Towards Fair Evaluation of Dialogue State Tracking by Flexible Incorporation of Turn-level Performances

Suvodip Dey, Ramamohan Kummara, Maunendra Sankar Desarkar
Indian Institute of Technology Hyderabad, India
cs19resch01003@iith.ac.in, cs19mds11004@iith.ac.in, maunendra@cse.iith.ac.in

Abstract

Dialogue State Tracking (DST) is primarily evaluated using Joint Goal Accuracy (JGA) defined as the fraction of turns where the ground-truth dialogue state exactly matches the prediction. Generally in DST, the dialogue state or belief state for a given turn contains all the intents shown by the user till that turn. Due to this cumulative nature of the belief state, it is difficult to get a correct prediction once a misprediction has occurred. Thus, although being a useful metric, it can be harsh at times and underestimate the true potential of a DST model. Moreover, an improvement in JGA can sometimes decrease the performance of turn-level or non-cumulative belief state prediction due to inconsistency in annotations. So, using JGA as the only metric for model selection may not be ideal for all scenarios. In this work, we discuss various evaluation metrics used for DST along with their shortcomings. To address the existing issues, we propose a new evaluation metric named **F**lexible **G**oal **A**ccuracy (**FGA**). FGA is a generalized version of JGA. But unlike JGA, it tries to give penalized rewards to mispredictions that are locally correct i.e. the root cause of the error is an earlier turn. By doing so, FGA considers the performance of both cumulative and turn-level prediction flexibly and provides a better insight than the existing metrics. We also show that FGA is a better discriminator of DST model performance.

1 Introduction

Dialogue State Tracking (DST) is at the core of task-oriented dialogue systems. It is responsible for keeping track of the key information exchanged during a conversation. With the growing popularity of task-based conversational agents, it is essential to review the evaluation of DST to appropriately measure the progress in this evolving area.

The task of DST is to predict the user intent through dialogue states (Henderson et al., 2014). Fig. 1 shows an example DST task from MultiWOZ (Budzianowski et al., 2018) dataset. Let U_t and S_t be the user and system utterances respectively at turn t. Then a typical conversation can be expressed as $D = \{U_0, (S_1, U_1), ...(S_n, U_n)\}$. The commonly used ground-truth dialogue state for DST is the belief state. Belief state B_t for turn t is defined as the set of *(domain, slot, slot-value)* triplets that have been extracted till turn t, thereby it is cumulative in nature. The objective of DST is to predict B_t given the dialogue history till turn t.

The primary metric for evaluating DST is Joint Goal Accuracy (JGA). It compares the predicted dialogue states to the ground truth B_t at each dialogue turn t (Henderson et al., 2014). As the belief state is cumulative, it is very unlikely for a model to get back a correct prediction after a misprediction. This is why it can provide an underestimated performance in certain cases. Besides, JGA completely ignores the performance of turn-specific local predictions. Let T_t be the turn-level belief state that contains all the intents or *(domain, slot, slot-value)* triplets expressed by the user only at turn t. Ideally, a model with higher JGA should also perform equally well to predict T_t. But, we observe that improving JGA can sometimes degrade the performance of predicting T_t mainly due to the presence of annotation inconsistencies in the available datasets. For example, in Fig. 1, the presence of *(hotel, area, centre)* and absence of *(attraction, name, all saints church)* in ground-truth B_2 and B_4 shows such inconsistencies. So, the generalization of the model may get compromised if the model selection is done only using JGA. Annotation inconsistencies and errors are common in real-world datasets. Hence, to provide a fair estimate, it requires not only track the performance of the cumulative belief state but also turn-level belief state as well.

In this work, we address these issues of JGA by proposing a novel evaluation metric for DST called **F**lexible **G**oal **A**ccuracy (**FGA**). The central idea of

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 318 - 324

FGA is to partially penalize a misprediction which is locally correct i.e. the source of the misprediction is some earlier turn. The main contributions of our work are as follows [1]:

- Detailed analysis of the existing DST metrics.
- Proposal of Flexible Goal Accuracy (FGA) than can keep track of both joint and turn-level performances simultaneously.
- Justification of FGA along with performance comparison on the MultiWOZ dataset.

