

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

Takyoung Kim<sup>1</sup>, Hoonsang Yoon<sup>1</sup>, Yukyung Lee<sup>1</sup>, Pilsung Kang<sup>1</sup>, Misuk Kim<sup>2</sup>

<sup>1</sup>Korea University, Seoul, Republic of Korea

<sup>2</sup>Sejong University, Seoul, Republic of Korea

<sup>1</sup>{takyoung\_kim, hoonsang\_yoon, yukyung\_lee, pilsung\_kang}@korea.ac.kr

<sup>2</sup>misuk.kim@sejong.ac.kr

## Abstract

Dialogue state tracking (DST) aims to extract essential information from multi-turn dialogue situations and take appropriate actions. A belief state, one of the core pieces of information, refers to the subject and its specific content, and appears in the form of `domain-slot-value`. The trained model predicts “accumulated” belief states in every turn, and joint goal accuracy and slot accuracy are mainly used to evaluate the prediction; however, we specify that the current evaluation metrics have a critical limitation when evaluating belief states accumulated as the dialogue proceeds, especially in the most used MultiWOZ dataset. Additionally, we propose **relative slot accuracy** to complement existing metrics. Relative slot accuracy does not depend on the number of predefined slots, and allows intuitive evaluation by assigning relative scores according to the turn of each dialogue. This study also encourages not solely the reporting of joint goal accuracy, but also various complementary metrics in DST tasks for the sake of a realistic evaluation.

## 1 Introduction

The dialogue state tracking (DST) module structures the belief state that appears during the conversation in the form of `domain-slot-value`, to provide an appropriate response to the user. Recently, multi-turn DST datasets have been constructed using the Wizard-of-Oz method to reflect more realistic dialogue situations (Wen et al., 2017; Mrkšić et al., 2017; Budzianowski et al., 2018). The characteristic of these datasets is that belief states are “accumulated” and recorded every turn. That is, the belief states of the previous turns are included in the current turn. It confirms whether the DST model tracks essential information that has appeared up to the present point.

Joint goal accuracy and slot accuracy are utilized in most cases to evaluate the prediction of accumu-

lated belief states. Joint goal accuracy strictly determines whether every predicted state is identical to the gold state, whereas slot accuracy measures the ratio of correct predictions. However, we determined that these two metrics solely focus on “penalizing states that fail to predict,” not considering “reward for well-predicted states.” Accordingly, as also pointed out in Rastogi et al. (2020a), joint goal accuracy underestimates the model prediction because of its error accumulation attribute, while slot accuracy overestimates it because of its dependency on predefined slots.

However, there is a lack of discussion on the metric for evaluating the most used MultiWOZ dataset, despite a recently published dataset (Rastogi et al., 2020b) proposing some metrics. To address the above challenge, we propose reporting the **relative slot accuracy** along with the existing metrics in MultiWOZ dataset. While slot accuracy has the challenge of overestimation by always considering all predefined slots in every turn, relative slot accuracy does not depend on predefined slots, and calculates a score that is affected solely by slots that appear in the current dialogue. Therefore, relative slot accuracy enables a realistic evaluation by rewarding the model’s correct predictions, a complementary approach that joint goal and slot accuracies cannot fully cover. It is expected that the proposed metric can be adopted to evaluate model performance more intuitively.

## 2 Current Evaluation Metrics

### 2.1 Joint Goal Accuracy

Joint goal accuracy, developed from Henderson et al. (2014b) and Zhong et al. (2018), can be said to be an ideal metric, in that it verifies that the predicted belief states perfectly match the gold label. Equation 1 expresses how to calculate the joint goal accuracy, depending on whether the slot values match each turn.

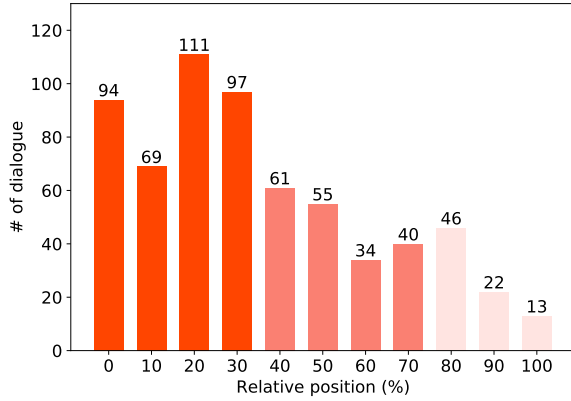


Figure 1: The relative position where joint goal accuracy of the turn is measured to be zero for the first time among the dialogues where joint goal accuracy of the last turn is zero. (642 of 999 MultiWOZ 2.1 test set with SOM-DST).

