# **SUBS: Subtree Substitution** for **Compositional** Semantic Parsing

**Jingfeng Yang**, Le Zhang, Diyi Yang

# Compositional Genarlization in Semantic Parsing

**Training Example 1:**
*Natural:* What is the largest city in the smallest state in the USA ?
*Formal (FunQL):* answer ( largest ( city ( loc_2 ( smallest ( state ( loc_2 ( countryid ( usa ) ) ) ) ) ) ) )

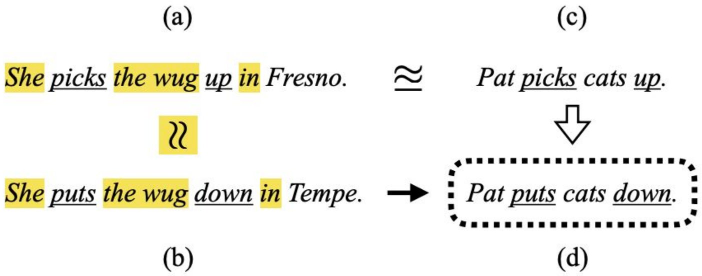
**Training Example 2:**
*Natural:* What is the population of the largest state ?
*Formal (FunQL):* answer ( population_1 ( largest ( state ( all ) ) ) )

**Test Example:**
*Natural:* What is the population of the largest city in the smallest state in the USA ?
*Formal (FunQL):* answer ( population_1 ( largest ( city ( loc_2 ( smallest ( state ( loc_2 ( countryid ( usa ) ) ) ) ) ) ) ) )

# Prior Work for Compositional Semantic Parsing

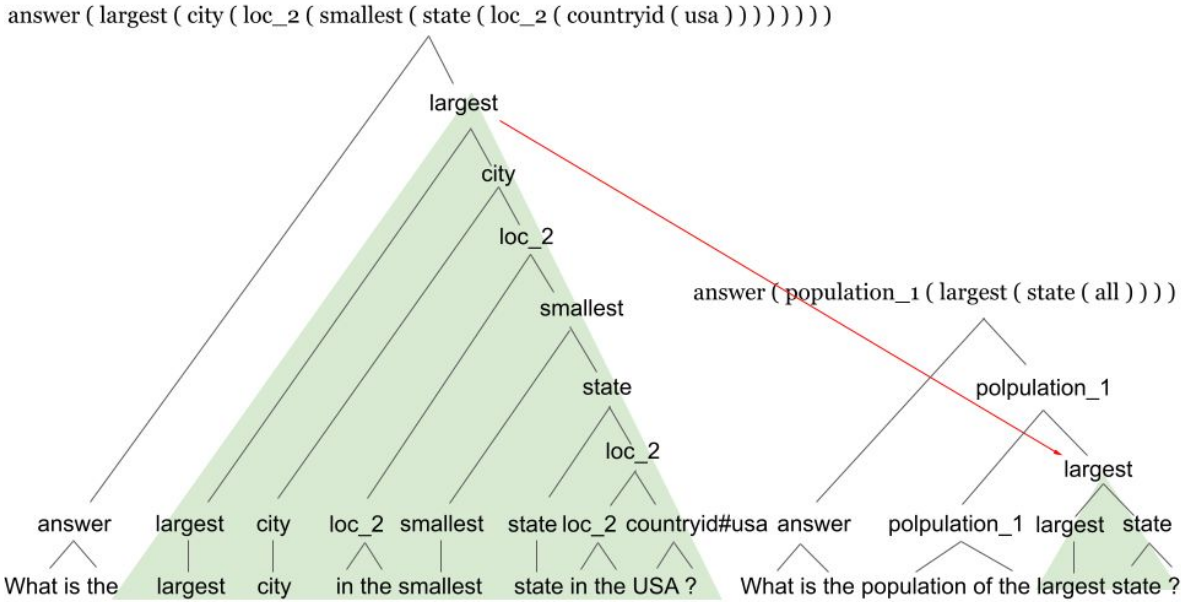
- Model Biases: Span-based Semantic Parsing (Herzig et al., 2021), Neural-Symbolic Stack Machines (Chen et al., 2020 ), Neural Module Networks (Gupta et al., 2019) etc.

