



# Granular Change Accuracy: A More Accurate Performance Metric for Dialogue State Tracking

---

Taha Aksu and Nancy F. Chen

# Performance Metrics' Weaknesses

## Conversation



## Dialog State

Slot	G	P1	P2
Hotel-internet	yes	✗	✓

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗
Hotel-day	Sunday	✓	✗
Hotel-people	6	✓	✗
Hotel-stay	4	✓	✗
Hotel-price	cheap	✓	✗
Hotel-type	guesthouse	✓	✗

## Number of Correct Predictions

P1 performs significantly better!

P1	P2
0	1

0	0
---	---

5	0
---	---

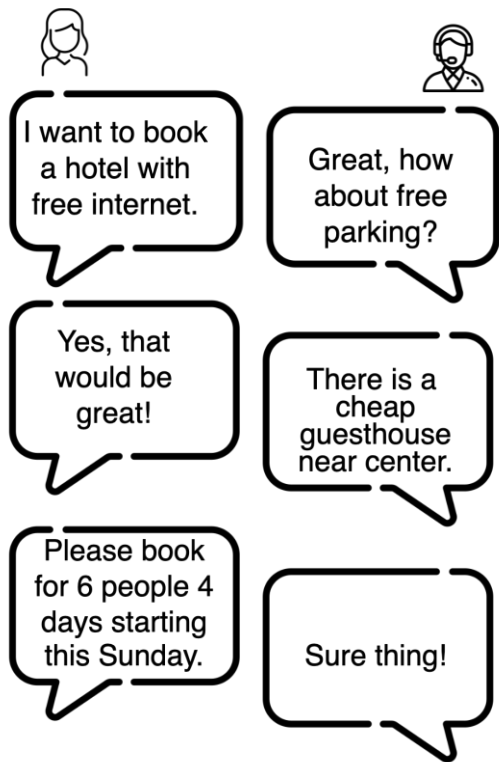
Total:

5/7

1/7

# Performance Metrics' Weaknesses

## Conversation



## Dialog State

Slot	G	P1	P2
Hotel-internet	yes	✗	✓

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗
Hotel-day	Sunday	✓	✗
Hotel-people	6	✓	✗
Hotel-stay	4	✓	✗
Hotel-price	cheap	✓	✗
Hotel-type	guesthouse	✓	✗

## DST Performance Metric Results

P1 performs significantly better!

	<u>P1</u>	<u>P2</u>
<u>JGA</u>	0.0	<b>33.33</b>
<u>SA</u>	<b>94.44</b>	92.22
<u>AGA</u>	23.81	<b>54.76</b>
<u>RSA</u>	23.81	<b>54.76</b>
<u>FGA</u>	13.12	<b>33.33</b>

- 4 out of 5 of the DST metrics evaluate P2 as the better model.
- SA gives inflated and very similar scores to both models.

# Performance Metrics' Weaknesses

## Conversation

## Dialog State

## Weaknesses



1 Slot

I want to book a hotel with free internet.

Great, how about free parking?

2 Slot

Yes, that would be great!

There is a cheap guesthouse near center.

7 Slot

Please book for 6 people 4 days starting this Sunday.

Sure thing!

Slot	G	P1	P2
Hotel-internet	yes	✗	✓

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗

Slot	G	P1	P2
Hotel-internet	yes	✗	✓
Hotel-parking	yes	✗	✗
Hotel-day	Sunday	✓	✗
Hotel-people	6	✓	✗
Hotel-stay	4	✓	✗
Hotel-price	cheap	✓	✗
Hotel-type	guesthouse	✓	✗

JGA=0

1

### Double-counting Scores

The same prediction is scored multiple times

JGA, FGA, SA, RSA, AGA

2

### Turn-centric Scores

Accuracy averaged over turns assuming uniform distribution

JGA, FGA, SA, RSA, AGA

3

### 0/1 Scores

Turns do not get partial scores

JGA, FGA

# What makes these metrics weak?

---

## JGA

- Calculates accuracy of perfect turns.
- Single error means a score of 0 for the turn.
- Mistaken predictions are propagated.
- Early mistakes are over penalized.