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

However, the joint goal accuracy underestimates the accumulated states because it scores the performances of later turn to zero if the model mispredicts even once in a particular turn, regardless of the model prediction quality at later turns. As illustrated in Figure 1, we measured the relative position of the turn causing this phenomenon for the dialogue. We used MultiWOZ 2.1 (Eric et al., 2019), and analyzed 642 samples from a total of 999 test sets in which the joint goal accuracy of the last turn is zero. The DST model selected for primary verification is the SOM-DST (Kim et al., 2020), which is one of the latest DST models. Accordingly, the relative position where joint goal accuracy first became zero was mainly at the beginning of the dialogue<sup>1</sup>. This means that the joint goal accuracy after the beginning of the dialogue is unconditionally measured as zero because of the initial misprediction, although the model may correctly predict new belief states at later turns. Failure to measure the performance of the latter part means that it cannot consider various dialogue situations provided in the dataset, which is a critical issue in building a realistic DST model.

<sup>1</sup>59 samples of the 642 samples have a joint goal accuracy of 1 in the middle, owing to a coincidental situation or differences in the analysis of annotation. Table A1 and Table A2 show the dialogue situation in detail, and Table A3 and Table A4 show the belief states accordingly. Refer to Appendix A.

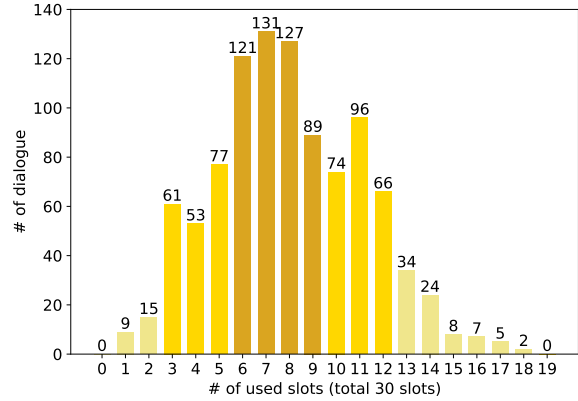


Figure 2: The number of predefined gold slots used in each dialogue (999 MultiWOZ 2.1 test set).

## 2.2 Slot Accuracy

Slot accuracy can compensate for situations where joint goal accuracy does not fully evaluate the dialogue situation. Equation 2 expresses how to calculate the slot accuracy.  $T$  indicates the total number of predefined slots for all the domains.  $M$  denotes the number of missed slots that the model does not accurately predict among the slots included in the gold state, and  $W$  denotes the number of wrongly predicted slots among the slots that do not exist in the gold state.

$$SA = \frac{T - M - W}{T} \quad (2)$$

Figure 2 illustrates the total number of annotated slots in MultiWOZ 2.1 to figure out the limitation of slot accuracy. Each value of  $x$ -axis in Figure 2 indicates the “maximum” number of slots that appear in a single dialogue, and we confirmed that approximately 85% of the test set utilized solely less than 12 of the 30 predefined slots in the experiment. Because the number of belief states appearing in the early and middle turns of the dialogue are smaller, and even fewer states make false predictions, calculating slot accuracy using Equation 2 reduces the influence of  $M$  and  $W$ , and the final score is dominated by the total slot number  $T$ . Accordingly, several previous studies still report the model performance using solely joint goal accuracy because slot accuracy excessively depends on the number of predefined slots, making the performance deviation among models trivial (refer to Table A5).

