Has it been difficult to compare your dialogue state tracking model beyond accuracy on an in-distribution test set?

Conduct comprehensive dialogue state tracking diagnostics to discover strengths and weaknesses and overlooked opportunities for improvement.

Know Thy Strengths: Comprehensive Dialogue State Tracking Diagnostics

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, Ahmad Beirami

Problem: sparse and uncoordinated comparisons beyond joint goal accuracy

Entity Bias (Qian et al. 2022)

`[user] I would like to leave from **London**.` → `train departure cambridge`

Unseen Entities (Heck et al. 2020)

Training entities
Cambridge, London, Liverpool

Test entities
Cambridge, London, Liverpool — **JGA: 92%**

`[user] Can you find me a five star hotel in LA with free parking?`

Test entities
Abu Dhabi, Dubai, Doha — **JGA: 15%**

`[user] I am uh looking for a 5 star hotel in um Pasadena no Los Angeles that offers free parking.`

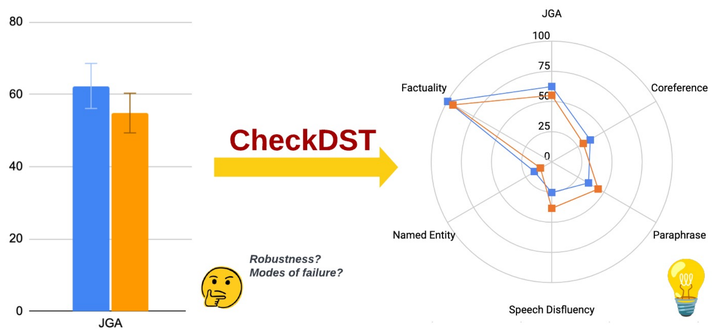
Language Variety (Heck et al. 2020)

`[user] I am looking for a 5 star hotel in Los Angeles that offers free parking.`

- Joint goal accuracy (JGA) on an in-distribution test set does not capture enough information as dialogue state tracking (DST) models are brittle.
- This is a known problem and previous work have done individual analysis beyond JGA. But they are sparse and uncoordinated, making it difficult to holistically compare strengths and weaknesses of DST models.

Solution: CheckDST

Toolkit that facilitates consolidated robustness evaluation and failure mode analytics



	Input Example	Example DST Prediction
Original	*I would like to leave from Cambridge.*	`train departure cambridge`
Paraphrase	*Please book me one ticket departing from Cambridge.*	`train departure cambridge`
Speech Disfluency	*I would like to uh leave from London no I meant Cambridge.*	`train departure cambridge`
Unseen Entity	*I would like to leave from mbadgceir.*	`train departure mbadgceir`
Coreference	*I need you to book the restaurant for Tuesday.* ... *I'm also looking for a train to London Kings Cross on the same day as the restaurant booking.*	... `restaurant day tuesday` `train day tuesday` ...
Hallucination	*I would like to leave from London.*	`train departure cambridge` `train departure london`

Original

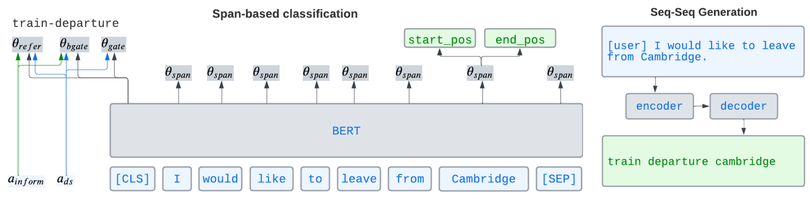
`[user] I would like to leave from Cambridge.` → `train departure cambridge`

Perturbed

`[user] Please book me one ticket departing from Cambridge.` → `train departure cambridge`

- Consolidate robustness evaluation and failure mode analytics to easily compare models more holistically.
- Measure robustness with Consistent JGA (cJGA): cJGA=1 when both original instance and perturbed instance are correctly predicted.
- Quantify failure modes and performance on challenging cases: hallucination frequency and coreference resolution

Experimental setup



Models
- Span-based classification models (SCLS)
- SimpleTOD-style Seq-Seq generation models (GEN)

Dataset
- Original train/val/test: MultiWOZ 2.3 (Han et al. 2020)
- Paraphrase & speech disfluency perturbations from LAUG (Liu et al. 2021)
- Named entity replacements (Huang et al. 2021)

