ACL 2022

22ND – 27TH MAY | 60TH MEETING | DUBLIN

# Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking

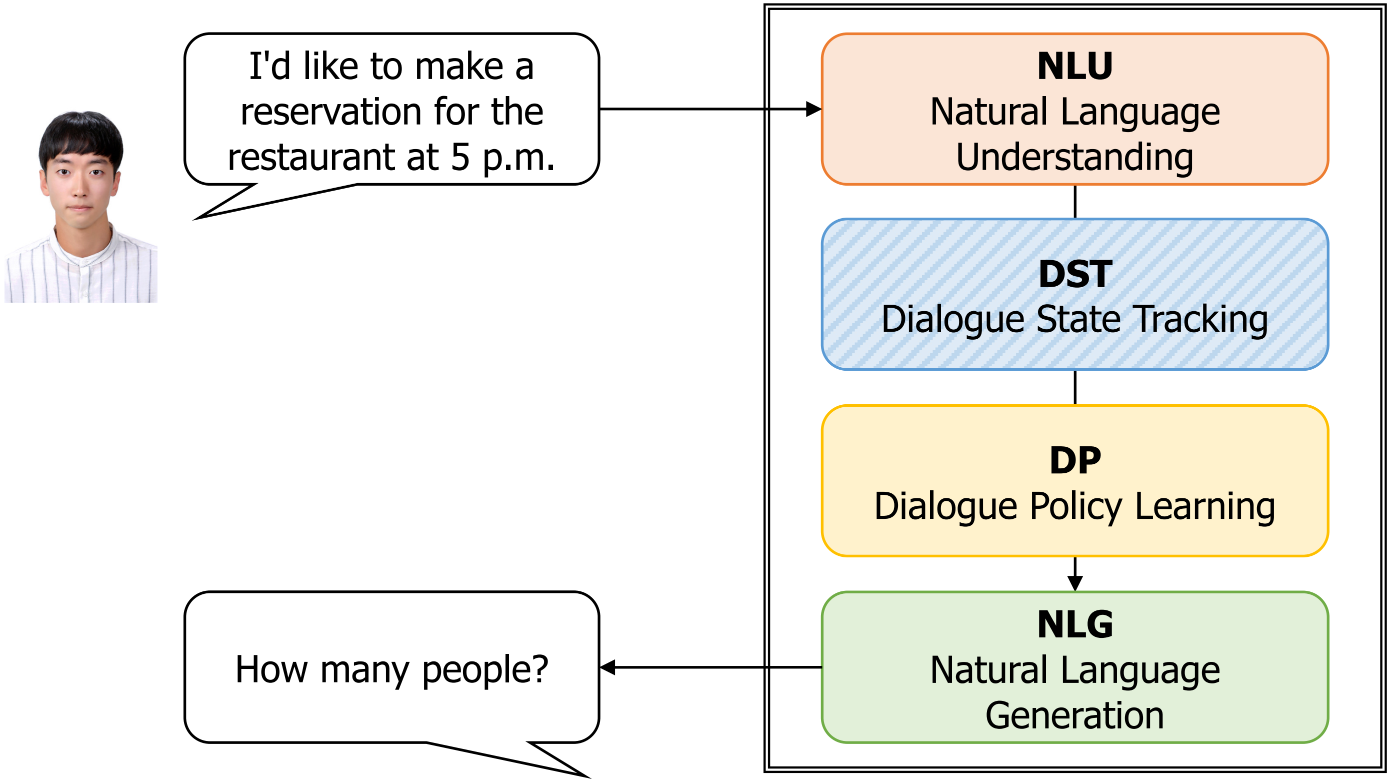
**Takyoung Kim** [1], **Hoonsang Yoon** [1], **Yukyung Lee** [1], **Pilsung Kang** [1], **and Misuk Kim** [2]

[1] Korea University, Seoul, Republic of Korea

[2] Sejong University, Seoul, Republic of Korea

# Dialogue State Tracking (DST)

- DST is a core component of a task-oriented dialogue system



I'd like to make a reservation for the restaurant at 5 p.m.

**NLU**
Natural Language Understanding

**DST**
Dialogue State Tracking

**DP**
Dialogue Policy Learning

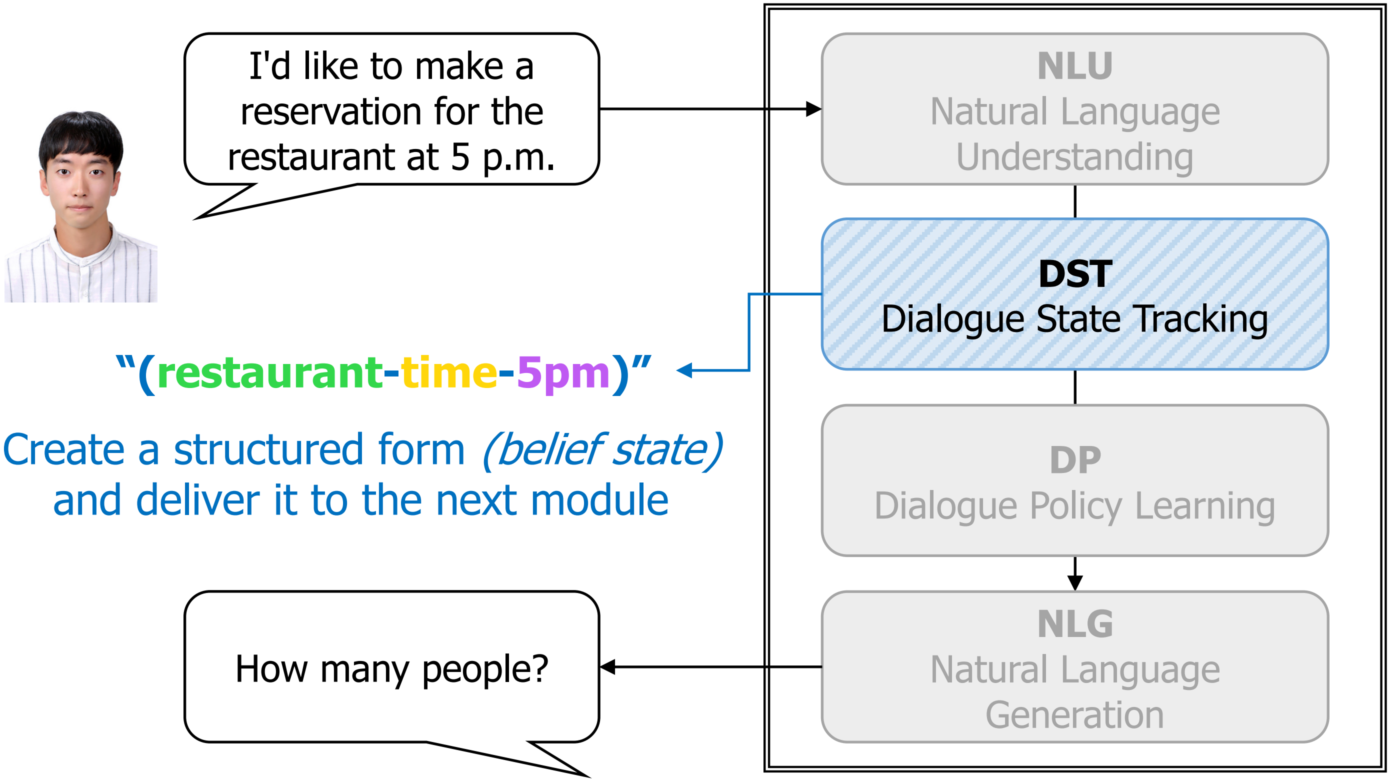
**NLG**
Natural Language Generation

How many people?