Furthermore, according to Table A6, we determined that slot accuracy tends to be too high. The slot accuracies of turns 0 and 1 show approximately 96% accuracy, despite the model not cor-

Type	Model	Joint Goal Acc.	Slot Acc.	F1 Score	Relative Slot Acc.
Open vocabulary	Transformer-DST (2021)	0.5446	0.9748	0.9229	0.8759
	TripPy (2020)	<b>0.6131</b>	0.9707	0.8573	0.8432
	SOM-DST (2020)	0.5242	0.9735	0.9179	0.8695
	Simple-TOD (2020)	0.5605	<b>0.9761</b>	<b>0.9276</b>	<b>0.8797</b>
	SAVN (2020)	0.5357	0.9749	0.9246	0.8769
	TRADE (2019)	0.4939	0.9700	0.9033	0.8520
	COMER (2019)	0.4879	0.9652	0.8800	0.8250
Ontology based	DST-STAR (2021)	0.5483	0.9754	0.9253	0.8780
	L4P4K2-DSGraph (2021)	0.5178	0.9690	0.9189	0.8570
	SUMBT (2019)	0.4699	0.9666	0.8934	0.8380

Table 1: Model performance of MultiWOZ 2.1 with various evaluation metrics. All reported performances are our re-implementation.

rectly predicting states at all. It becomes difficult to compare various models in detail, if each model shows a high performance, even though nothing is adequately predicted. In addition, as the turn progresses, there are no rewards for a situation in which the model tracks the belief state without any challenges. The case correctly predicting two out of three in turn 4, and the case correctly predicting three out of four in turn 5 exhibit the same slot accuracy. Therefore, the slot accuracy measured according to Equation 2 differs from our intuition.

### 2.3 Other Metric

Recently, Rastogi et al. (2020b) proposed a metric called average goal accuracy. The main difference between the average goal accuracy and the proposed relative slot accuracy is that the average goal accuracy only considers the slots with non-empty values in the gold states of each turn, whereas the proposed relative slot accuracy considers those in both gold and predicted states. Since average goal accuracy ignores the predicted states, it cannot properly distinguish a better model from a worse model in some specific situations. We will discuss it in more detail in Section 4.1.

## 3 Relative Slot Accuracy

As can be observed in Equation 2, slot accuracy has the characteristic that the larger the number of predefined slots ( $T$ ), the smaller the deviation between the prediction results. The deviation among DST models will be even more minor when constructing datasets with various dialogue situations, because the number of predefined slots will continually in-

crease. It is not presumed to be an appropriate metric in terms of scalability.

Therefore, we propose relative slot accuracy, that is not affected by predefined slots, and is evaluated with adequate rewards and penalties that fit human intuition in every turn. Equation 3 expresses how to calculate the relative slot accuracy, and  $T^*$  denotes the number of unique slots appearing in the predicted and gold states in a particular turn.

$$RSA = \frac{T^* - M - W}{T^*}, \text{ where } 0 \text{ if } T^* = 0 \quad (3)$$

Relative slot accuracy rewards well-predicted belief states by measuring the scores in accumulating turns. Further discussions on the relative score will be discussed in Section 4.1.

## 4 Experiments

We measured MultiWOZ 2.1, an improved version of MultiWOZ 2.0 (Budzianowski et al., 2018), which has been adopted in several studies, according to Table A5. Five domains (i.e., *hotel*, *train*, *restaurant*, *attraction*, and *taxi*) are adopted in the experiment, following Wu et al. (2019), and there are a total of 30 `domain-slot` pairs. We selected the DST models in Table A5 that perform the MultiWOZ experiment with the original authors’ reproducible code<sup>2</sup>. Additionally, we reported the F1 score, which can be calculated using the current predicted and gold states.

<sup>2</sup>Implementation codes for Simple-TOD and TripPy are from <https://github.com/salesforce/coco-dst>.

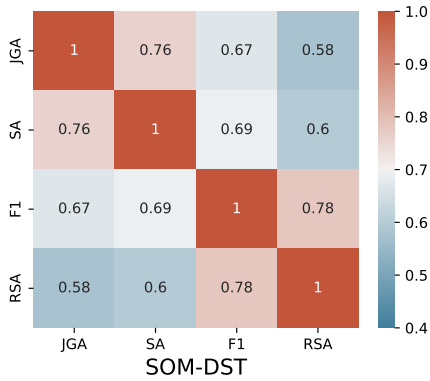


Figure 3: Correlation matrix of evaluation performance of total 7,368 turns in 999 MultiWOZ 2.1 test set using SOM-DST. Results for other models are included in Figure A1.