## FGA

- Modifies JGA with a decreasing penalty for propagated turns.
- Still harshly evaluates turns for a single mistake.

## SA

- Calculates accuracy across all predefined slots which results in overestimation of performance.

## AGA and RSA

- **AGA** calculates a recall value for all turns with non-empty ground truth belief states and returns the average.
- **RSA** modifies SA to calculate accuracy over just the active slots (very similar to AGA).
- Both AGA and RSA still computes the average over each turn, resulting in the repeated counting of errors and correct predictions for each turn, *i.e. under/over-estimation*.

# A more accurate DST Metric: Granular Change Accuracy

---

## Weaknesses

1. Calculates performance over slots only when their value is modified in the last turn.

1

### **Double-counting Scores**

The same prediction is scored multiple times

JGA, FGA,  
SA, RSA, AGA

2. Takes the average over the total number of modifications.

2

### **Turn-centric Scores**

Accuracy averaged over turns assuming uniform distribution

JGA, FGA,  
SA, RSA, AGA

3. Calculates performance over slots rather than turns.

3

### **0/1 Scores**

Turns do not get partial scores

JGA, FGA

# Granular Change Accuracy (GCA)

Slot label	Ground Truth   Prediction	Count Type
Train Destination	Cambridge   "None"	Missed
Train Source	Birmingham   London	Wrong
Train Arrive-by	17:00   17:00	Correct
Train Leave-at	"None"   14:00	Overshot

## Label Precision

$$L_P = \frac{C + W}{P}$$

## Label Recall

$$L_R = \frac{C + W}{G}$$

## Value Precision

$$V_P = \frac{C}{P}$$

## Value Recall

$$V_R = \frac{C}{G}$$

- $G = C + W + M$  is the total number of gold values.
- The combined counts  $C + W$  pertain to instances where the slot detection was correct, even if the subsequent value prediction might not be accurate.

# Granular Change Accuracy (GCA)

Slot label	Ground Truth   Prediction	Count Type
Train Destination	Cambridge   "None"	Missed
Train Source	Birmingham   London	Wrong
Train Arrive-by	17:00   17:00	Correct
Train Leave-at	"None"   14:00	Overshot

## Label Precision

$$L_P = \frac{C + W}{P}$$

## Label Recall

$$L_R = \frac{C + W}{G}$$

## Value Precision

$$V_P = \frac{C}{P}$$

## Value Recall

$$V_R = \frac{C}{G}$$

Final metric is formed similar to re-knowned F1 metric:

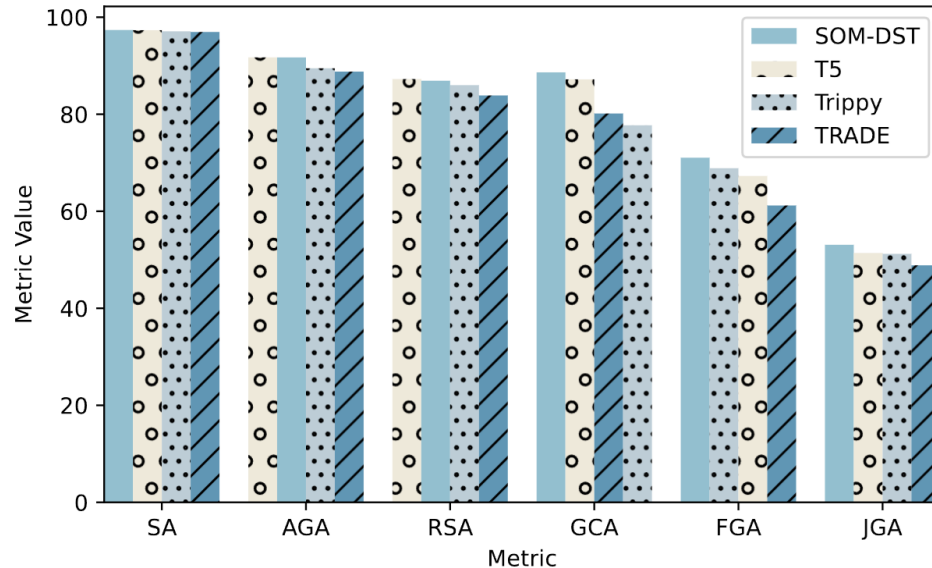
$$GCA = \text{Harmonic\_mean}(V_P, V_R, L_P, L_R)$$