2 Discussion on existing DST metrics

2.1 Joint goal accuracy

Joint accuracy or joint goal accuracy (JGA) checks whether the set of predicted belief states exactly matches the ground truth for a given user turn (Henderson et al., 2014; Wu et al., 2019). Let B_t and B'_t be the set of ground-truth and predicted belief states at turn t. Then the prediction of turn t is considered to be correct if and only if B_t exactly matches B'_t. Fig. 1 shows an illustration of the predicted belief state where the predictions of B'_t are generated using SOM-DST (Kim et al., 2020). In the example, there are 2 out of 6 correct predictions of B'_t that result in a JGA score of 33.33% for the whole conversation.

Although joint goal accuracy is a convenient metric to evaluate DST, it has certain limitations. The main source of the issue is the cumulative nature of ground-truth B_t. As a result, once a misprediction has occurred, it is difficult to get back a correct prediction in subsequent turns. For example, in Fig. 1, the prediction goes wrong in Turn 2 which affects all the later predictions. So, it is very likely to get a JGA of zero if the model somehow mispredicts the first turn. Therefore, JGA can undermine the true potential of a DST model and provide an underestimated performance.

In addition, JGA does not take into account turn-level performances. For instance, in Fig. 1, Turn 3 and 5 are locally correct but JGA will mark them 0 since B_t and B'_t has not matched exactly. Normally, it is expected that increasing the exact matches will also reflect in turn-level matches. But we observed that sometimes increasing exact matches can decrease turn-level matches mainly due to annotation inconsistencies. So, one should be careful while using only joint accuracy for model selection. Besides, the available DST datasets (like MultiWOZ) contain a lot of annotation errors (Zang et al., 2020). For example in turn 4, the model has predicted the intent *(attraction, name, all saints church)*. Although the prediction looks rational, the triplet is absent in the ground-truth. So, if a mismatch occurs due to an annotation error, it is highly probable that all the subsequent turns will be marked incorrect leading to an underestimated performance.

Hence, using joint goal accuracy for evaluating DST works fine if there are no annotation errors and the sole purpose is to improve the prediction of cumulative belief state. Otherwise, there is a need to include turn-level performance in order to obtain a fair evaluation of a DST model.

Turn		Conversation Details	Exact match	Turn match
0	U_0	Hi, I am traveling to Cambridge and could use some help for sure. I am so excited to see some local tourist attractions.	✓	✓
	B_0	{ }		
	B'_0	{ }		
1	S_1	We have 79 attractions to choose from, anything specific that you would like to tell us to help narrow it down?	✓	✓
	U_1	I'm looking for a hotel called cityroomz.		
	B_1	{**hotel**: {**name**: cityroomz}}		
	B'_1	{**hotel**: {**name**: cityroomz}}		
2	S_2	Cityroomz is a 0-star hotel in the center of town. Its address is Sleeperz Hotel, Station Road.	×	×
	U_2	Can you please book a room for 4 people for 2 nights starting on wednesday?		
	B_2	{**hotel**: {**area**: centre, **day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz, **stars**: 0}}		
	B'_2	{**hotel**: {**day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz}}		
3	S_3	Booking was successful.Reference number is : WGUYAGN2 anything else i can help?	×	✓
	U_3	Thanks. I am also looking for places to go in town. Perhaps an attraction in the city centre.		
	B_3	{**attraction**: {**area**: centre}, **hotel**: {**area**: centre, **day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz, **stars**: 0}}		
	B'_3	{**attraction**: {**area**: centre}, **hotel**: {**day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz}}		
4	S_4	I have the all saints church located at jesus lane and it's free entrance.	×	×
	U_4	That sounds perfect. Thanks!		
	B_4	{**attraction**: {**area**: centre}, **hotel**: {**area**: centre, **day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz, **stars**: 0}}		
	B'_4	{**attraction**: {**area**: centre, **name**: all saints church}, **hotel**: {**day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz}}		
5	S_5	Can I help you with anything else?	×	✓
	U_5	No thanks. That's all I need. Goodbye.		
	B_5	{**attraction**: {**area**: centre}, **hotel**: {**area**: centre, **day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz, **stars**: 0}}		
	B'_5	{**attraction**: {**area**: centre, **name**: all saints church}, **hotel**: {**day**: wednesday, **people**: 4, **stay**: 2, **name**: cityroomz}}		