#### 4.1 Results and Discussion

Table 1 presents the overall results. Regarding slot accuracy, the difference between the largest and smallest values is solely 1.09%. It can be one of the reasons that several researchers do not report it. Meanwhile, relative slot accuracy can explicitly highlight the deviation among models by showing a 5.47% difference between the largest and smallest values. Furthermore, the correlation with joint goal accuracy, a mainly adopted metric, and relative slot accuracy with respect to each turn is lower than the correlation with joint goal accuracy and slot accuracy, as illustrated in Figure 3. Specifically, it can be compared with a different perspective when using the proposed reward-considering evaluation metric.

**Domain-specific Evaluation** We reported the joint goal, slot, and relative slot accuracies per domain utilizing the SOM-DST model in Table 2. Relative slot accuracy derives a specific score in the turn configuration and prediction ratio of each domain by excluding slots that do not appear in the conversation. For example, the *taxi* domain shows a low score, meaning that it has relatively several cases of incorrect predictions, compared to the number of times slots belonging to the *taxi* domain appear. Because slot accuracy cannot distinguish the above trend, the score of the *hotel* domain is lower than that of the *taxi* domain. In summary, relative slot accuracy enables relative comparison according to the distribution of the domain in a dialogue.

Domain	Joint Goal Acc.	Slot Acc.	Relative Slot Acc.
hotel	0.4923	0.9731	0.8493
train	0.7162	0.9874	0.9176
restaurant	0.6589	0.9858	0.8977
attraction	0.6811	0.9878	0.8421
taxi	0.5701	0.9798	0.7828

Table 2: Per-domain performance of SOM-DST prediction.

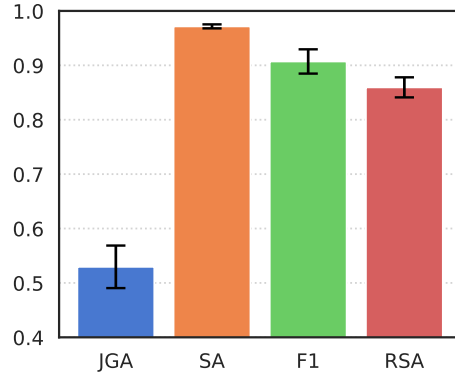


Figure 4: The mean and standard deviation of model performance reported in Table 1.

**Dependency on Predefined Slots** As discussed in Section 2.2, slot accuracy requiring total predefined slots is not a scalable method for evaluating the current dialogue dataset that contains a few domains in each dialogue. For example, when evaluating a dialogue sample that solely deals with the *restaurant* domain, even domains that never appear at all (i.e., *hotel*, *train*, *attraction*, and *taxi*) are involved in measuring performance, making deviations among different models trivial. However, relative slot accuracy can evaluate the model’s predictive score without being affected by slots never seen in the current dialogue, which is a more realistic way, considering that each dialogue contains its own turn and slot composition. Figure 4 illustrates the mean and standard deviations of the model performance in Table 1. As can be observed from the results, the relative slot accuracy has a higher deviation than the slot accuracy, enabling a detailed comparison among the methodologies.

**Reward on Relative Dialogue Turn** Relative slot accuracy is able to reward the model’s correct prediction by measuring the accuracy on a relative basis for each turn. Table A6 compares the slot and relative slot accuracies. The relative slot accuracy from turns 0 – 3 is measured as 0 because it cal-

Type	Belief State	Joint Goal Acc.	Average Goal Acc.	Relative Slot Acc.
Gold State	restaurant-area-centre restaurant-food-indian restaurant-people-2	-	-	-
Prediction of Model A	restaurant-area-centre restaurant-food-chinese attraction-area-centre	0	0.3333	0.2500
Prediction of Model B	restaurant-area-centre restaurant-food-chinese restaurant-name-nusha attraction-area-centre attraction-pricerange-cheap	0	0.3333	0.1667

Table 3: A situation that average goal accuracy cannot distinguish between two models. States with blue denote correct prediction, and as defined in Section 2.2, states with orange and pink denote respective  $M$  and  $W$ .

culates the score based on the unique state of the current turn according to Equation 3. In addition, regarding slot accuracy in turns 4, 5, and 6, there is no score improvement for the additional well-predicted state by the model, whereas the score increases when the newly added state is matched in the case of relative slot accuracy. Therefore, relative slot accuracy can provide an intuitive evaluation reflecting the current belief state recording method, in which the number of slots accumulates incrementally as the conversation progresses.

