CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers

Shiyang Li*, **Semih Yavuz***, Kazuma Hashimoto, Jia Li, Tong Niu,
Nazneen Rajani, Xifeng Yan, Yingbo Zhou, Caiming Xiong

Multi-Domain Dialogue State Tracking

	Dialogue Flow	Turn-level Belief State	Dialogue-level Belief State
Turn 1	[System]: Hello, how can I help you? [User]: I need to find a restaurant in the center.	<restaurant, area, center>	<restaurant, area, center>
Turn 2	[System]: I have many options! Do you have any preference? [User]: It needs to serve British food and I'd like a reservation for 18:00.	<restaurant, food, British>, <restaurant, book time, 18:00>	<restaurant, area, center>, <restaurant, food, British>, <restaurant, book time, 18:00>

MultiWOZ Benchmark

Belief Tracking

Model	MultiWOZ 2.0		MultiWOZ 2.1	
	Joint Accuracy	Slot	Joint Accuracy	Slot
MDBT (Ramadan et al., 2018)	15.57	89.53		
GLAD (Zhong et al., 2018)	35.57	95.44		
GCE (Nouri and Hosseini-Asl, 2018)	36.27	98.42		
Neural Reading (Gao et al, 2019)	41.10			
HyST (Goel et al, 2019)	44.24			
SUMBT (Lee et al, 2019)	46.65	96.44		
SGD-baseline (Rastogi et al, 2019)			43.4	
TRADE (Wu et al, 2019)	48.62	96.92	46.0	
COMER (Ren et al, 2019)	48.79			
MERET (Huang et al, 2020)	50.91	97.07		
DSTQA (Zhou et al, 2019)	51.44	97.24	51.17	97.21
SUMBT+LaRL (Lee et al. 2020)	51.52	97.89		
DS-DST (Zhang et al, 2019)			51.2	
LABES-S2S (Zhang et al, 2020)			51.45	
DST-Picklist (Zhang et al, 2019)	54.39		53.3	
MinTL-BART (Lin et al, 2020)	52.10		53.62	
SST (Chen et al. 2020)			55.23	
TripPy (Heck et al. 2020)			55.3	
SimpleTOD (Hosseini-Asl et al. 2020)			56.45	
ConvBERT-DG + Multi (Mehri et al. 2020)			58.7	
TripPy + CoCoAug (Li and Yavuz et al. 2020)			60.53	

- From ~15% in 2018
- To ~60% in 2020

How well can SOTA DST generalize beyond i.i.d?

data	attraction-name	hotel-name	restaurant-name	taxi-departure	taxi-destination	train-departure	train-destination
dev	94.5	96.4	97.3	98.6	98.2	99.6	99.6
test	96.2	98.4	96.8	95.6	99.5	99.4	99.4

Table 1: The percentage (%) of domain-slot values in dev/test sets covered by training data.

Takeaway-1:

- Possible values for the slots significantly overlap.

slot name	data	area	book day	book time	food	name	price range
book people	train	1.9	38.8	39.2	2.1	16.4	1.5
	dev	1.9	38.9	38.9	1.9	16.3	2.2
	test	2.7	36.9	37.7	1.6	18.7	2.4

Table 2: Co-occurrence distribution(%) of *book people* slot with other slots in *restaurant* domain within the same user utterance. It rarely co-occurs with particulars slots (e.g., *food*), which hinders the evaluation of DST models on realistic user utterances such as "*I want to book a Chinese restaurant for 8 people.*"

Takeaway-2:

- Slot co-occurrence distribution of evaluation sets closely follow training data.

Problem and Objective

Problem:

- Held-out accuracy is often useful, but it usually overestimates (Ribeiro et. al.) the model's actual generalization capability due to i.i.d property of datasets

Objective:

- How can we evaluate DST models beyond the held-out accuracy?

This work:

- A principled, model-agnostic approach to evaluate the generalization capability of DST models to (1) **unseen slot values**, (2) **less frequent but realistic slot combinations**

Ribeiro et. al. Beyond accuracy: Behavioral testing of NLP models with CheckList. ACL'20 Best Paper.

