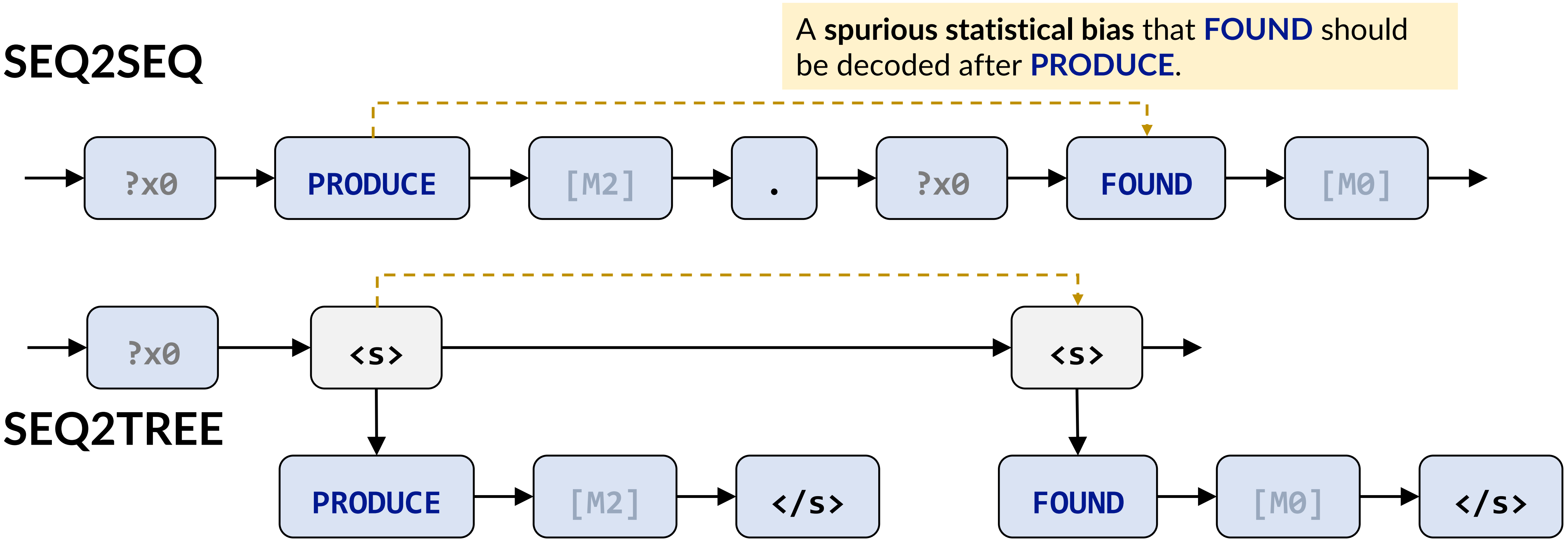
```
SELECT DISTINCT ?x0 WHERE {
          ?x0 PRODUCE [M2] .
          ?x0 FOUND [M0] .
          ?x0 NATIONALITY ns:m.0d0vqn
}
```

```
SELECT DISTINCT ?x0 WHERE {
          ?x0 FOUND [M0] .
          ?x0 PRODUCE [M2] .
          ?x0 NATIONALITY ns:m.0d0vqn
}
```

```
SELECT DISTINCT ?x0 WHERE {
          ?x0 NATIONALITY ns:m.0d0vqn
          ?x0 PRODUCE [M2] .
          ?x0 FOUND [M0] .
}
```