**Comparison to Average Goal Accuracy** Relative slot accuracy can compare DST model performances more properly than average goal accuracy, as mentioned in Section 2.3. Table 3 describes how these two metrics result in different values for the same model predictions. In this example, average goal accuracy cannot consider additional belief states incorrectly predicted by Model B, resulting in the same score between the two models. In contrast, relative slot accuracy can give a penalty proportional to the number of wrong predictions because it includes both gold and predicted states when calculating the score. Consequently, relative slot accuracy has a more elaborated discriminative power than the average goal accuracy.

## 5 Conclusion

This paper points out the challenge that the existing joint goal and slot accuracies cannot fully evaluate the accumulating belief state of each turn in the MultiWOZ dataset. Accordingly, the relative slot accuracy is proposed. This metric is not affected by unseen slots in the current dialogue situation, and compensates for the model’s correct predic-

tion. When the DST task is scaled up to deal with more diverse conversational situations, a realistic model evaluation will be possible using relative slot accuracy. Moreover, we suggest reporting various evaluation metrics to complement the limitations of each metric in future studies, not solely reporting the joint goal accuracy.

## Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00034, Clustering technologies of fragmented data for time-based data analysis) and Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2020040126-0001)

## References

- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Guan-Lin Chao and Ian R. Lane. 2019. [BERT-DST: scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer](#). In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1468–1472. ISCA.
- Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. [Schema-guided multi-domain](#)



- dialogue state tracking with graph attention neural networks. In *AAAI*.
- Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. **Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines**. *CoRR*, abs/1907.01669.
- Yue Feng, Yang Wang, and Hang Li. 2021. **A sequence-to-sequence approach to dialogue state tracking**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1714–1725, Online. Association for Computational Linguistics.
- Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. **Dialog state tracking: A neural reading comprehension approach**. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.
- Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. **Hyst: A hybrid approach for flexible and accurate dialogue state tracking**. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1458–1462. ISCA.
- Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geischauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. **TripPy: A triple copy strategy for value independent neural dialog state tracking**. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.
- Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.
- Matthew Henderson, Blaise Thomson, and Steve Young. 2014b. **Word-based dialog state tracking with recurrent neural networks**. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, U.S.A. Association for Computational Linguistics.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. **A Simple Language Model for Task-Oriented Dialogue**. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.
- Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sangwoo Lee. 2020. **Efficient dialogue state tracking by selectively overwriting memory**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.
- Hung Le, Richard Socher, and Steven C.H. Hoi. 2020. **Non-autoregressive dialog state tracking**. In *International Conference on Learning Representations*.
- Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. **SUMBT: Slot-utterance matching for universal and scalable belief tracking**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.
- Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021. **Knowledge-aware graph-enhanced gpt-2 for dialogue state tracking**.
- Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. **Neural belief tracker: Data-driven dialogue state tracking**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. **Schema-guided dialogue state tracking task at dstc8**.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. **Scalable and accurate dialogue state tracking via hierarchical sequence generation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1876–1885, Hong Kong, China. Association for Computational Linguistics.
- Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. **Building a conversational agent overnight with dialogue self-play**.
- Yexiang Wang, Yi Guo, and Siqi Zhu. 2020. **Slot attention with value normalization for multi-domain dialogue state tracking**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3019–3028, Online. Association for Computational Linguistics.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. **A network-based end-to-end trainable task-oriented dialogue system**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational*

*Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. [Transferable multi-domain state generator for task-oriented dialogue systems](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Peng Wu, Bowei Zou, Ridong Jiang, and AiTi Aw. 2020. [GCDST: A graph-based and copy-augmented multi-domain dialogue state tracking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1063–1073, Online. Association for Computational Linguistics.

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021. [Slot self-attentive dialogue state tracking](#). In *Proceedings of the Web Conference 2021, WWW '21*, page 1598–1608, New York, NY, USA. Association for Computing Machinery.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. [MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking base-lines](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Yan Zeng and Jian-Yun Nie. 2021. [Jointly optimizing state operation prediction and value generation for dialogue state tracking](#).

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. [Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking](#). In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. [Global-locally self-attentive encoder for dialogue state tracking](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467, Melbourne, Australia. Association for Computational Linguistics.

Li Zhou and Kevin Small. 2019. [Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering](#). *CoRR*, abs/1911.06192.