- Data Augmentation and then Seq2seq Model:
  - Synchronous Context-Free Grammar (SCFG) (Jia et al., 2016).
  - Good-Enough Compositional Data Augmentation (GECA) (Andreas et al., 2019):


(a)
(c)
She picks the wug up in Fresno.
≈
Pat picks cats up.
≈
She puts the wug down in Tempe.
Pat puts cats down.
(b)
(d)

Limitations of prior Data Augmentation: identify only simple replaceable spans!

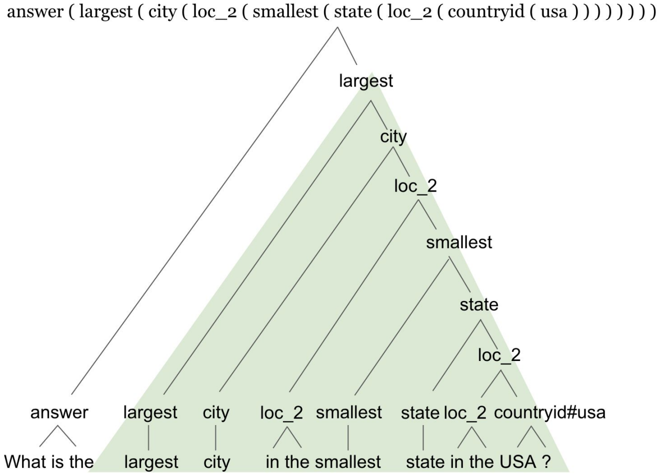
# Subtree Substitution (SUBS) Data Augmentation



**Subtree Substitution Result:**
What is the population of the largest city in the smallest state in the USA ?
answer ( population_1 ( largest ( city ( loc_2 ( smallest ( state ( loc_2 ( countryid ( usa ) ) ) ) ) ) ) ) )

# Dataset and Tree Source

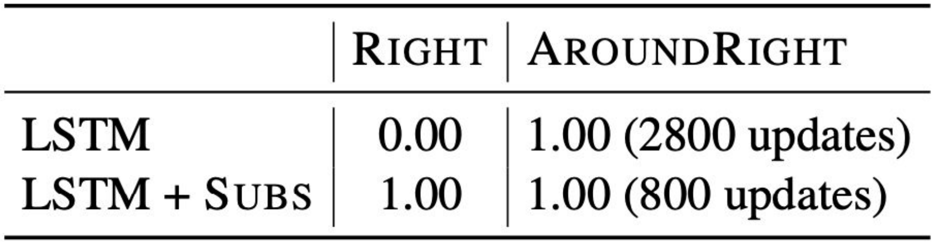
- Dataset:
  - SCAN (“turn around left” -> “LTURN LTURN LTURN LTURN”)
  - GeoQuery
- Span Tree:



  - Induced by Span-based Semantic Parsing (Herzig et al., 2021)
  - Semi-automatically annotated gold trees which requires only manually designed domain-specific lexicon and rules (Herzig et al., 2021).

# Results - SCAN

Better performance and faster convergence on the diagnostic dataset.