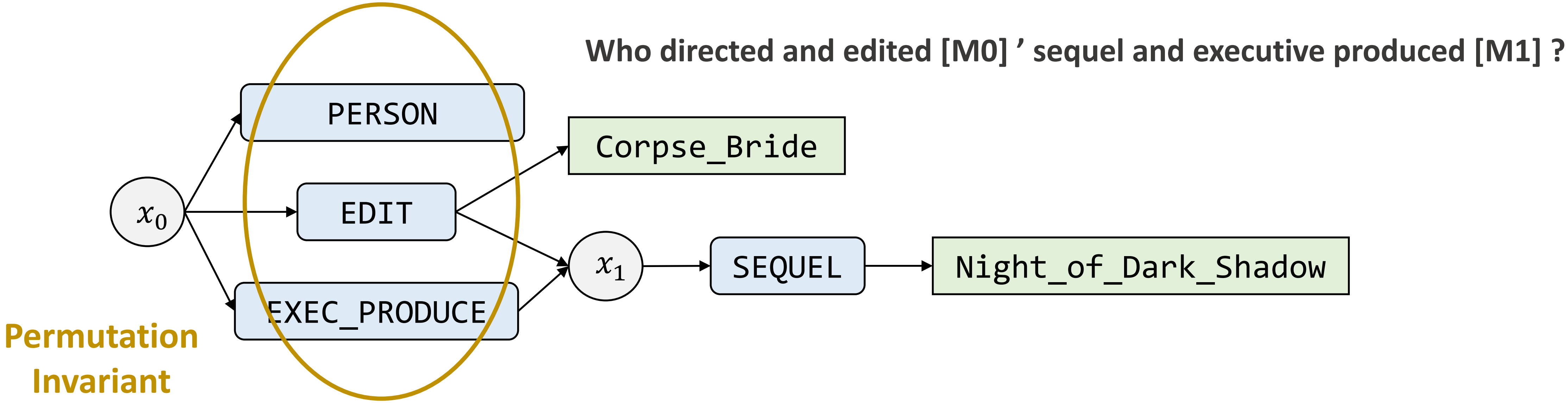
Decoding Order

Imposing additional ordering constraints increases learning complexity, thus limiting compositional generalization.