Figure 1: Illustration of DST task. "Exact Match" compares Ground truth belief state B_t and Predicted belief state B'_t. "Turn Match" indicates the correctness of turn-level non-cumulative belief state prediction. Arrows represent the propagation of errors.

[1] Code is available at github.com/SuvodipDey/FGA

2.2 Slot Accuracy

Slot accuracy (SA) is a relaxed version of JGA that compares each predicted *(domain, slot, slot-value)* triplet to its ground-truth label individually (Wu et al., 2019). Let S be the set of unique domain-slot pairs in the dataset. Let B_t and B'_t be the set of ground-truth and predicted belief states respectively. Then slot accuracy at turn t is defined as

$$SA = \frac{|S| - |X| - |Y| + |P \cap Q|}{|S|}, \quad (1)$$

where $X = (B_t \setminus B'_t)$, $Y = (B'_t \setminus B_t)$, P is the set of unique domain-slot pairs from X, and Q is the set of unique domain-slot pairs from Y. Basically, in Equation 1, $|X|$ and $|Y|$ represent the number of false negatives and false positives respectively. Note that if the value of a ground-truth domain-slot pair is wrongly predicted then this misprediction will be counted twice (once in both X and Y). The term $|P \cap Q|$ in the above equation helps to rectify this overcounting. In MultiWOZ, the value of $|S|$ is 30. For Turn 2 in our running example, since $|B_1 \setminus B'_1| = 2$ and $|B'_1 \setminus B_1| = 0$, slot accuracy is equal to $\frac{(30-2-0-0)}{30}$ i.e. 93.33%. Slot accuracy for the entire conversation in Fig. 1 is 94.44%.

The value of slot accuracy can be very misleading. For instance, even if the prediction of Turn 2 is wrong in Fig. 1, we get a slot accuracy of 93.33% which is extremely high. Basically, slot accuracy overestimates the DST performance. Let us exhibit this fact by considering the case where we predict nothing for all turns i.e. $B'_t = \emptyset, \forall t$. Then, slot accuracy simplifies to $\frac{|S|-|B_t|}{|S|}$. It is natural that $|B_t| << |S|$ because a conversation will typically have only a small number of domain-slot pairs *live* at any time. As a result, slot accuracy remains on the higher side ($\approx 81\%$ for MultiWOZ 2.1) even if we predict nothing. For datasets with a larger number of domain/slots, since $|S|$ is large, slot accuracy will be close to 1 for almost all scenarios. Thus, slot accuracy is a poor metric to evaluate DST.

2.3 Average Goal accuracy

Average goal accuracy (AGA) is a relatively newer metric proposed to evaluate the SGD dataset (Rastogi et al., 2020). Here, the slots that have a non-empty assignment in the ground-truth dialogue state are only considered during evaluation. Let $N_t \subseteq B_t$ be the set of ground-truth triplets having non-empty slot-values. Then AGA is computed as $\frac{|N_t \cap B'_t|}{|N_t|}$ where B'_t is the predicted belief state for turn t. The turns having $N_t = \emptyset$ are ignored during the computation of AGA. In Fig. 1, AGA for turn 2 is 4/6, and 76.19% for the entire conversation.

This metric has mainly two limitations. Firstly, AGA is only recall-oriented and thereby does not consider the false positives. Ignoring the false positives makes this metric insensitive to extraneous triplets in the predicted belief state. However, this issue can be easily addressed by redefining AGA as $\frac{|N_t \cap B'_t|}{|N_t \cup B'_t|}$. But there still exists a second major problem with AGA. Note that even if a turn is completely wrong, AGA for that turn can still be higher because of the correct predictions in the previous turns. For example, even if turn 2 and 4 are incorrect, we get an AGA of 4/6 and 5/7 respectively which clearly indicates an overestimation.