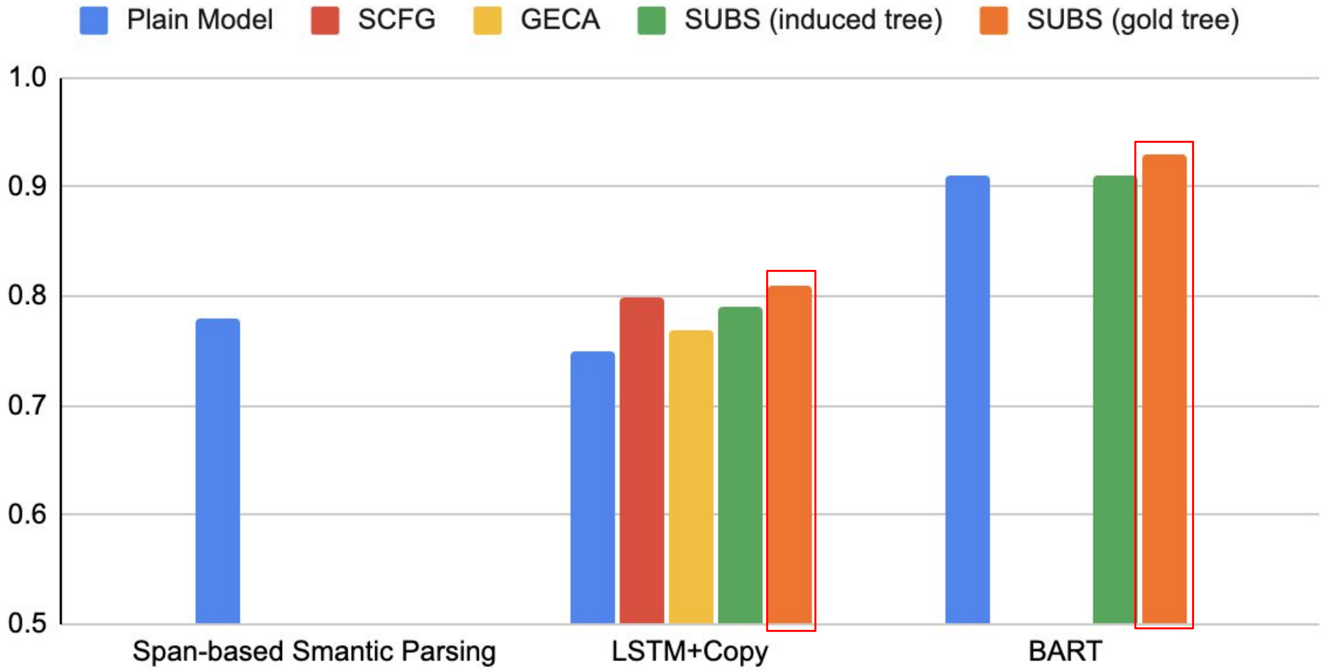


| | Right | AroundRight |
|---|---|---|
| LSTM | 0.00 | 1.00 (2800 updates) |
| LSTM + Subs | 1.00 | 1.00 (800 updates) |

Table 1: Accuracy of diagnostic experiments on Scan.