*NeurIPS 2020 Tutorial: Deeper Conversational AI*

# Dialogue State Tracking (DST)

- DST is a core component of a task-oriented dialogue system
- "Belief state" presents **domain**, **slot**, and **value** of specific dialogue situation



I'd like to make a reservation for the restaurant at 5 p.m.

**NLU**
Natural Language Understanding

**DST**
Dialogue State Tracking

**DP**
Dialogue Policy Learning

**NLG**
Natural Language Generation

**"(restaurant-time-5pm)"**

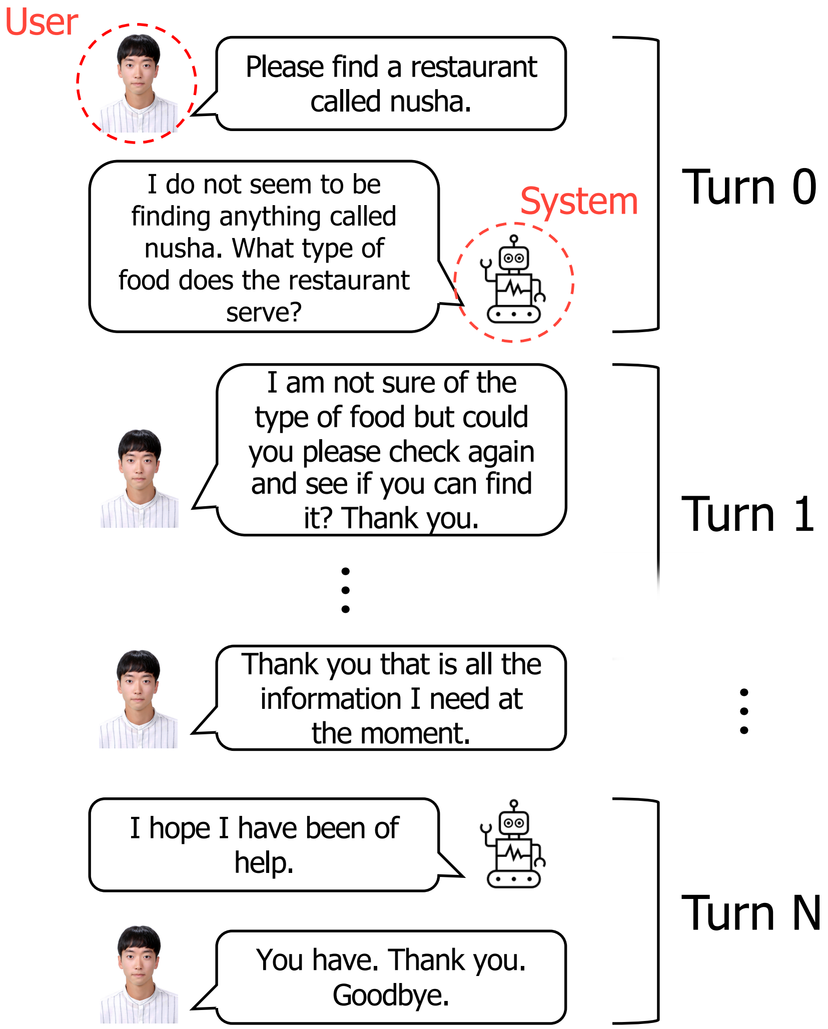
Create a structured form *(belief state)* and deliver it to the next module

How many people?

*NeurIPS 2020 Tutorial: Deeper Conversational AI*

# Dialogue State Tracking (DST)

- Most used MultiWOZ dataset has accumulated multi-turn structure



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

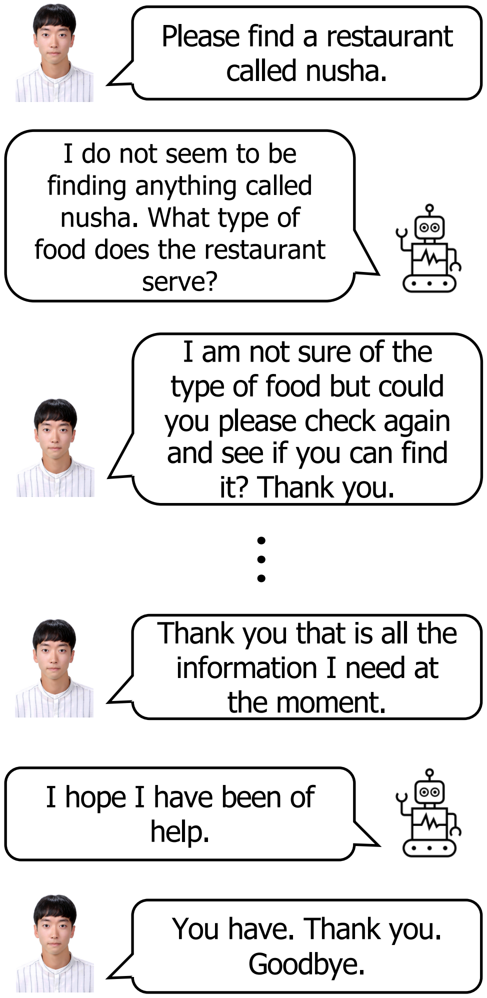
Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

# Dialogue State Tracking (DST)

- Most used MultiWOZ dataset has accumulated multi-turn structure



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

| Turn | Gold State (domain-slot-value) | |
|---|---|---|
| 0 | - | → Evaluate |
| 1 | - | → Evaluate |
| 2 | attraction-name-nusha | → Evaluate |
| 3 | attraction-name-nusha | → Evaluate |
| 4 | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian | → Evaluate |
| 5 | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | → Evaluate |
| 6 | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | → Evaluate |

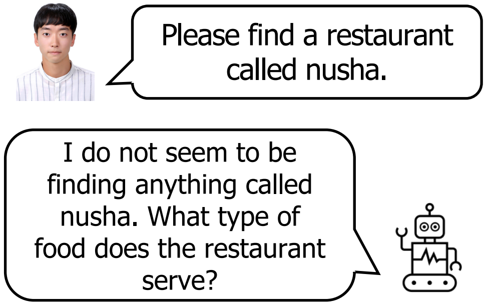
# Dialogue State Tracking (DST)