Su Zhu, Jieyu Li, Lu Chen, and Kai Yu. 2020. [Efficient context and schema fusion networks for multi-domain dialogue state tracking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 766–781, Online. Association for Computational Linguistics.

## A Complementary discussions of joint goal accuracy

Our findings show that if the model makes an incorrect prediction, the error accumulates until the end of the dialogue, and the joint goal accuracy remains at zero. In this section, we discuss a few cases of 59 dialogues that do not show the trend among 642 dialogues selected in Section 2.1; however, it is important to note that these few cases have negligible effect on the trend in Figure 1, solely changing the position where the joint goal accuracy first becomes zero.

We sampled dialogues of the MultiWOZ 2.1 test set in Table A1 and Table A2, and marked values appearing in the dialogue in bold. Table A3 and Table A4 indicate the corresponding belief states of each dialogue. In the first dialogue presented in Table A1, the joint goal accuracy is measured as 1 at turn 2. In this case, the model incorrectly predicted the `restaurant-pricerange` slot at turns 0 and 1, and then the utterance about the slot appeared by chance. In a general case, the wrong prediction of the `restaurant-pricerange` slot at turn 0 will accumulate to the last turn. However, in this case, another incorrect prediction at turn 3 will cause error accumulation in this dialogue.

The second dialogue presented in Table A2, reports the incorrect prediction according to the interpretation of annotations at turn 4. In other words, because the dialogue about the `hotel-internet` slot appears over turns 4 and 5, it is solely an error depending on the prediction timing of the model. Because the correct belief state was predicted right from turn 5, it cannot be said to be an error accumulation phenomenon; however, the model did not predict the `hotel-pricerange` slot at turn 6, which is the last turn in this case.

In conclusion, it can be determined that the model does not seem to accumulate erroneous predictions because of an accidental situation or interpretation of annotations, but this does not negate the error accumulation phenomenon. Furthermore, the fact that the starting point of making the joint goal accuracy of subsequent turns to 0 mainly occurs at the beginning of the dialogue does not change.

Turn	Dialogue History
0	<i>System:</i> “ ” <i>User:</i> “can you help me find a nice restaurant ?”
1	<i>System:</i> “sure ! what kind of food do you like ?” <i>User:</i> “i was thinking some <b>indian</b> food would be great .”
2	<i>System:</i> “i have 22 indian restaurant -s do you have a preference for area of town ?” <i>User:</i> “no , i <b>do not care</b> where it is . i am thinking i would like an <b>expensive</b> restaurant .”
3	<i>System:</i> “would you like to try <b>curry garden</b> ?” <i>User:</i> “that is fine book me a table for <b>6</b> on <b>sat</b> at <b>17:30</b> . i also need a <b>train for the same</b> and should <b>leave leicester</b> for <b>cambridge</b> ”
4	<i>System:</i> “alright , i have made your requested booking at curry garden , and the reference number is hk9ycl6z . as for the train , what time would you like to depart ?” <i>User:</i> “it does not matter as long as i am there by <b>13:45</b> leaving leicester going to cambridge , i’ll need the reference number too please”
5	<i>System:</i> “i have found tr6210 leaving leicester at 11:09 on saturday and arriving in cambridge at 12:54 . i have booked 6 seats . reference number is lr5i1rzv . anything else i can do ?” <i>User:</i> “that will be all for today . thanks so much for all your help . goodbye”

Table A1: Sample dialogue of MultiWOZ 2.1 test set (PMUL4234.json).

Turn	Dialogue History
0	<i>System:</i> “ ” <i>User:</i> “i would like help finding a train headed to <b>cambridge</b> .”
1	<i>System:</i> “i will be happy to help you find a train . can you tell me where you will be departing from ?” <i>User:</i> “departing from <b>london kings cross</b> on <b>tuesday</b> .”
2	<i>System:</i> “when would you like to leave or arrive by ?” <i>User:</i> “i need to <b>arrive by 18,30</b> .”
3	<i>System:</i> “take train tr1434 , which will arrive at 18:08 . shall i book you for that train ?” <i>User:</i> “can i get the price for a ticket , first ?”
4	<i>System:</i> “sure ! the ticket is 23.60 pounds .” <i>User:</i> “thanks ! i am also looking for a hotel called <b>archway house</b> . can you tell me if they have <b>free wifi</b> ?”
5	<i>System:</i> “they do . would you like to book a room ?” <i>User:</i> “i would first like to know what their price range and hotel type are , thank you .”
6	<i>System:</i> “archway house is a <b>moderate</b> -ly priced guesthouse . would you like their address or perhaps to book a room there ?” <i>User:</i> “thank you , but no . you’ve already helped me with everything i needed today .”