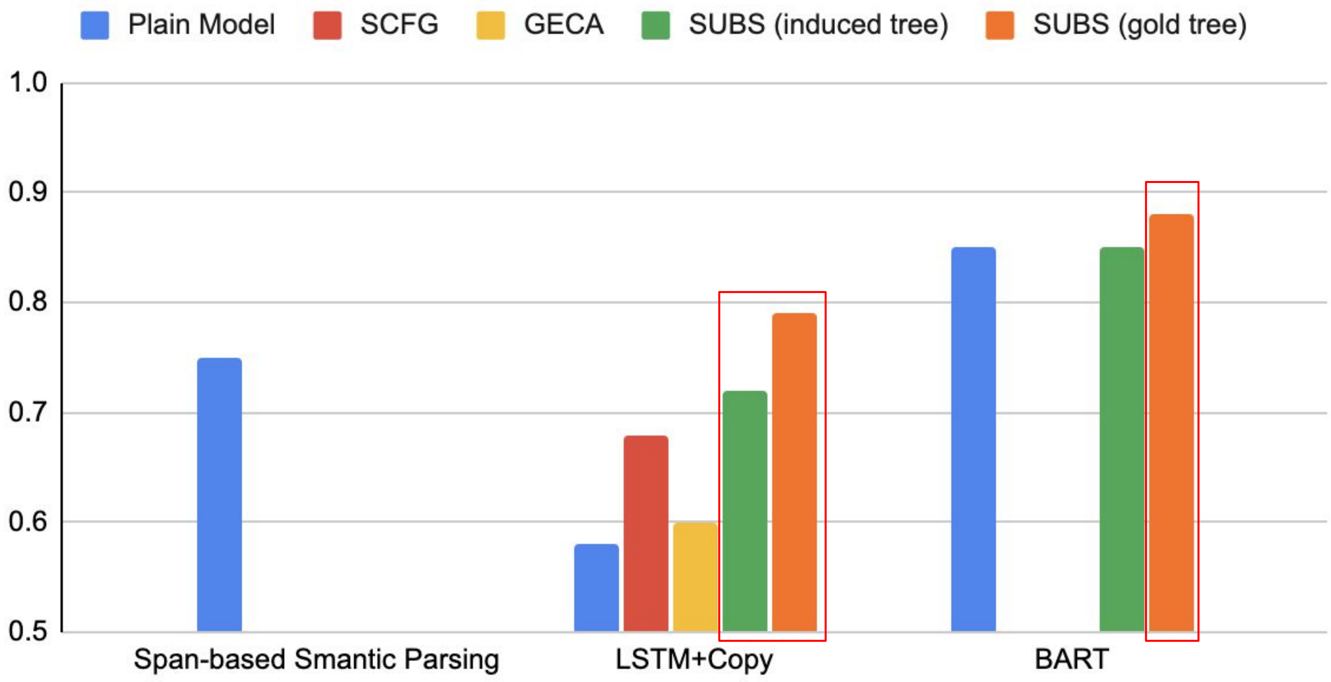
# Results - GeoQuery i.i.d. Split

Data augmentation boost the performance , especially in LSTM based models.



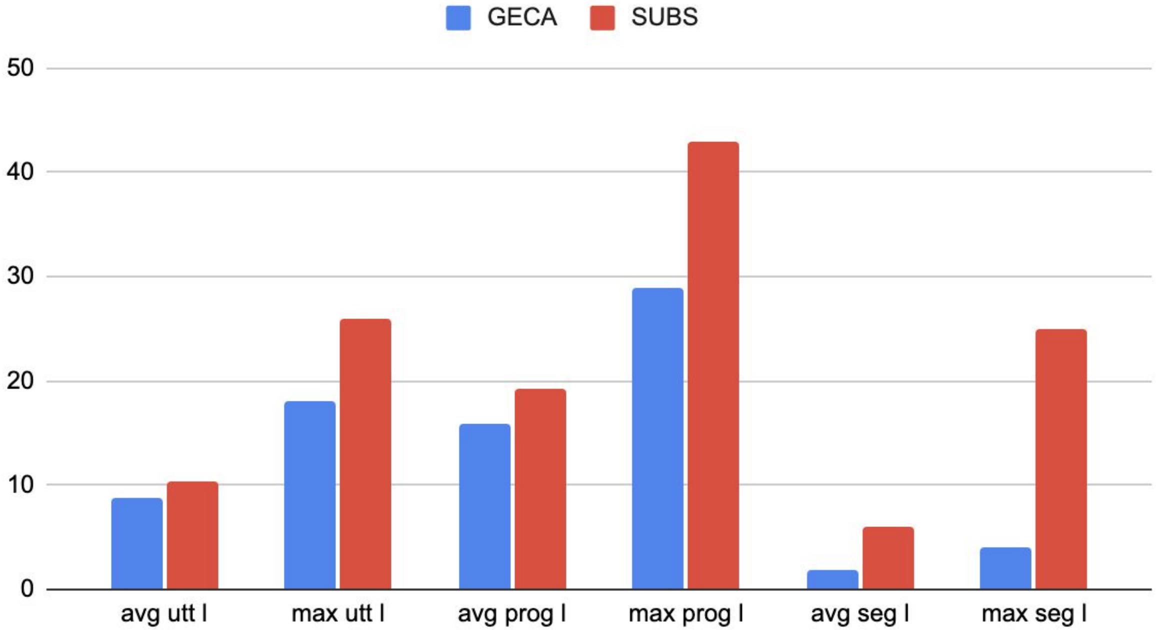
# Results - GeoQuery Compositional Split

SUBS data augmentation is better than others for compositional generalization!