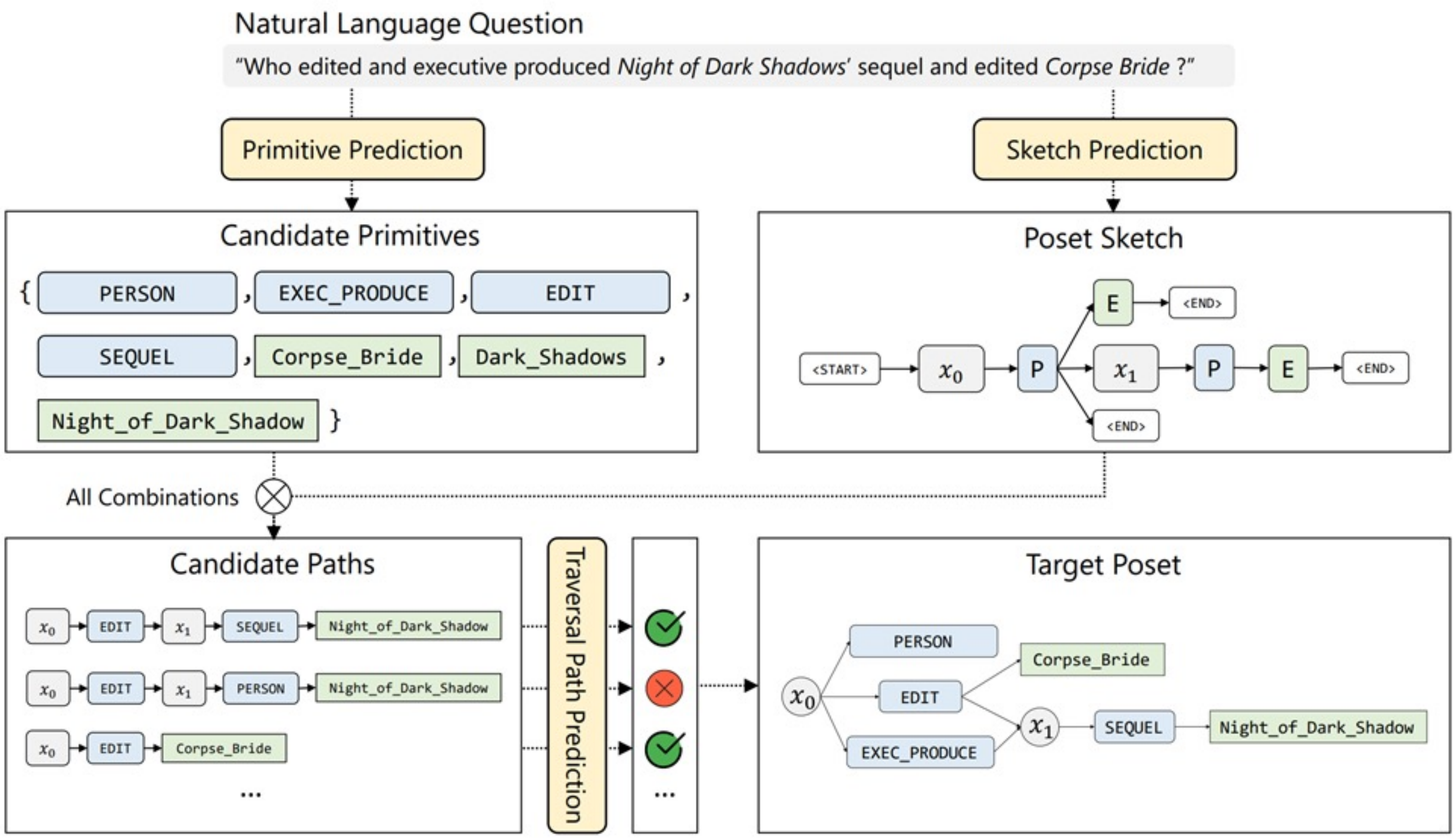


Semantic Meaning as Poset

- Poset (Partially Ordered SET)
  - Poset is a set with a partial order relation (reflexive, antisymmetric, and transitive).
  - Every poset can take the form of a DAG (Directed Acyclic Graph).
- Decode a poset, rather than a sequence/tree.



Hierarchical Poset Decoding



Evaluate on CFQ

Models	MCD1	MCD2	MCD3
LSTM+Attention (Keysers et al., 2020)	28.9%	5.0%	10.8%
Transformer (Keysers et al., 2020)	34.9%	8.2%	10.6%
Universal Transformer (Keysers et al., 2020)	37.4%	8.1%	11.3%
LSTM+Attention (with simplified SPARQL expression)	42.2%	14.5%	21.5%
Transformer (with simplified SPARQL expression)	53.0%	19.5%	21.6%
Seq2Tree (Dong and Lapata, 2016)	24.3%	4.1%	6.3%
CGPS (Li et al., 2019)	4.81%	1.04%	1.82%
Hierarchical Poset Decoding	**79.6%**	**59.6%**	**67.8%**
with Seq2Seq-based sketch prediction	74.3%	45.7%	50.2%
with Seq2Tree-based sketch prediction	75.7%	40.9%	51.1%
w/o Hierarchical Mechanism	21.3%	6.4%	10.1%

Part II. Meaning

Hierarchical Poset Decoding for Compositional Generalization in Language [NeurIPS'20]

- Poset structure in semantics is a key factor for compositional generalization in language.
- Hierarchical Poset Decoding on the formal language can significantly enhance the compositional generalization (CFQ 18.9→69.0).

Content

Monolingual Data on Compositionality

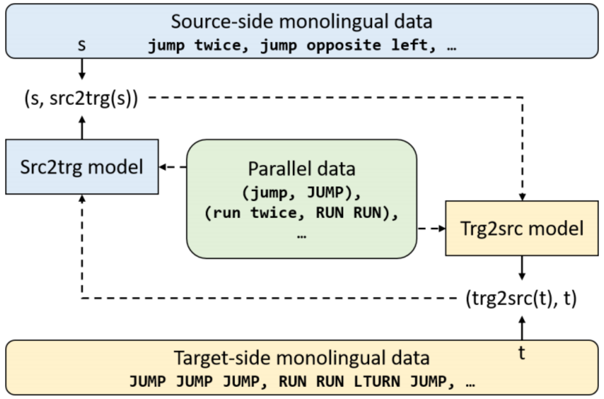
- NL-Code parallel data are limited and expensive.
- NL/Code monolingual data are cheap and abundant.