Table A2: Sample dialogue of MultiWOZ 2.1 test set (MUL2270.json).



Turn	Predicted State	Gold State	Joint Goal Acc.
0	restaurant-pricerange- <b>expensive</b>	-	0
1	restaurant-pricerange- <b>expensive</b> restaurant-food-indian	restaurant-food-indian	0
2	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare	1
3	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>sunday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>tuesday</b> train-departure-leicester	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>saturday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>saturday</b> train-departure-leicester train-book people- <b>6</b>	0
4	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>sunday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>tuesday</b> train-departure-leicester train-arriveby-13:45 train-leaveat- <b>dontcare</b>	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>saturday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>saturday</b> train-departure-leicester train-arriveby-13:45 train-book people- <b>6</b>	0
5	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>sunday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>tuesday</b> train-departure-leicester train-arriveby-13:45 train-leaveat- <b>dontcare</b>	restaurant-pricerange-expensive restaurant-food-indian restaurant-area-dontcare restaurant-book day- <b>saturday</b> restaurant-book people-6 restaurant-book time-17:30 restaurant-name-curry garden train-destination-cambridge train-day- <b>saturday</b> train-departure-leicester train-arriveby-13:45 train-book people- <b>6</b>	0

Table A3: SOM-DST prediction of MultiWOZ 2.1 test sample (PMUL4234.json).

Turn	Predicted State	Gold State	Joint Goal Acc.
0	train-destination-cambridge	train-destination-cambridge	1
1	train-destination-cambridge	train-destination-cambridge	1
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
2	train-destination-cambridge	train-destination-cambridge	1
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
	train-arriveby-18:30	train-arriveby-18:30	
3	train-destination-cambridge	train-destination-cambridge	1
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
	train-arriveby-18:30	train-arriveby-18:30	
4	train-destination-cambridge	train-destination-cambridge	0
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
	train-arriveby-18:30	train-arriveby-18:30	
	hotel-name-archway house	hotel-name-archway house hotel-internet-yes	
5	train-destination-cambridge	train-destination-cambridge	1
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
	train-arriveby-18:30	train-arriveby-18:30	
	hotel-name-archway house	hotel-name-archway house	
	hotel-internet-yes	hotel-internet-yes	
6	train-destination-cambridge	train-destination-cambridge	0
	train-day-tuesday train-departure-london kings cross	train-day-tuesday train-departure-london kings cross	
	train-arriveby-18:30	train-arriveby-18:30	
	hotel-name-archway house	hotel-name-archway house	
	hotel-internet-yes	hotel-internet-yes hotel-pricerange-moderate	

Table A4: SOM-DST prediction of MultiWOZ 2.1 test sample (MUL2270.json).

Method	Metric	Dataset
DST-STAR (Ye et al., 2021)	JGA	MultiWOZ 2.0 (Budzianowski et al., 2018), MultiWOZ 2.1 (Eric et al., 2019)
Seq2Seq-DU (Feng et al., 2021)	JGA	SGD (Rastogi et al., 2020b), MultiWOZ 2.1, MultiWOZ 2.2 (Zang et al., 2020)
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 (Wen et al., 2017), MultiWOZ 2.1, Sim-M, Sim-R (Shah et al., 2018)
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 (Henderson et al., 2014a)
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

Table A5: Evaluation metrics used for performance comparison among the methodologies. We focused on metrics evaluating the belief state of each turn. For convenience, the name of each metric is abbreviated. JGA: Joint Goal Accuracy, SA: Slot Accuracy.