# Experiments

---

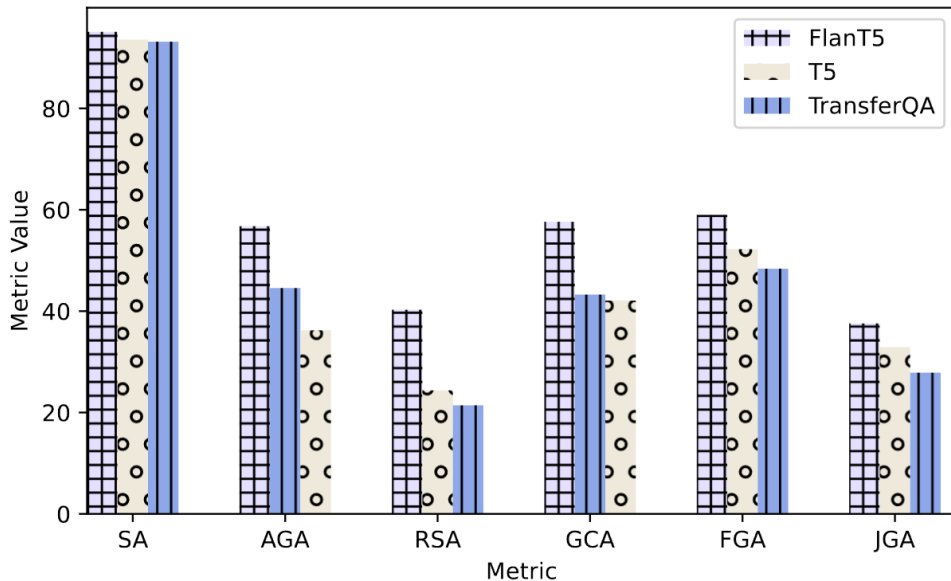
- We conduct experiments with **MultiWOZ 2.1** and **SGD** datasets in various different settings:
  - Full-shot
  - Few-shot
  - Zero-shot
- We evaluate 6 DST models across these settings:
  - **TRADE**
  - **SOM-DST**
  - **T5**
  - **Trippy**
  - **TransferQA**
  - **FlanT5**