- Joint goal accuracy (JGA) and slot accuracy (SA) are mainly used

| Method | Metric | Dataset |
|---|---|---|
| DST-STAR (Ye et al., 2021) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| Seq2Seq-DU (Feng et al., 2021) | JGA | SGD, MultiWOZ 2.1, MultiWOZ 2.2 |
| L4P4K2-DSGraph (Lin et al., 2021) | JGA, SA | MultiWOZ 2.0 |
| Transformer-DST (Zeng and Nie, 2021) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| NA-DST (Le et al., 2020) | JGA, SA | MultiWOZ 2.0, MultiWOZ 2.1 |
| TripPy (Heck et al., 2020) | JGA | WOZ 2.0, MultiWOZ 2.1, Sim-M, Sim-R |
| SOM-DST (Kim et al., 2020) | JGA, SA | MultiWOZ 2.0, MultiWOZ 2.1 |
| Simple-TOD (Hosseini-Asl et al., 2020) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| GCDST (Wu et al., 2020) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| CSFN-DST (Zhu et al., 2020) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| SAVN (Wang et al., 2020) | JGA, SA | MultiWOZ 2.0, MultiWOZ 2.1 |
| SST (Chen et al., 2020) | JGA, SA | MultiWOZ 2.0, MultiWOZ 2.1 |
| DS-DST (Zhang et al., 2020) | JGA | MultiWOZ 2.0, MultiWOZ 2.1 |
| DSTQA (Zhou and Small, 2019) | JGA, SA | WOZ 2.0, MultiWOZ 2.0, MultiWOZ 2.1 |
| SUMBT (Lee et al., 2019) | JGA | WOZ 2.0, MultiWOZ 2.0 |
| DST-Reader (Gao et al., 2019) | JGA | MultiWOZ 2.0 |
| BERT-DST (Chao and Lane, 2019) | JGA | WOZ 2.0, Sim-M, Sim-R, DSTC2 |
| TRADE (Wu et al., 2019) | JGA, SA | MultiWOZ 2.0 |
| Hyst (Goel et al., 2019) | JGA | MultiWOZ 2.0 |
| COMER (Ren et al., 2019) | JGA | WOZ 2.0, MultiWOZ 2.0 |

# Joint Goal Accuracy (JGA)

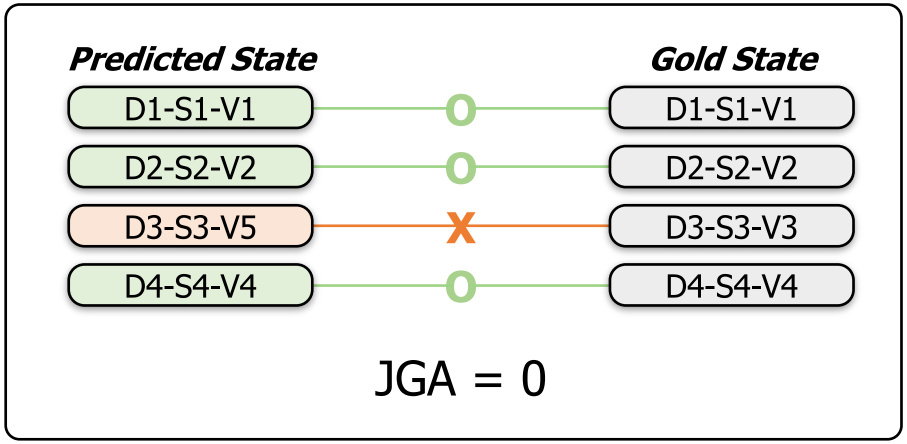
$$JGA = \begin{cases} 1 & \text{if predicted state} = \text{gold state} \\ 0 & \text{otherwise} \end{cases}$$



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

Decides whether the model's prediction "perfectly" matches with the ground truth



| Predicted State | | Gold State |
|---|---|---|
| D1-S1-V1 | O | D1-S1-V1 |
| D2-S2-V2 | O | D2-S2-V2 |
| D3-S3-V5 | X | D3-S3-V3 |
| D4-S4-V4 | O | D4-S4-V4 |

JGA = 0

# Joint Goal Accuracy (JGA)

$$JGA = \begin{cases} 1 & \text{if predicted state} = \text{gold state} \\ 0 & \text{otherwise} \end{cases}$$

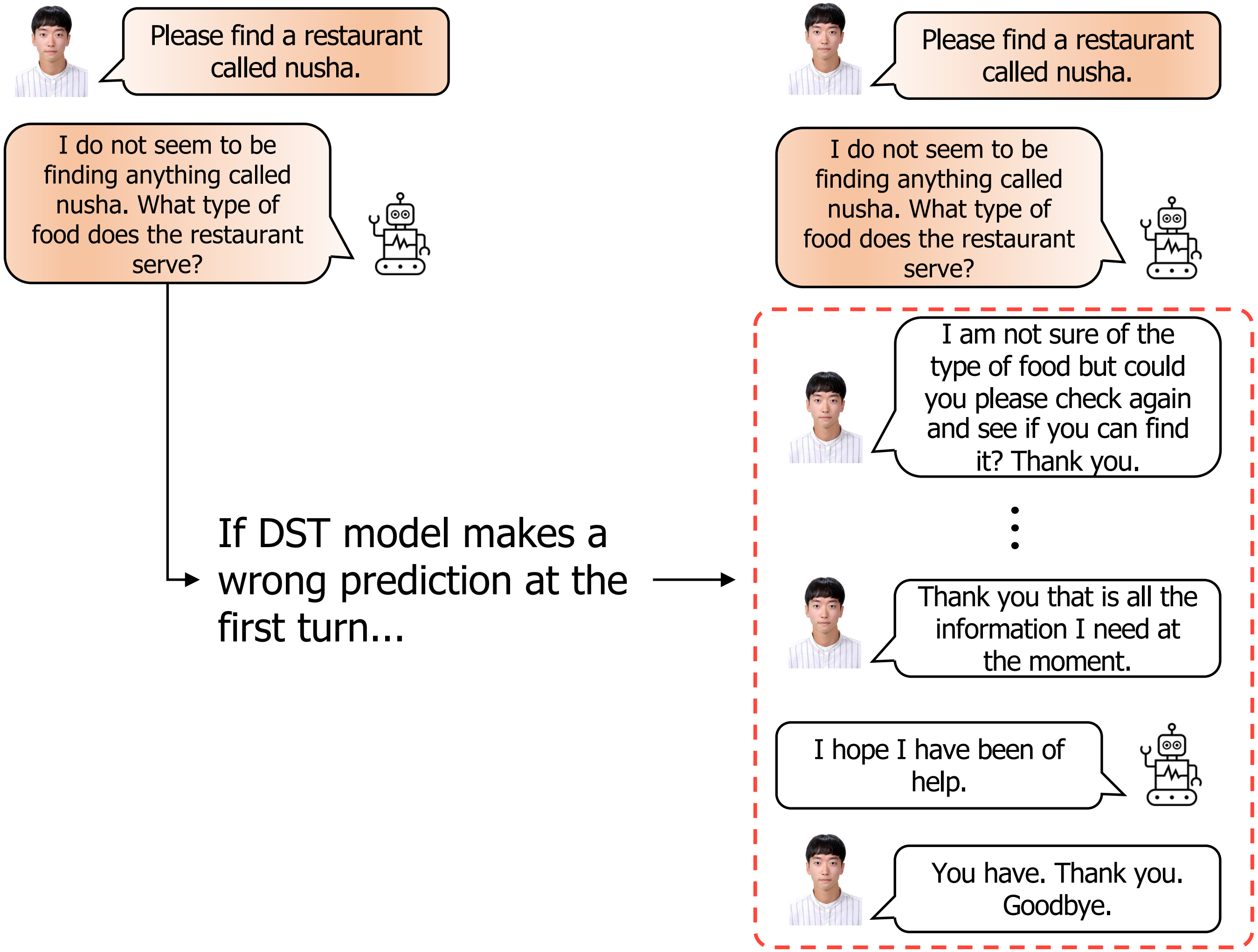
Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

If DST model makes a wrong prediction at the first turn...

# Joint Goal Accuracy (JGA)

$$JGA = \begin{cases} 1 & \text{if predicted state} = \text{gold state} \\ 0 & \text{otherwise} \end{cases}$$



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

If DST model makes a wrong prediction at the first turn...

Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

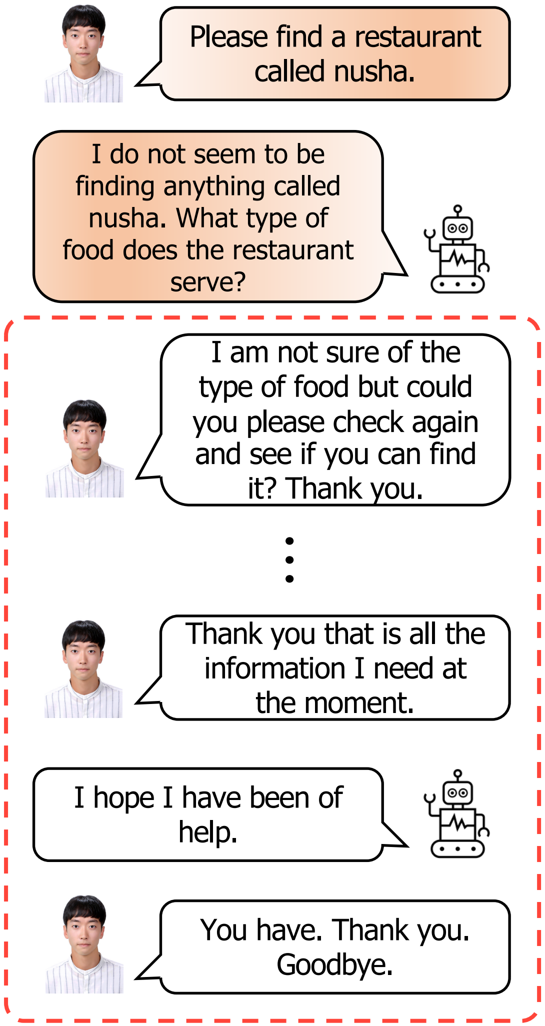
I hope I have been of help.

You have. Thank you. Goodbye.

JGA does not consider subsequent dialogues (JGA = 0)

# Joint Goal Accuracy (JGA)

- Error propagates through later turns
- JGA is too strict to evaluate various dialogue situations



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

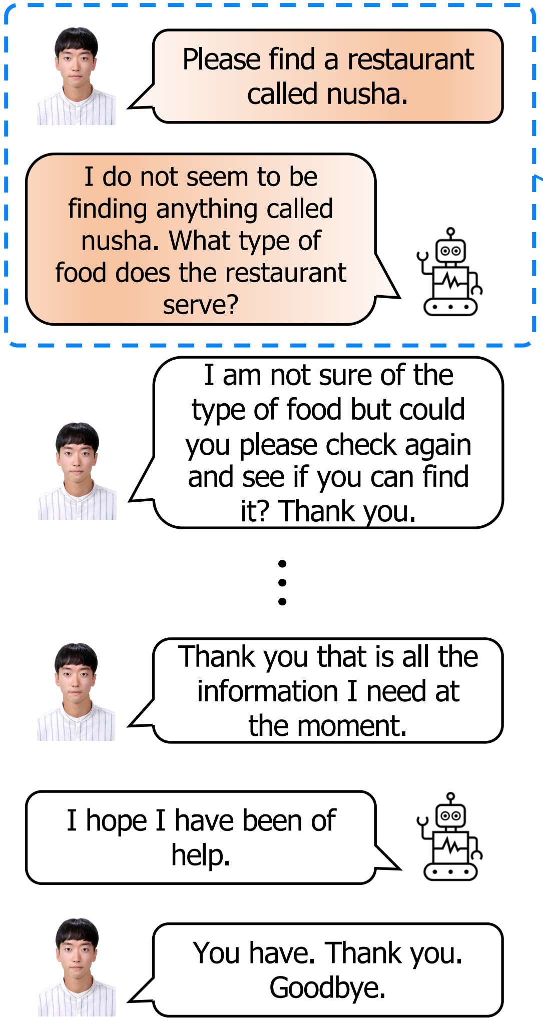
I hope I have been of help.

You have. Thank you. Goodbye.

| Turn | Pred State (domain-slot-value) | Gold State (domain-slot-value) | JGA |
|---|---|---|---|
| 0 | restaurant-name-nusha | - | 0 |
| 1 | restaurant-name-nusha | - | 0 |
| 2 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 3 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 4 | restaurant-area-centre<br>restaurant-food-indian | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian | 0 |
| 5 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | 0 |
| 6 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | 0 |

# Joint Goal Accuracy (JGA)

- Most wrong predictions of model happen in the beginning of the dialogue



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

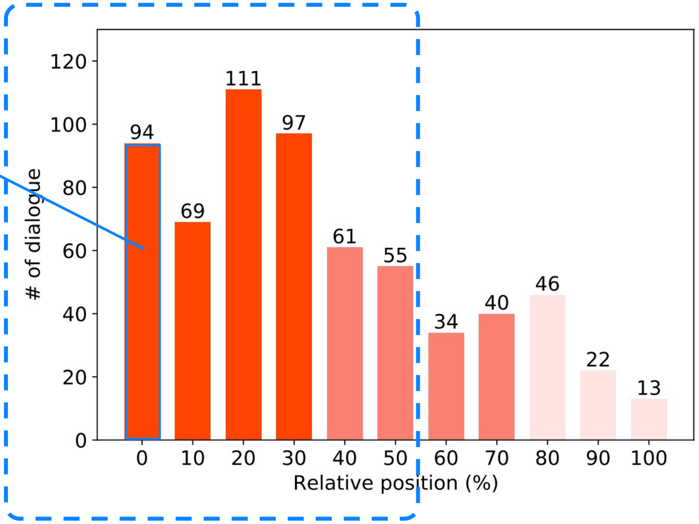
I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.



✓ Most dialogues in MultiWOZ suffer from this problem

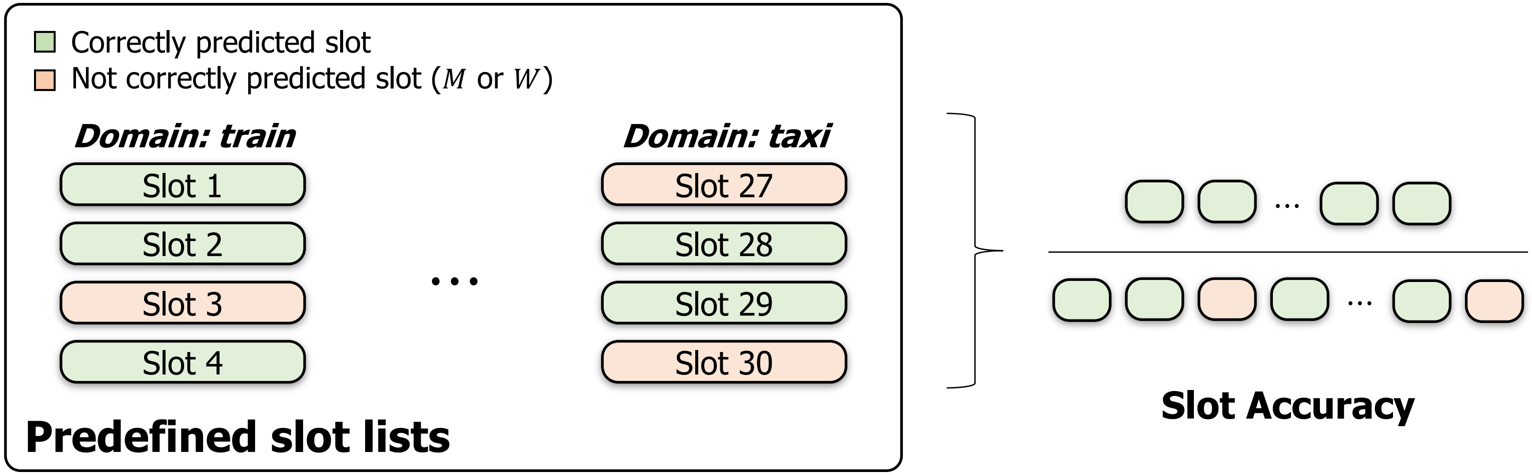
✓ Cannot evaluate the overall flow of dialogue situation