# Analysis of Augmented Data

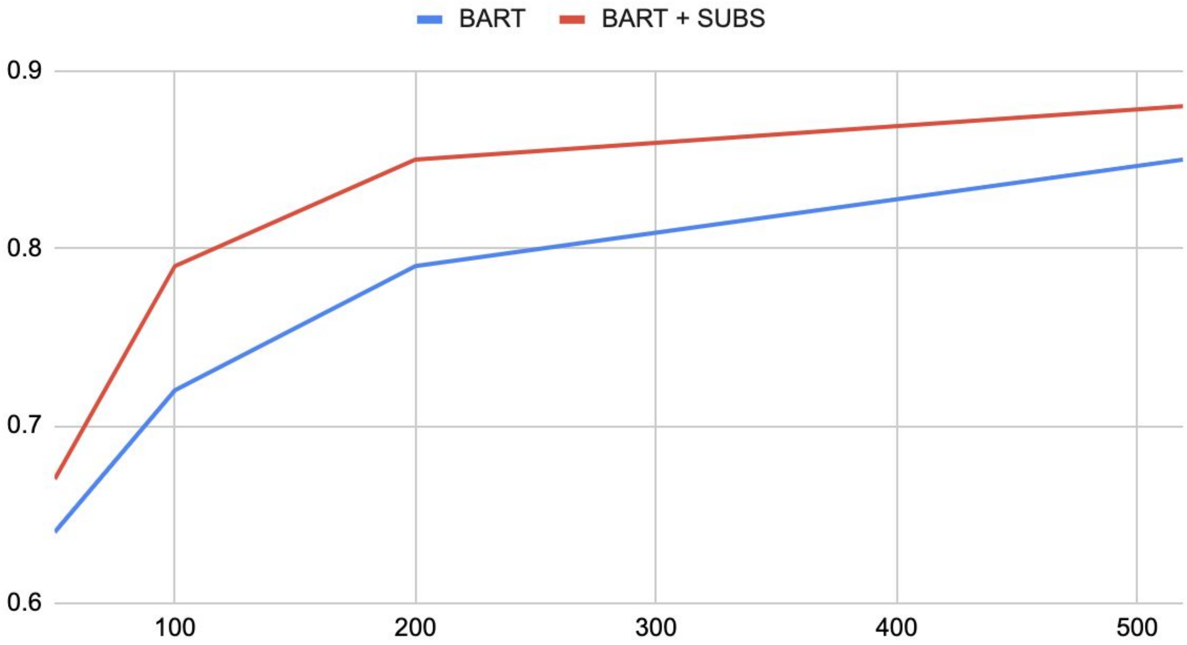
Compared with GECA, SUBS can identify and exchange much more complex structures, and produce more complex utterance and program pairs.



avg: average max: maximal utt: utterance prog: program seg: segment l: length

# Few-shot Settings

The improvement of SUBS is even larger in the few-shot setting!



# SUBS: Subtree Substitution (for Compositional Semantic Parsing)

- Takeaways:
  - Subtree Substitution as a Compositional data augmentation method can help compositional generalization in semantic parsing.
  - Subtree Substitution can identify more complex structures as exchangeable elements, compared with other augmentation methods.

- Authors:
  - Jingfeng Yang
  - Le Zhang
  - Diyi Yang

- Github:
  - https://github.com/GT-SALT/SUBS