3 Flexible Goal Accuracy

From the previous discussion, it is evident that despite a few limitations, joint goal accuracy is superior to the other two metrics. This is why with the objective to obtain a better evaluation metric for DST, we address the shortcomings of JGA by proposing a new metric called Flexible goal accuracy (FGA). The description of FGA is presented in the next part of this section, whereas its working is described as a pseudo-code in Algo. 1.

For a given a turn t, an error in belief state prediction (i.e. $B_t \neq B'_t$) can occur in two ways: 1) the source of the error is turn t itself i.e. the turn-level prediction is wrong, 2) the turn-level prediction of turn t is correct but the source of the error is some earlier turn $t_{err} \prec t$. FGA works differently from JGA only for type 2 errors. Unlike JGA, FGA does not penalize type 2 errors completely. It assigns a penalized score based on the distance between the error turn (t_{err}) and the current turn (t) and the penalty is inversely proportional to this distance ($t - t_{err}$). The main idea is to forget the mistakes with time in order to attain a fair judgment of a DST model offline.

We decide the correctness of a turn-level match using the logic shown in line 10 of Algo. 1. A turn $t > 0$ is locally correct if ($T'_t \subseteq B_t$ and $T_t \subseteq B'_t$) where $T_t = B_t \setminus B_{t-1}$ and $T'_t = B'_t \setminus B'_{t-1}$. In other words, a turn-level or local match indicates that all the intents shown by the user in a particular turn have been correctly detected without any false positives. Just comparing T_t and T'_t to check a turn-level or local match can be erroneous because it will not credit the model for error corrections.

Algorithm 1: FGA for single conversation

Input: B = list of groun-truth belief states,
B' = list of predicted belief states,
N = #turns
Output: Flexible goal accuracy

1 $T = \{0, 1, \ldots, N-1\}$, $t_{err} \leftarrow -\infty$, f = 0
2 **for** $t \in T$ **do**
3 $w \leftarrow 1$
4 **if** $B_t \neq B'_t$ **then**
5 **if** $t = 0$ **then**
 /* Type 1 error */
6 $w \leftarrow 0$, $t_{err} \leftarrow t$
7 **else**
8 $T_t \leftarrow B_t \setminus B_{t-1}$
9 $T'_t \leftarrow B'_t \setminus B'_{t-1}$
10 **if** $T'_t \not\subseteq B_t$ *or* $T_t \not\subseteq B'_t$ **then**
 /* Type 1 error */
11 $w \leftarrow 0, t_{err} \leftarrow t$
12 **else**
 /* Type 2 error */
13 $x \leftarrow (t - t_{err})$
14 $w \leftarrow 1 - \exp(-\lambda x)$
15 $f \leftarrow f + w$
16 **return** f/N

For the penalty function, we use the CDF of exponential distribution (shown in Line 14 of Algo. 1) parameterized by λ where $\lambda \geq 0$. Clearly, the strictness of FGA is inversely proportional to λ. Note that $\lambda = 0$ will reduce FGA to JGA (strict metric) whereas $\lambda \to \infty$ will report only the accuracy on turn-level matches (relaxed metric). Finding the appropriate λ for a specific DST task should be done carefully in order to match the desired evaluation criteria. However, we can take a theoretical stand and approximate the hyper-parameter value as $\lambda = -ln(1-p)/t_f$ where t_f is the number of turns that it will take to forget a mistake by factor p where $(0 \leq p < 1)$. For example, if t_f=6 and p=0.95, then λ=0.499. So, the strictness of FGA is directly proportional to t_f and inversely proportional to p. If the dataset is clean, one can alternatively find the best λ through a human evaluation, although it would require additional human effort. Hence, we can flexibly set the strictness criteria of FGA through the hyper-parameter λ according to our requirement.