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi* )
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots

  (not existing in gold states)



🟩 Correctly predicted slot
🟧 Not correctly predicted slot ($M$ or $W$)

***Domain: train***

Slot 1
Slot 2
Slot 3
Slot 4

…

***Domain: taxi***

Slot 27
Slot 28
Slot 29
Slot 30

**Predefined slot lists**

**Slot Accuracy**

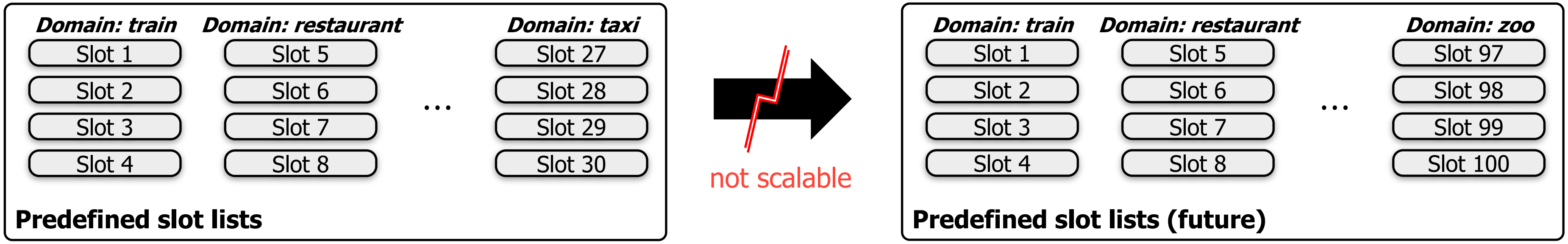
Measures the <u>proportion of correct slots</u> over total predefined slots

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi* )
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots

  (not existing in gold states)



| ***Domain: train*** | ***Domain: restaurant*** | | ***Domain: taxi*** |
|---|---|---|---|
| Slot 1 | Slot 5 | | Slot 27 |
| Slot 2 | Slot 6 | … | Slot 28 |
| Slot 3 | Slot 7 | | Slot 29 |
| Slot 4 | Slot 8 | | Slot 30 |

**Predefined slot lists**

not scalable

| ***Domain: train*** | ***Domain: restaurant*** | | ***Domain: zoo*** |
|---|---|---|---|
| Slot 1 | Slot 5 | | Slot 97 |
| Slot 2 | Slot 6 | … | Slot 98 |
| Slot 3 | Slot 7 | | Slot 99 |
| Slot 4 | Slot 8 | | Slot 100 |

**Predefined slot lists (future)**

$$\frac{30 - 1 - 2}{30} = 0.9333$$

$$\frac{100 - 1 - 2}{100} = 0.98$$

Highly depends on the total number of predefined slots

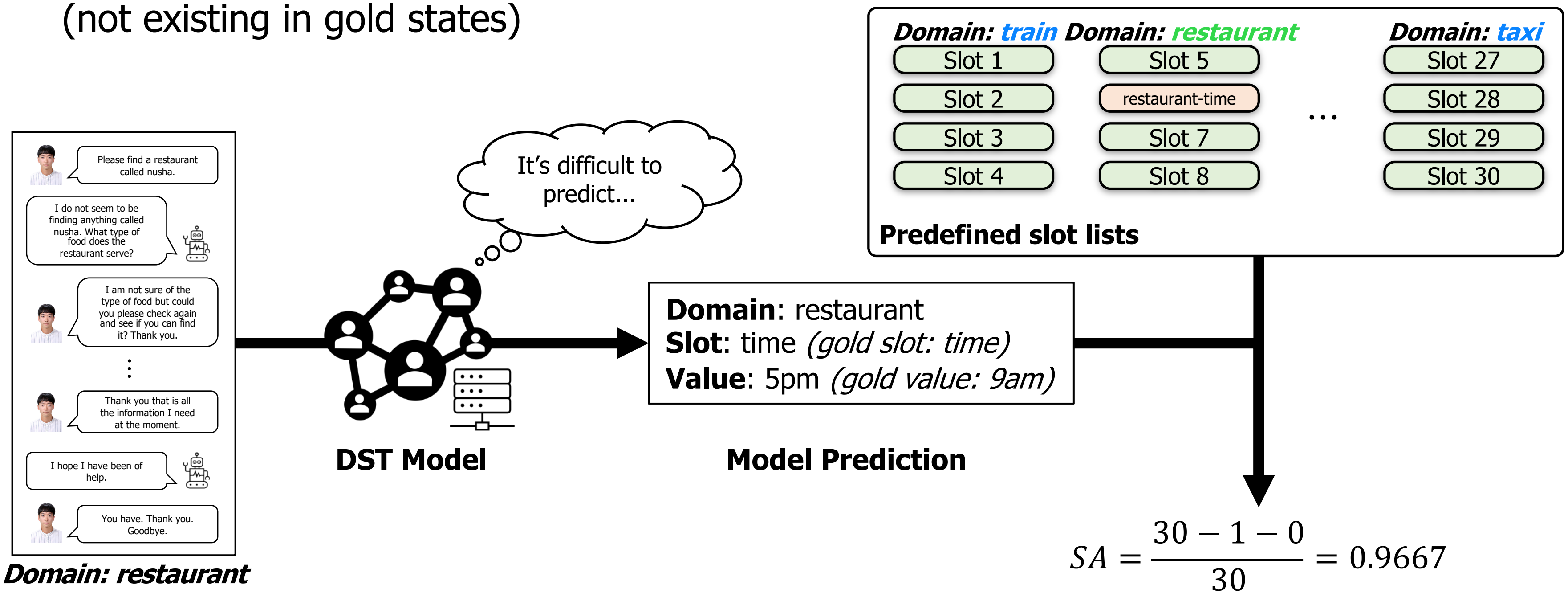
(Performance deviation among models decreases when tasks are added)

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi*)
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots

  (not existing in gold states)



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

***Domain: restaurant***

It's difficult to predict...