CoCo Pipeline -- Training

Training Phase

CoCo Pipeline -- Training

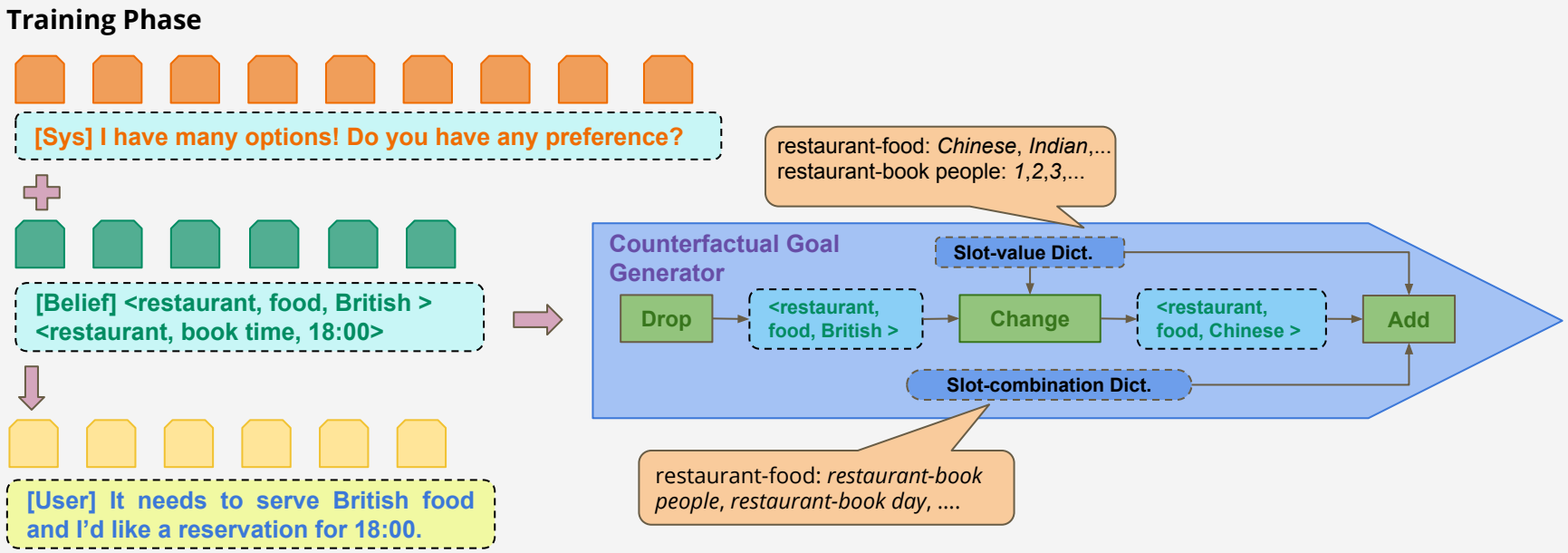
Training Phase

OBJECTIVE

$$p_\theta(U_t^{\text{usr}}|U_t^{\text{sys}}, L_t) = \prod_{k=1}^{n_t} p_\theta(U_{t,k}^{\text{usr}}|U_{t,<k}^{\text{usr}}, U_t^{\text{sys}}, L_t)$$

$$\mathcal{J}_{\text{gen}} = -\sum_{k=1}^{n_t} \log p_\theta(U_{t,k}^{\text{usr}}|U_{t,<k}^{\text{usr}}, U_t^{\text{sys}}, L_t)$$