# Full-shot Results



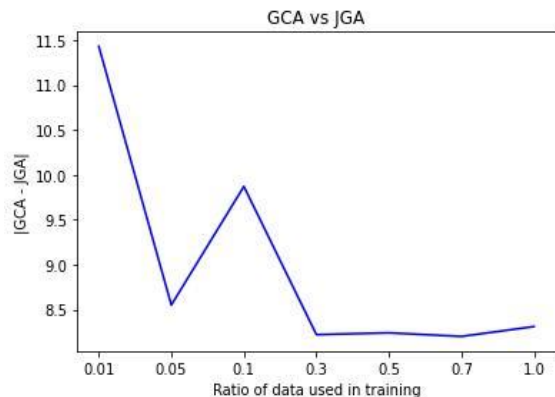
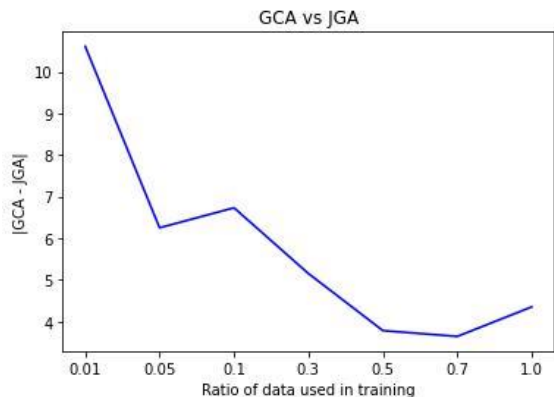
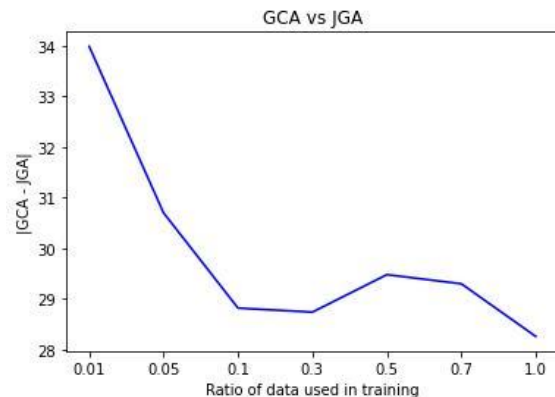
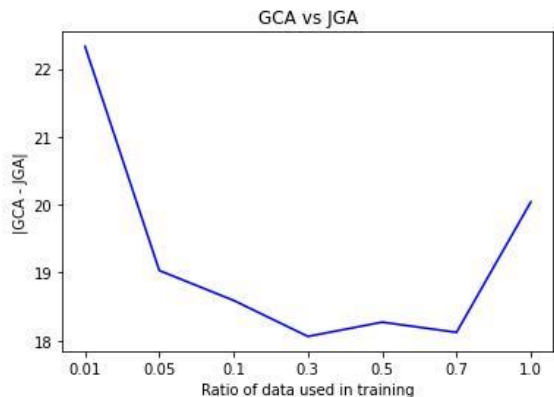
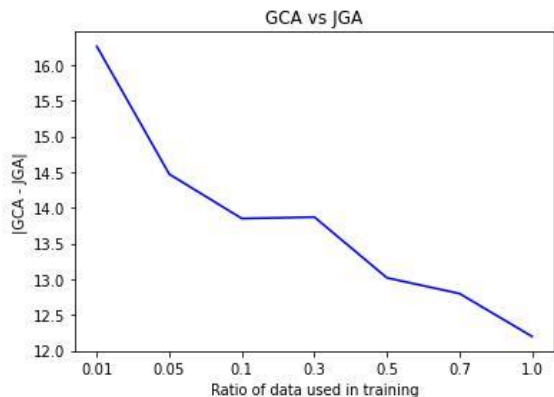
- JGA and FGA tend to produce lower performance scores owing to their binary scoring approach.
- SA and AGA exhibit inflated scores, reflecting their tendency to overestimate.
- GCA and RSA scores are positioned between these two extremes.

# Zero-shot Results



- Double-counting results in lower scores for JGA.
- FGA's turn-centric scoring counts most empty turns correctly boosting the final score of a model.
- Whenever a turn does not have any active slots RSA scores that turn as 0. This drags the average score of the dialogue down (due to turn-centric scoring).
- GCA counts each mistake once at the first encounter and calculates the performance by aggregating accuracy over the model's actions rather than turns taking the middle ground.

# Few-shot



👁️ When training data is scarcer JGA exhibits increased differences in comparison to GCA.

💡 The identified weaknesses become more pronounced when evaluating models trained on limited data.

# Fine-grained Analysis

---

To analyze edge cases, we examined 20 predictions of TRADE and SOM-DST models where FGA and GCA show the largest disagreement:

1

FGA overestimates the performance when errors are accumulated in a few turns, or when the mistakes are **not uniformly distributed**.

2

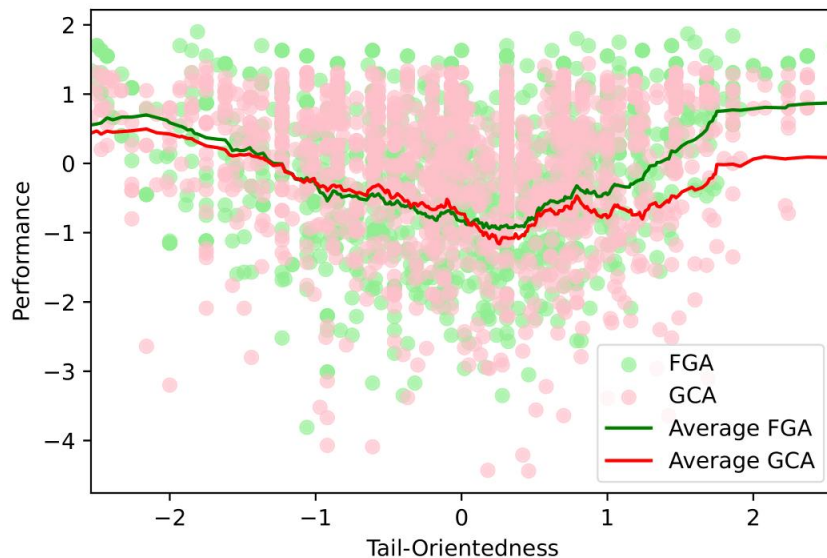
The effect is prominent if these accumulations occur in the later part of the dialog, or when the mistakes show a **tail-oriented distribution**.