**DST Model**

**Domain**: restaurant
**Slot**: time *(gold slot: time)*
**Value**: 5pm *(gold value: 9am)*

**Model Prediction**

***Domain: train*** Slot 1, Slot 2, Slot 3, Slot 4
***Domain: restaurant*** Slot 5, restaurant-time, Slot 7, Slot 8
…
***Domain: taxi*** Slot 27, Slot 28, Slot 29, Slot 30

**Predefined slot lists**

$$SA = \frac{30 - 1 - 0}{30} = 0.9667$$

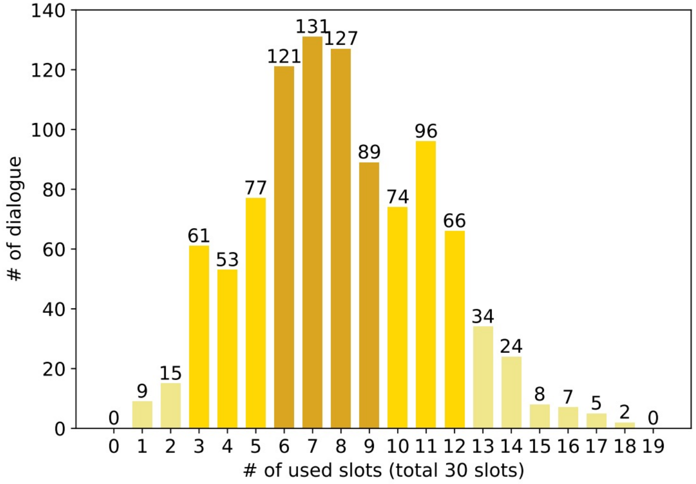
Unrelated situations can affect model performance *(default status: correct)*

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi*)
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots

  (not existing in gold states)



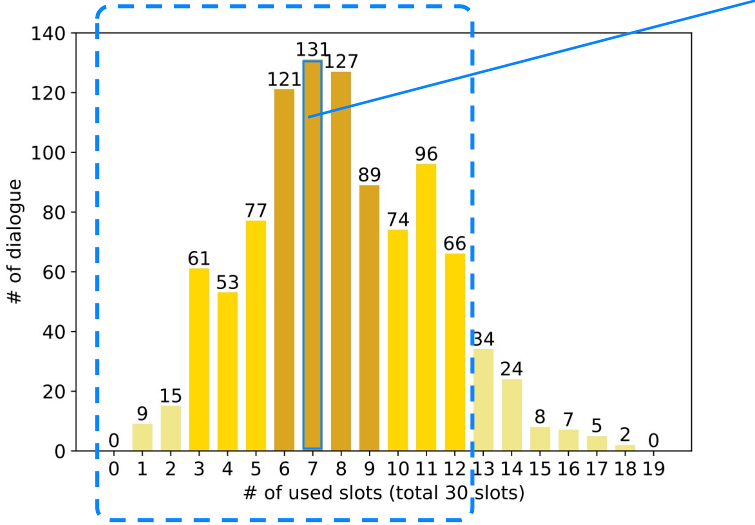
The "maximum" number of slots that appear in a single dialogue

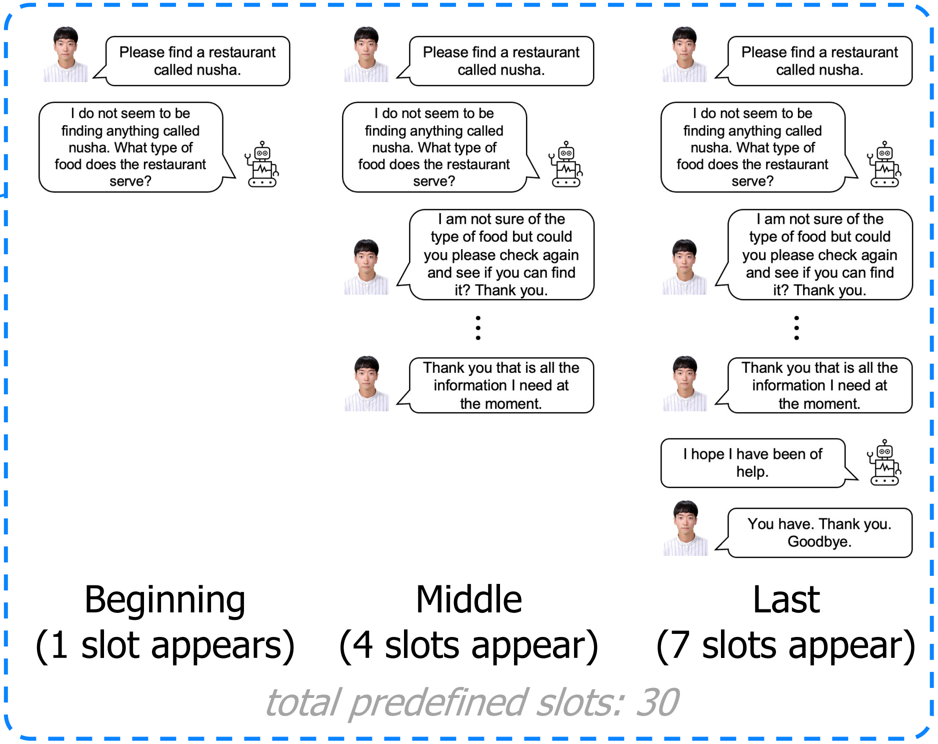
: Most dialogues in MultiWOZ dataset do not have enough slots in each

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi* )
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots (not existing in gold states)





Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

Beginning (1 slot appears)

Middle (4 slots appear)

Last (7 slots appear)

*total predefined slots: 30*

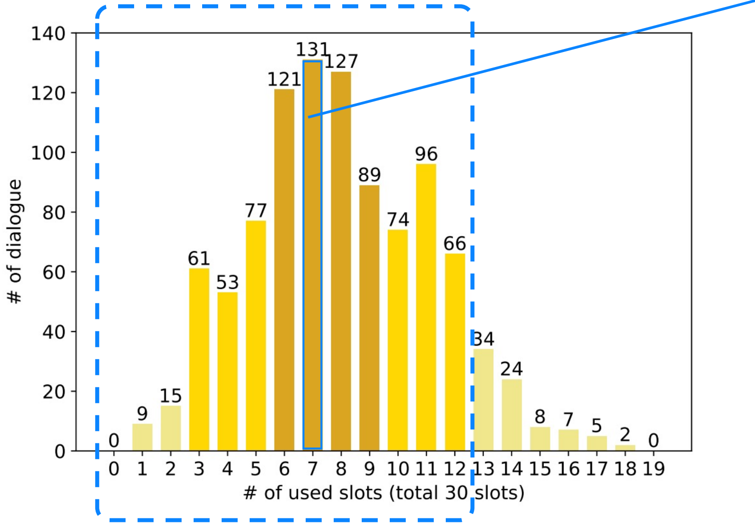
The "maximum" number of slots that appear in a single dialogue

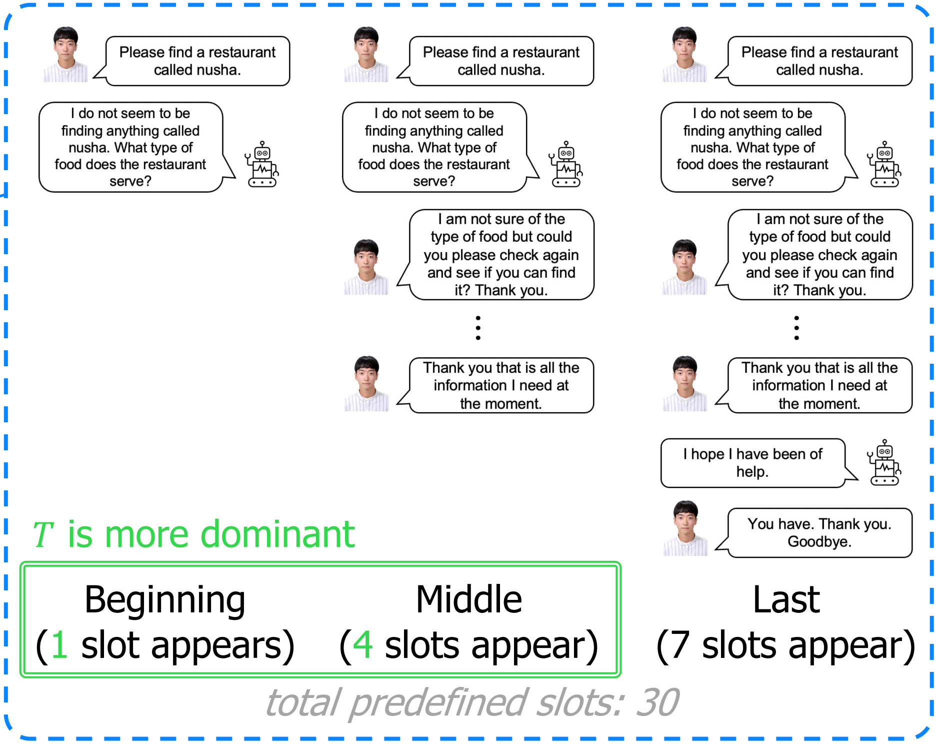
: Most dialogues in MultiWOZ dataset do not have enough slots in each