In our running example (Fig. 1), the FGA score for each turn with $\lambda = 0.5$ is {1, 1, 0, 0.39, 0, 0.39} which results in a FGA score of 46.33% for the entire conversation. We can observe two things from these numbers. Firstly, it is not overestimating in comparison to SA and AGA. Secondly, it gives a better estimate than JGA in keeping track of both exact and turn-level matches simultaneously. Hence, FGA can provide a relatively balanced estimate than the existing metrics even in the presence of annotation errors and inconsistencies.

4 Result and Analysis

In this section, we report the performance of FGA along with the other metrics on four different DST models: TRADE (Wu et al., 2019), Hi-DST (Dey and Desarkar, 2021), SOM-DST (Kim et al., 2020), and Trippy (Heck et al., 2020). We use the MultiWOZ 2.1 dataset (Eric et al., 2020) as most of the recent progress in DST are showcased on this dataset. The results are reported in Table 1. Since the MultiWOZ dataset covers many domains (hotel, restaurant, taxi, train, attraction) where each domain may have different levels of tolerance (intuitively train, taxi booking may be strict whereas information seeking about attraction, restaurant domains may be lenient), an overall common/single strictness setting for the entire dataset may be difficult to reach at. Hence, we reported the FGA score for multiple values of hyper-parameter λ rather than showing the result for a single value. For the same reason, we did not try to find the best λ for evaluating the MultiWOZ dataset.

From Table 1, we can observe that Trippy has the best JGA. Currently, most of the state-of-the-art DST performances are shown using Trippy. However, we can notice that Trippy does not have the same performance gain for turn-level matches. It has lesser turn-level matches than SOM-DST and Hi-DST. This behavior of Trippy can be a side-effect of boosting the JGA using its intricate featurization. In contrast, Hi-DST optimizes explicitly for turn-level non-cumulative belief states, thereby achieving better turn-level accuracy at the expense of JGA. Among the four models, SOM-DST performs well for both objectives because of their sophisticated selective overwrite mechanism. Now, by comparing the numbers of Table 1, we can infer that FGA does a better job in providing a fair estimate while considering both exact and turn-level matches. Moreover, we can also notice that FGA acts as a better discriminator of DST models in comparison to the existing metrics.

Human Evaluation: We conducted a human

Model	#Turns	#M1	#M2	JGA	SA	AGA	$FGA_{0.25}$	$FGA_{0.5}$	$FGA_{0.75}$	FGA_{1}
TRADE	7368	3600	5287	48.86%	96.96%	88.79%	56.58%	61.19%	64.16%	66.18%
Hi-DST	7368	3622	5903	49.16%	96.70%	90.74%	61.31%	67.69%	71.47%	73.91%
SOM-DST	7368	3912	6084	53.09%	97.36%	91.71%	64.94%	71.04%	74.61%	76.88%
Trippy	7368	3926	5875	53.28%	97.30%	90.75%	63.24%	68.67%	71.97%	74.13%

Table 1: Comparison of DST metrics. "M1" and "M2" represents exact and turn-level matches respectively. "FGA_x" indicates the FGA value calcualated using λ=x.

evaluation involving 11 evaluators on 100 randomly picked conversations from the MultiWOZ 2.1 test data. For each turn in a conversation, we provided the system and user utterances along with the ground-truth and predicted belief states. The predictions were generated using SOM-DST. For each conversation, the evaluators were asked to report their satisfaction (1) or dissatisfaction (0) with the performance of the model in keeping track of user intent throughout the conversation. Pearson correlation coefficient of JGA and FGA (with $\lambda = 0.5$) with human ratings came out to be 0.33 and 0.37 respectively. This shows that FGA is slightly better correlated than JGA with human evaluation.

5 Conclusion

In this work, we analyzed the limitations of existing DST metrics. We argued that joint accuracy can underestimate the power of a DST algorithm, whereas slot and average goal accuracy can overestimate it. We addressed the issues of joint accuracy by introducing Flexible goal accuracy (FGA) which tries to give partial credit to mispredictions that are locally correct. We justified that FGA provides a relatively balanced estimation of DST performance along with better discrimination property. In conclusion, FGA is a practical and insightful metric that can be useful to evaluate future DST models.