Unlabeled documents

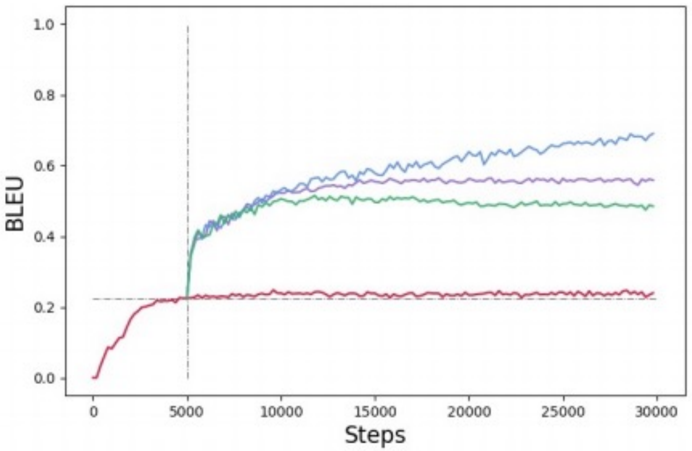
Unlabeled programs

Revisiting Iterative Back-Translation

- Dual structure for exploiting large-scale monolingual data (natural language utterances & programs)



Workflow of iterative back-translation



Self-cleaning of pseudo-parallel data

Models	MCD1	MCD2	MCD3
LSTM+Attn	28.9 ± 1.8	5.0 ± 0.8	10.8 ± 0.6
Transformer	34.9 ± 1.1	8.2 ± 0.3	10.6 ± 1.1
Uni-Transformer	37.4 ± 2.2	8.1 ± 1.6	11.3 ± 0.3
CGPS	13.2 ± 3.9	1.6 ± 0.8	6.6 ± 0.6
T5-11B	61.4 ± 4.8	30.1 ± 2.2	31.2 ± 5.7
GRU+Attn (Ours)	32.6 ± 0.22	6.0 ± 0.25	9.5 ± 0.25
+mono30	**64.8 ± 4.4**	**57.8 ± 4.9**	**64.6 ± 4.9**
+mono100	83.2 ± 3.1	71.5 ± 6.9	81.3 ± 1.6
+transductive	88.4 ± 0.7	81.6 ± 6.5	88.2 ± 2.2

Results on CFQ: good compositional generalization

Reference

[1]. Lake et al. Human few-shot learning of compositional instructions. In CogSci 2019.

[2]. Lake & Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In ICML 2018.

[3]. Keysers et al. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. In ICLR 2020.

[4]. Russin et al. Compositional generalization by factorizing alignment and translation. In EMNLP 2020 Student Workshop.

[5]. Kim & Linzen COGS: A compositional generalization challenge based on semantic interpretation. In EMNLP 2020.

[6]. Ruis et al. A Benchmark for Systematic Generalization in Grounded Language Understanding. In NeurIPS 2020.

[7]. Li et al. Compositional Generalization for Primitive Substitutions. In EMNLP 2019.

[8]. Brenden M. Lake. Compositional generalization through meta sequence-to-sequence learning. In NeurIPS 2019.

[9]. Gordon et al. Permutation Equivariant Models for Compositional Generalization in Language. In ICLR 2020.

[10]. Jacob Andreas. Good-Enough Compositional Data Augmentation. In ACL 2020.

[11]. M. Baroni. Linguistic generalization and compositionality in modern artificial neural networks. In Phil. Trans. R. Soc. B. 2019.

[12]. Dong and Lapata. Language to Logical Form with Neural Attention. In ACL 2016.

Thanks & QA