Turn	Predicted State	Gold State	Slot Acc.	Relative Slot Acc.
0	restaurant-name-nusha	-	0.9667	0
1	restaurant-name-nusha	-	0.9667	0
2	restaurant-name-nusha	attraction-name-nusha	0.9333	0
3	restaurant-name-nusha	attraction-name-nusha	0.9333	0
4	restaurant-area-centre restaurant-food-indian	attraction-name-nusha restaurant-area-centre restaurant-food-indian	0.9667	0.6667
5	restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	attraction-name-nusha restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	0.9667	0.7500
6	restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	attraction-name-nusha restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	0.9667	0.8000
7	restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	attraction-name-nusha restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	0.9667	0.8000
8	restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	attraction-name-nusha restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	0.9667	0.8000
9	restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	attraction-name-nusha restaurant-name-saffron brasserie restaurant-area-centre restaurant-food-indian restaurant-pricerange-expensive	0.9667	0.8000

Table A6: SOM-DST prediction of MultiWOZ 2.1 test sample (PMUL4648.json). The joint goal accuracy of every turn is 0 because of belief states with red color. When calculating score, the number of total slots is set to 30, which is of *hotel*, *train*, *restaurant*, *attraction*, and *taxi* domains in MultiWOZ 2.1. Relative slot accuracy can be calculated just using slot-values appearing in the dialogue, not being affected by unused information.

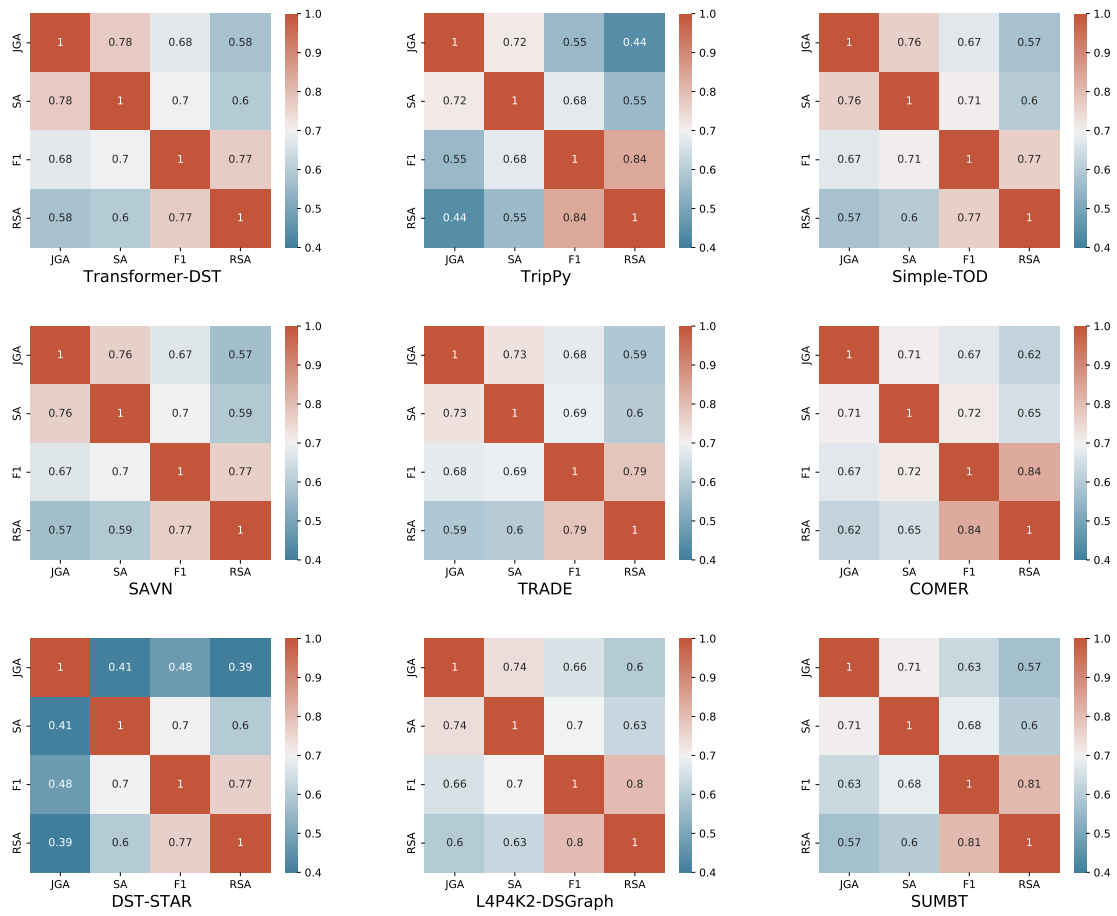


Figure A1: Correlation matrices of evaluation performance using various DST models.