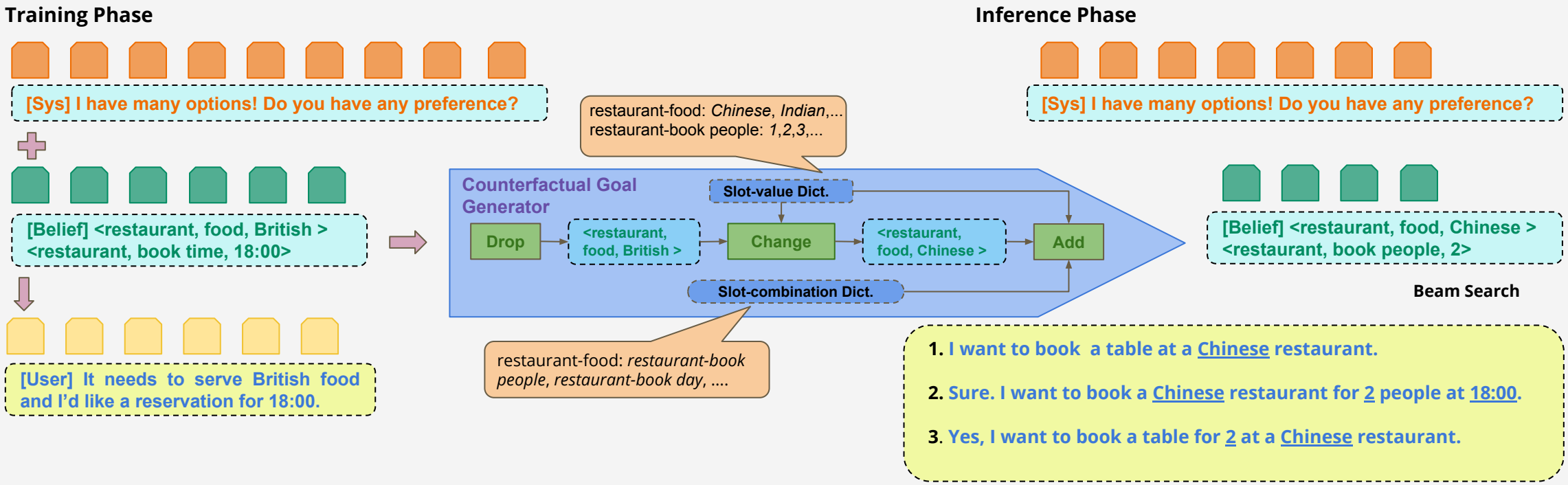
CoCo Pipeline -- Goal Generation



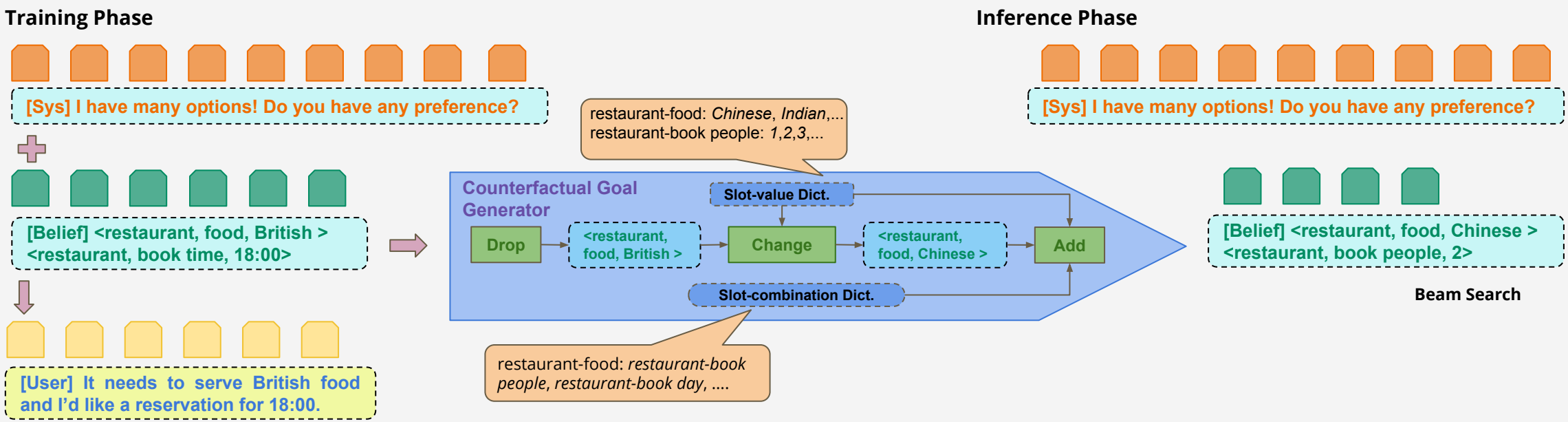
Meta Operations:

- **Drop** a slot
- **Add** a new slot that is consistent with dialogue history
- **Change** the value of a slot

CoCo Pipeline -- Utterance Generation

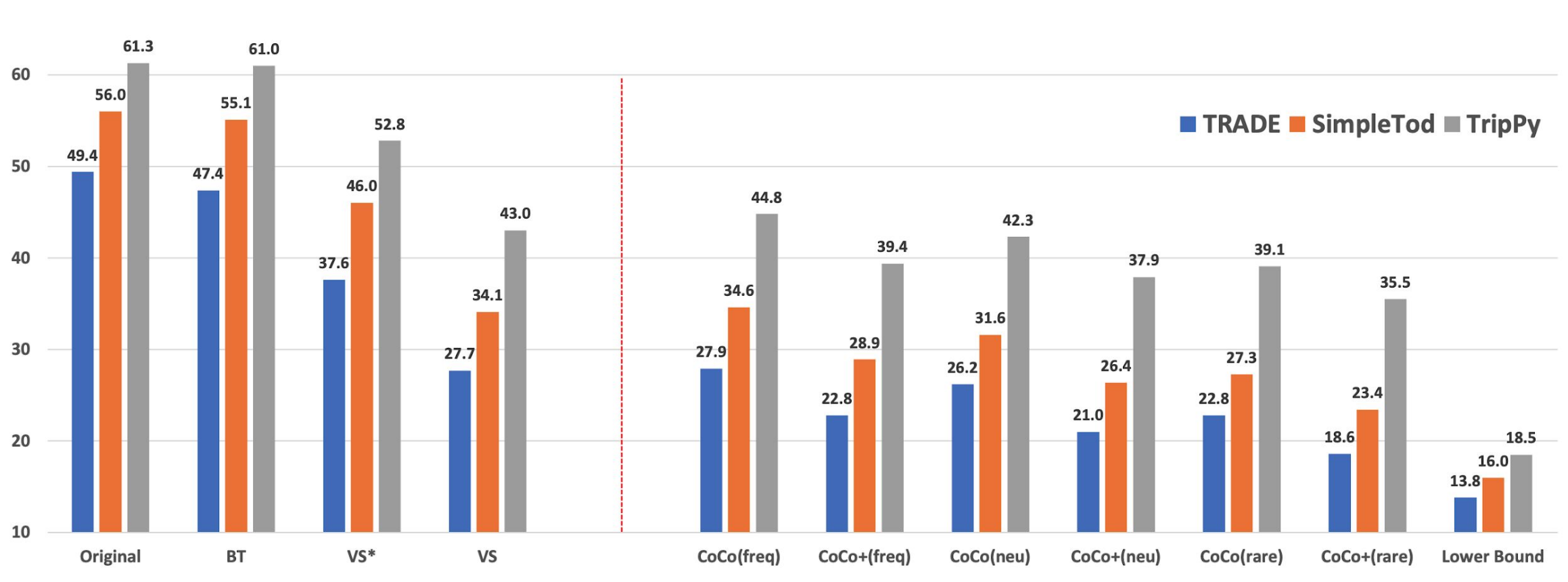


CoCo Pipeline -- Filtering



1. ~~I want to book a table at a Chinese restaurant.~~
   [Classifier Filter: ✓, Slot-Value Match Filter: ×]
2. ~~Sure. I want to book a Chinese restaurant for 2 people at 18:00.~~
   [Classifier Filter: ×, Slot-Value Match Filter: ✓]
3. Yes, I want to book a table for 2 at a Chinese restaurant.
   [Classifier Filter: ✓, Slot-Value Match Filter: ✓]

Main Result



- **BT** indicates back-translation
- **VS** denotes Value-Substitution
- **VS*** uses in-domain value dictionary
- **Freq, neu, rare** indicate the slot-combination dictionary used