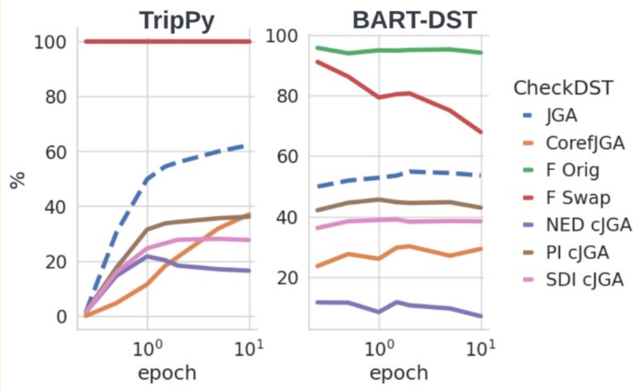
Metrics
- Paraphrase invariant cJGA (PI cJGA)
- Speech disfluency invariant cJGA (SDI cJGA)
- Named Entity Directional cJGA (NED cJGA)
- CorefJGA: JGA on cases that require coreference resolution
- Factuality (F): 1 – hallucination frequency

CheckDST quickly quantifies relative strengths and weaknesses

between different models and also the same model at different points of training

	Model	JGA	CorefJGA	PI cJGA	SDI cJGA	NED cJGA	F_{orig}	F_{swap}
SCLS	TripPy (2020)	$62.3_{0.2}$	$37.0_{0.7}$	$36.2_{0.1}$	$27.7_{0.5}$	$16.5_{0.4}$	100_{0}	100_{0}
	ConvBERT-DG (2020)	$62.0_{0.1}$	$36.6_{0.7}$	$35.9_{0.2}$	$29.5_{0.4}$	$16.2_{0.1}$	100_{0}	100_{0}
GEN	BART-DST (2020)	$52.5_{0.2}$	$25.5_{1.2}$	$43.0_{0.8}$	$36.5_{1.3}$	$9.8_{0.7}$	$94.8_{0.2}$	$75.4_{1.8}$
	SOLOIST (2021a)	$54.8_{0.4}$	$30.4_{1.1}$	$44.5_{0.6}$	$38.5_{0.5}$	$10.7_{0.4}$	$94.9_{0.1}$	$81.1_{1.6}$
	MUPPET-DST (2021)	$54.9_{0.4}$	$29.9_{1.9}$	$45.4_{0.4}$	$39.1_{0.7}$	$7.8_{1.4}$	$94.6_{0.1}$	$68.4_{3.8}$

1. Inter-model comparison
- ↑JGA ≠ ↑Robustness
- SCLS vs GEN: each have **clear strengths and weaknesses**
- SCLS is more robust to unseen named entities and coreferences.
- GEN: more robust to language variety



2. Intra-model comparison
- ↑JGA ≠ ↑Robustness
- Both SCLS and GEN models share same trend of **dropping robustness as training progresses**, despite gains in JGA.
- This trend starts even from the start for GEN models and thus encourages few-shot / zero-shot approaches.

TripPy (Heck et al. 2020)

restaurant name: *osdijkal* (Epoch 1) → restaurant name: (none) (Epoch 10)

SCLS gives up on prediction

BART-DST (Lewis et al. 2020, Qian et al. 2021)