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.

Suvodip Dey and Maunendra Sankar Desarkar. 2021. Hi-DST: A hierarchical approach for scalable and extensible dialogue state tracking. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 218–227, Singapore and Online. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, 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. 2014. 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, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo 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.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8689–8696.

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.

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 baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

A Appendix

A.1 MultiWOZ Dataset

MultiWOZ (Budzianowski et al., 2018) is a popular DST corpus that contains both single and multi-domain conversations. For this work, we used MultiWOZ 2.1 (Eric et al., 2020) which is an updated version of the original MultiWOZ 2.0 dataset. In addition to the original dataset, MultiWOZ 2.1 contains fixes to some noisy annotations. Table 2 shows few elementary statistics of the dataset.

Data	#Conversations	#Turns	Avg. turns
Train	8420	56668	6.73
Dev	1000	7374	7.37
Test	999	7368	7.37

Table 2: Elementary statistics of MultiWOZ 2.1 dataset. "Avg. turns" indicate average turns per conversation.

A.2 Result generation procedure

We generated results for four DST models - Trade (Wu et al., 2019) [2], Hi-DST (Dey and Desarkar, 2021) [3], SOM-DST (Kim et al., 2020) [4], and Trippy (Heck et al., 2020) [5]. We used their official code to train them on MutiWOZ 2.1 dataset. All four models generate an inference file that contains the predicted belief states for the test set. We used these inference files to compute the values of different metrics shown in Table 1. As we trained all the models from scratch, the results may not be exactly the same as those reported in the original paper.

A.3 Human evaluation format

For each randomly picked conversation for human evaluation, we prepared a file that logged the utterances, ground-truth, and predicted belief state for each turn. Additionally, we indicated whether the ground truth exactly matched the predicted belief state to speed up the evaluation process. A sample file format is shown in Fig. 2.

[2] github.com/jasonwu0731/trade-dst
[3] github.com/SuvodipDey/Hi-DST
[4] github.com/clovaai/som-dst
[5] gitlab.cs.uni-duesseldorf.de/general/dsml/trippy-public

```
Dialogue ID : MUL0379.json
----------
Turn: 0
Sys :
Usr : I am looking to get to the Rajmahal restaurant please, how do I get there?

GT  : {'restaurant': {'name': 'rajmahal'}}
PR  : {'restaurant': {'name': 'rajmahal'}}
Matched : True
----------
Turn: 1
Sys : Would you like for me to book you a taxi to the restaurant?
Usr : I need you to book the restaurant for me if that's okay. For 2 people at 19:45 on tuesday is what I request. Can I get the reference number too?

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}}
Matched : True
----------
Turn: 2
Sys : Okay I booked it and your reference number is 8D21ZMGT. Have a great day.
Usr : Actually, I'm also looking for a train. I need to go to London Kings Cross on the same day as the restaurant booking.

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'departure': 'london kings cross'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'day': 'tuesday', 'destination': 'london kings cross'}}
Matched : False
----------
Turn: 3
Sys : No problem. Would you like to specify where you're departing from and what time you'd like?
Usr : I am departing from London Kings Cross and need to go to Cambridge. I want to arrive by 09:15.

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'departure': 'london kings cross', 'destination': 'cambridge'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
Matched : False
----------
Turn: 4
Sys : I have several options to get you where you are going that arrive before 9:15. Which day would you be traveling?
Usr : I will be traveling on Tuesday.

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
Matched : True
----------
Turn: 5
Sys : There are two trains for that search. Would you look me to book you the one that leaves at 05:17?
Usr : What are the travel times for those trains?

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
Matched : True
----------
Turn: 6
Sys : They are both 51 minutes.
Usr : Thank you, that should be all for today.

GT  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
PR  : {'restaurant': {'day': 'tuesday', 'people': '2', 'time': '19:45', 'name': 'rajmahal'}, 'train': {'arriveby': '09:15', 'day': 'tuesday', 'departure': 'london kings cross', 'destination': 'cambridge'}}
Matched : True
----------
```

Figure 2: Data format for human evaluation