Human Evaluation

	Human likeliness	Correctness
Human	87%	85%
CoCo(ori)	90%	91%
CoCo(freq)	90%	99%
CoCo(neu)	79%	98%
CoCo(rare)	82%	96%

Human-likeness:

- 3 annotators give a score of 0/1 to both original user utterances and various CoCo-generated utterances conditioned the original turn-level belief states
- Majority voting on 3 scores

Correctness:

- 3 annotators gives a score of 0/1 to whether the utterance perfectly reflects the turn-level belief state
- Majority voting on 3 scores

- Human evaluations validate that CoCo leads to **high-fidelity**, **human-like** conversations

CoCo as Data Augmentation (CoCoAug)

Model	Joint Goal Accuracy
DSTreader (Gao et al., 2019)	36.40%†
TRADE (Wu et al., 2019)	45.60% †
MA-DST (Kumar et al., 2020)	51.04% †
NA-DST (Le et al., 2020)	49.04% †
DST-picklist (Zhang et al., 2019a)	53.30% †
SST (Chen et al., 2020)	55.23% †
MinTL(T5-small) (Lin et al., 2020)	50.95% §
SimpleTOD (Hosseini-Asl et al., 2020)	55.76% §
ConvBERT-DG+Multi (Mehri et al., 2020)	58.70% §¶
TripPy (Heck et al., 2020)	55.04%*
+ CoCoAug (1×)	56.00%
+ CoCoAug (2×)	56.94%
+ CoCoAug (4×)	59.73%
+ CoCoAug (8×)	**60.53%**

Multi-Fold Generation:

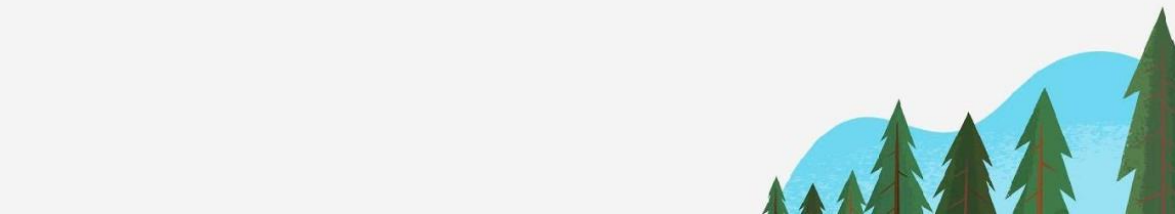
- Sample multiple counterfactual goals
- Apply the same beam-search + filtering pipeline
- Make sure that slot values do not overlap with the test examples

Takeaways:

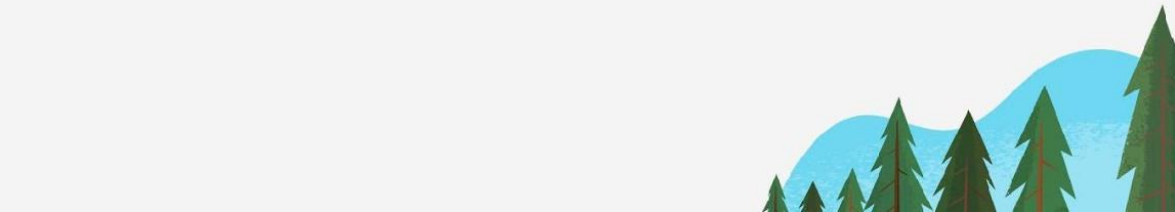
- TripPy significantly benefits from additional conversations generated by CoCo for training
- Accuracy consistently increases as the augmentation size grows
- Using 8x data augmentation provides absolute performance improvement of 5.5%

Conclusion

- SOTA DST models are significantly worse on the CoCo-generated conversations
  - **limitations of relying only on the held-out accuracy**
- CoCo as data augmentation leads to significant gains on TripPy
  - **+ 5.5% improvement on MultiWOZ test set**
- We hope this work inspires further research on developing and using stronger evaluation strategies towards building **more generalizable TOD systems**



Thanks!



- CODE: https://github.com/salesforce/coco-dst
- PAPER: https://openreview.net/pdf?id=eom0IUrF__F
- Blog-Post: https://blog.einstein.ai/coco-dst/