# Tail-Orientedness (TO)

- To examine the impact of tail-oriented mistakes on FGA and GCA evaluation, we introduce a new measure:

$$TO = \frac{E_t - \left(\frac{n-1}{2}\right)}{n} \quad E_t = \frac{\sum_{i=0}^m t_i}{m}$$

- It calculates the average distance of each mistake's turn from the middle turn of the dialog.
- Despite GCA consistently yielding higher results for lower values FGA shows similar or even higher scores at the right-hand side of the figure.
- This suggests that as dialog state mistakes become more tail-oriented, FGA tends to overestimate the performance.

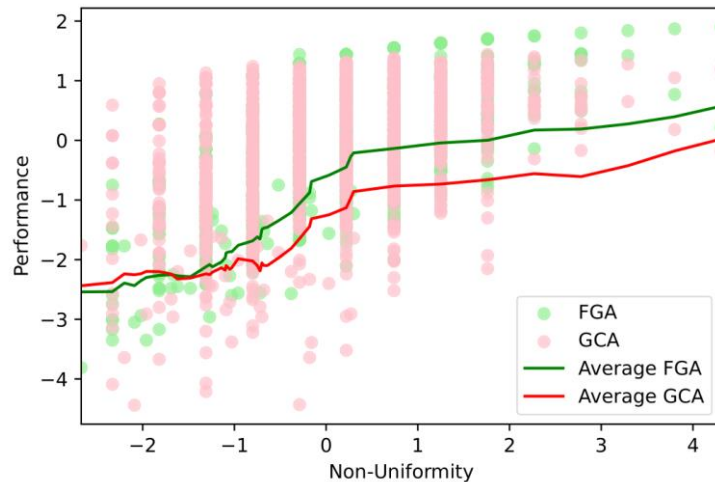


# Non-uniformity (NU)

- In a similar manner we define a non-uniformity measure:

$$NU = \frac{\sum_{t=0}^n |m_t - E_m|}{E_m} \quad E_m = \frac{m}{n}$$

- FGA generally exhibits lower values compared to GCA on left hand side of the figure, however, one can observe GCA going higher as NU values increase.
- This suggests that FGA is adversely affected by the uniform spread of prediction errors.



# Correlations of FGA and GCA with TO and NU

---

- Finally, we further calculate the Pearson Correlation Coefficients of both FGA and GCA with both spurious traits across dialogs.
- The correlations between TO and FGA / GCA are 0.08/-0.05
- The correlations between NU and FGA / GCA are 0.59/0.40 respectively.
- The differences between these correlations are significant according to Zou's confidence interval tests\*.
- FGA's correlation with both features is significantly stronger with a 95% confidence level.
- These results show that GCA is less susceptible to spurious features than FGA.

# Summary

---

- DST metrics may over/under estimate performance under common scenarios.
- These happen mostly due to three main weaknesses shared across current metrics:
  1. Double counting predictions
  2. Turn-centric scores
  3. 0/1 scores
- We propose a new metric, GCA, which addresses these weaknesses by design.
- We show through a comprehensive set of experiments that GCA is less prone to spurious features that earlier metrics has shown.

# Thanks!

---



Code:

[https://github.com/cuthalienn/Granular\\_Change\\_Accuracy](https://github.com/cuthalienn/Granular_Change_Accuracy)



Paper:

<https://arxiv.org/abs/2403.11123v1>

# THANKS!

# Questions?

Contact me: [taksu@u.nus.edu](mailto:taksu@u.nus.edu)