# Slot Accuracy (SA)

$$SA = \frac{T - M - W}{T}$$

- $T$: Total number of predefined slots (*30* in *train*, *hotel*, *restaurant*, *attraction*, and *taxi* )
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots (not existing in gold states)





Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

$T$ is more dominant

Beginning (1 slot appears)

Middle (4 slots appear)

Last (7 slots appear)

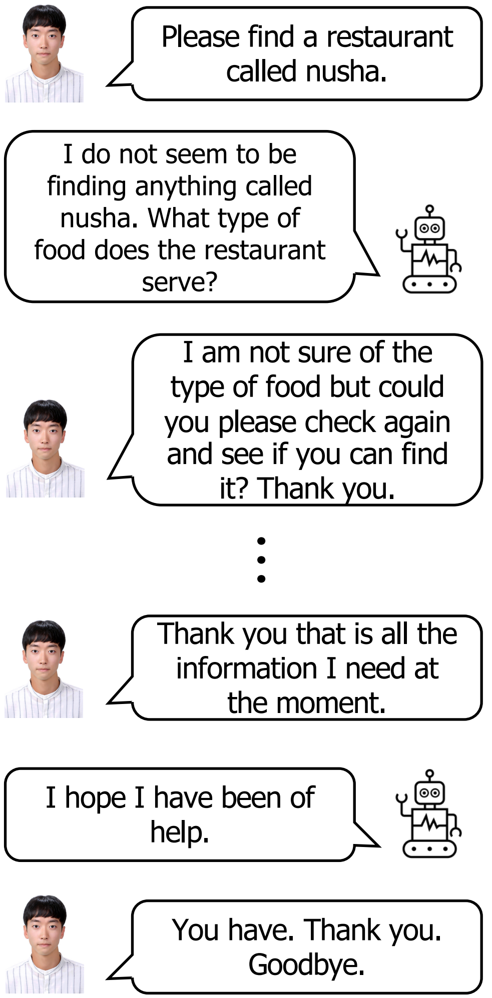
*total predefined slots: 30*

The "maximum" number of slots that appear in a single dialogue

: Most dialogues in MultiWOZ dataset do not have enough slots in each

# Slot Accuracy (SA)

- SA excessively depends on predefined slots not appearing in current dialogue
- Show unnecessarily high score → not aligned with human intuition



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

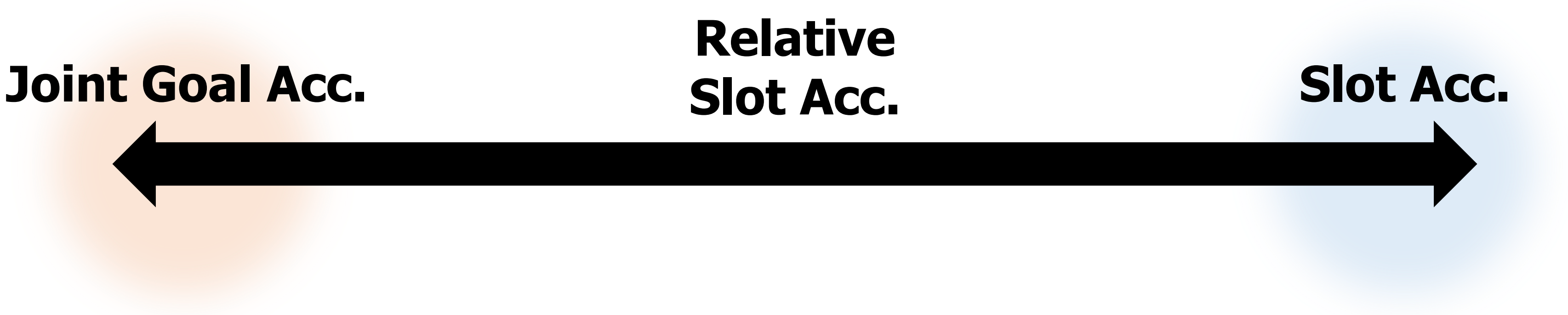
You have. Thank you. Goodbye.

| Turn | Pred State (domain-slot-value) | Gold State (domain-slot-value) | SA |
|---|---|---|---|
| 0 | restaurant-name-nusha | - | 0.9667 |
| 1 | restaurant-name-nusha | - | 0.9667 |
| 2 | restaurant-name-nusha | attraction-name-nusha | 0.9333 |
| 3 | restaurant-name-nusha | attraction-name-nusha | 0.9333 |
| 4 | restaurant-area-centre<br>restaurant-food-indian | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian | 0.9667 |
| 5 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | 0.9667 |
| 6 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | 0.9667 |

# Proposed: Relative Slot Accuracy (RSA)

$$RSA = \frac{T^* - M - W}{T^*}, \qquad \text{where 0 if } T^* = 0$$

- $T^*$: Number of unique slots appearing in the predicted and gold states
- $M$: Number of mispredicted slots (existing in gold states)
- $W$: Number of wrongly predicted slots (not existing in gold states)



**Joint Goal Acc.**

**Relative Slot Acc.**

**Slot Acc.**

Simple but effectively complement limitations of existing metrics

# Proposed: Relative Slot Accuracy (RSA)

$$RSA = \frac{T^* - M - W}{T^*}, \qquad \text{where } 0 \text{ if } T^* = 0$$

Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

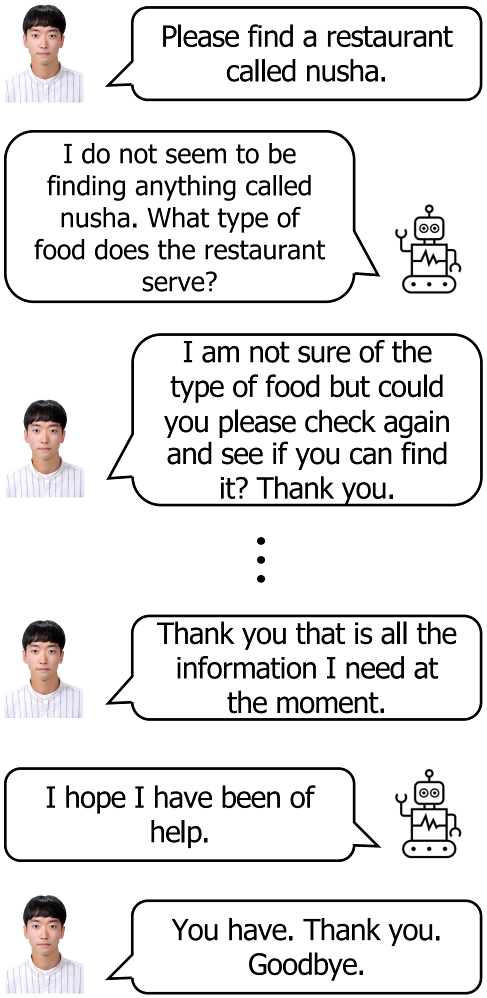
You have. Thank you. Goodbye.