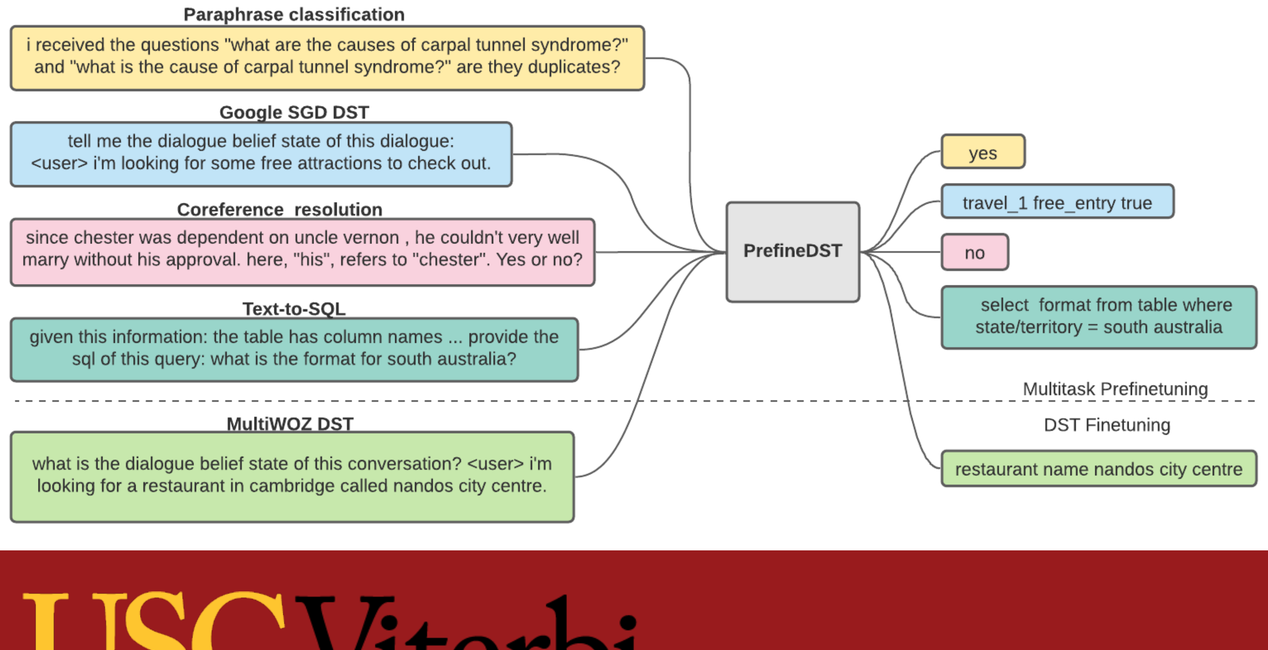
restaurant name: *osdijkal* (Epoch 1) → restaurant name: *oskal* (Epoch 10)

GEN incorrectly copies subsequence of original slot value but predicts correct slot type → more robust behavior

3. Failure mode comparison
- Same drop in robustness but different patterns of failure.

CheckDST can guide holistic performance gains!

PrefineDST targets exposed weaknesses with pre-finetuning to attain comprehensive improvements.



Model	JGA	CorefJGA	PI cJGA	SDI cJGA	NED cJGA	F_{orig}	F_{swap}
BART-DST	$52.5_{0.2}$	$25.5_{1.2}$	$43.0_{0.8}$	$36.5_{1.3}$	$9.8_{0.7}$	$94.8_{0.2}$	$75.4_{1.8}$
SOLOIST	$54.8_{0.4}$	$30.4_{1.1}$	$44.5_{0.6}$	$38.5_{0.5}$	$10.7_{0.4}$	$94.9_{0.1}$	$81.1_{1.6}$
MUPPET-DST	$54.9_{0.4}$	$29.9_{1.9}$	$45.4_{0.4}$	$39.1_{0.7}$	$7.8_{1.4}$	$94.6_{0.1}$	$68.4_{3.8}$
PrefineDST	$55.7_{0.3}$	$30.5_{0.5}$	$46.1_{1.0}$	$41.8_{1.0}$	$11.0_{0.5}$	$95.2_{0.2}$	$76.7_{1.4}$

	Average Δ from Best ↓
TripPy (Heck et al., 2020)	4.55
ConvBERT-DG (Mehri et al., 2020)	4.57
SimpleTOD (Hosseini-Asl et al., 2020)	6.93
SOLOIST (Peng et al., 2021a)	4.98
MUPPET-DST (Aghajanyan et al., 2021)	5.40
PrefineDST	**3.92**

- CheckDST exposes failure modes and brittleness, **guiding the development of more robust DST models.**
- **PrefineDST is pre-finetuned with tasks that intuitively addresses these weaknesses**, such as paraphrase understanding and various parsing tasks that require correctly copying spans from the input.
- **PrefineDST shows most well-rounded performance** and shows promise towards robust DST models.

USC Viterbi Information Sciences Institute

CUTE LAB NAME

Meta AI Google AI

Website: justin-cho.com/checkdst
Data: github.com/wise-east/checkdst
Contact: Justin Cho (hd.justincho@gmail.com)