Model does not predict any slots → Penalize

| Turn | Pred State (domain-slot-value) | Gold State (domain-slot-value) | RSA |
|---|---|---|---|
| 0 | restaurant-name-nusha | - | 0 |
| 1 | restaurant-name-nusha | - | 0 |
| 2 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 3 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 4 | restaurant-area-centre<br>restaurant-food-indian | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian | 0.6667 |
| 5 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | 0.7500 |
| 6 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | 0.8000 |

# Proposed: Relative Slot Accuracy (RSA)

$$RSA = \frac{T^* - M - W}{T^*}, \qquad \text{where } 0 \text{ if } T^* = 0$$



Please find a restaurant called nusha.

I do not seem to be finding anything called nusha. What type of food does the restaurant serve?

I am not sure of the type of food but could you please check again and see if you can find it? Thank you.

⋮

Thank you that is all the information I need at the moment.

I hope I have been of help.

You have. Thank you. Goodbye.

Model does predict slots incrementally → Reward

| Turn | Pred State (domain-slot-value) | Gold State (domain-slot-value) | RSA |
|---|---|---|---|
| 0 | restaurant-name-nusha | - | 0 |
| 1 | restaurant-name-nusha | - | 0 |
| 2 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 3 | restaurant-name-nusha | attraction-name-nusha | 0 |
| 4 | restaurant-area-centre<br>restaurant-food-indian | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian | 0.6667 |
| 5 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive | 0.7500 |
| 6 | restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | attraction-name-nusha<br>restaurant-area-centre<br>restaurant-food-indian<br>restaurant-pricerange-expensive<br>restaurant-name-saffron brasserie | 0.8000 |

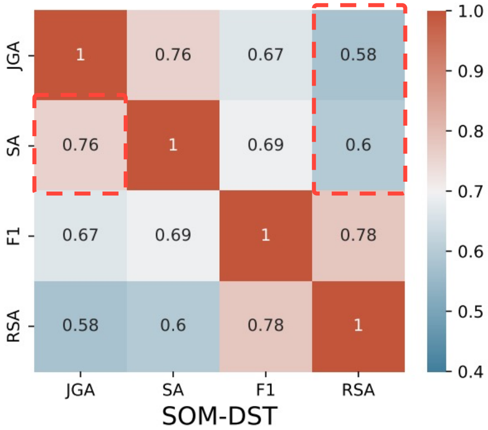
# Proposed: Relative Slot Accuracy (RSA)

- Performance comparison of various DST models
- Comparison can be made differently with relative slot accuracy

| Type | Model | Joint Goal Acc. | Slot Acc. | F1 Score | Relative Slot Acc. |
|---|---|---|---|---|---|
| Open Vocabulary | Transformer-DST (Zeng and Nie, 2021) | 0.5446 | 0.9748 | 0.9229 | 0.8759 |
| | TripPy (Heck et al., 2020) | 0.6131 | 0.9707 | 0.8573 | 0.8432 |
| | SOM-DST (Kim et al., 2020) | 0.5242 | 0.9735 | 0.9179 | 0.8695 |
| | Simple-TOD (Hosseini-Asl et al., 2020) | 0.5605 | 0.9761 | 0.9276 | 0.8797 |
| | SAVN (Wang et al., 2020) | 0.5357 | 0.9749 | 0.9246 | 0.8769 |
| | TRADE (Wu et al., 2019) | 0.4939 | 0.9700 | 0.9033 | 0.8520 |
| | COMER (Ren et al., 2019) | 0.4879 | 0.9652 | 0.8800 | 0.8250 |
| Ontology based | DST-STAR (Ye et al., 2021) | 0.5483 | 0.9754 | 0.9253 | 0.8780 |
| | L4P4K2-DSGraph (Lin et al., 2021) | 0.5178 | 0.9690 | 0.9189 | 0.8570 |
| | SUMBT (Lee et al., 2019) | 0.4699 | 0.9666 | 0.8934 | 0.8380 |

# Proposed: Relative Slot Accuracy (RSA)

- RSA is less correlated with other accuracy metrics
- Different results when considering accumulated multi-turn situation

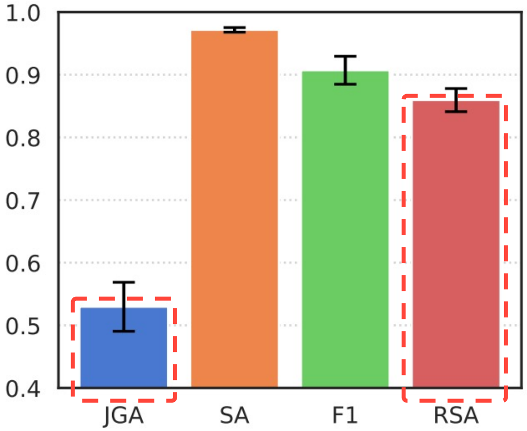


✓ We can expect high SA when JGA is high

✓ We cannot expect high RSA when JGA and SA are high

# Proposed: Relative Slot Accuracy (RSA)

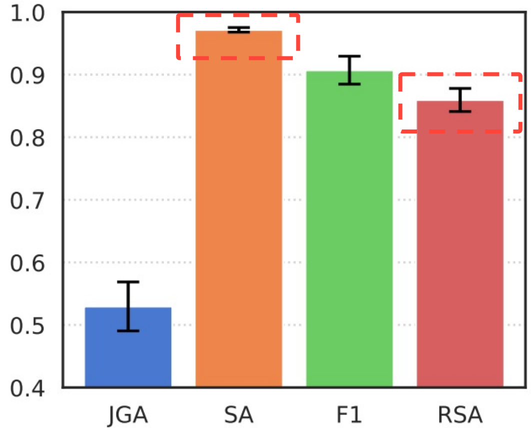
- RSA is less strict than JGA: evaluation with a flexible manner
- RSA has larger deviation than SA: more detailed performance comparison



Enables realistic comparison among models

# Proposed: Relative Slot Accuracy (RSA)

- RSA is less strict than JGA: evaluation with a flexible manner
- RSA has larger deviation than SA: more detailed performance comparison



Enables realistic comparison among models

# Summary

- Pointed out the limitation of current evaluation metrics in DST
  - ✓ Joint goal accuracy (JGA) underestimates dialogue situations
  - ✓ Slot accuracy (SA) overestimates dialogue situations

- Propose **relative slot accuracy (RSA)** to complement these metrics
  - ✓ Does not depend on the number of predefined slots
  - ✓ Aligned with human intuition by rewarding and penalizing according to the model prediction in accumulated multi-turn structure

# 🤗 Thank you for listening 🤗

Contact: takyoung_kim@korea.ac.kr

ArXiv: https://arxiv.org/abs/2203.03123