# DST Human Evaluation guideline

**What is Dialogue State Tracking (DST)?**
Dialogue State Tracking (DST) is an important task for a task-oriented dialogue system to track the user intent in an ongoing conversation. It is an integral part of a task-oriented chatbot like online booking chatbot. The intents captured by the DST are used to run a query in the backend database (just like a booking agent) to search and book the desired entity.

**Belief State:** In DST, user intent at each turn is represented by a belief state**. Belief state for a given turn t contains all the intents shown by the user till turn t.** So, a belief state is cumulative in nature since it contains all the intents till the current turn. In a belief state, an intent is represented by a **(domain, slot, slot-value)** triplet. The domain is the topic of conversation. In our dataset, there are only five domains - **attraction, hotel, restaurant, train, and taxi**. Slots are the attributes of domains on which the conversation is happening. Slot values are the value of the attributes. In our dataset, there are only a limited number of slots for each domain -

- **Attraction**: 'name', 'type', 'area'
- **Hotel**: 'name', 'type', 'parking', 'area', 'day', 'stay', 'internet', 'people', 'stars', 'pricerange'
- **Restaurant**: 'name', 'food', 'area', 'day', 'time', 'people', 'pricerange'
- **Taxi**: 'arriveby', 'departure', 'leaveat', 'destination'
- **Train**: 'arriveby', 'day', 'leaveat', 'destination', 'departure', 'people'

Example:-
Turn 0:
*User : Could you find me an attraction in the east part of town?*
*Belief State : {(attraction, area, east)}*

In the above example, the user wants to find the attraction in the east part of town. So, the belief state for this turn contains {(attraction, area, east)} as shown as above.

*Turn 1:*
*System: Definitely, My favourite place in the east is the Funky Fun House. It's funky and fun.*
*User : Sure, I will go with Funky Fun House. Can I have a phone number please?*
*Belief State : {(attraction, area, east), (attraction, name , Funky Fun House)}*

In this turn, the system gave Funky Fun House as result for user query in previous turn and user is fine with it. Belief State is updated as shown above by appending user intent (attraction, name, Funky Fun House).

**DST Objective:** The objective of DST is to capture the belief state correctly at each turn.

**Note:** A user can mention slots outside these predefined slots but those should not appear in the belief state. More detailed description regarding each slot can be found in slots_descriptions.json file.

**What is the objective of this human evaluation?**
The task of this human evaluation is to rate the performance of the chatbot on detecting the user intentions throughout the conversation. You have to judge the performance of the model on the entire conversation and report your satisfaction (1) or dissatisfaction (0).

**Data format**
For each conversation turn, you will be provided with the actual belief state (GT) and the predicted belief state (PR) generated by a model.

Sample turn data:
*Turn: 0*
*Sys :*
*Usr : I'm looking for a moderately priced place to stay.*

*GT : {'hotel': {'pricerange': 'moderate'}}*
*PR : {'hotel': {'pricerange': 'moderate'}}*

*Matched : True*

Turn - the current turn number.
Sys - System generated response.
Usr - User input
GT : Golden Truth Belief State.
PR : System Predicted Belief State.
Matched: True when GT and PR are exactly the same, false otherwise.

Inorder to save the space, the belief state is aggregated by domain in the evaluation data.

**What are the characteristics of belief state in our task?**

- Belief state is used to form a sql query using which the desired items are searched or booked. A booking agent first gathers as much information as possible and then recommends or books. Similarly, the booking chatbot tries to gather as much information using the belief state and then query the backend when required. This is why the belief state is kept cumulative.

- Although the belief state is cumulative, there are few instances where an intent or (domain,slot,value) triplet from the previous turn can get deleted in the current turn. For example, if the user came to book a train ticket for 2 people but somehow at the end the agent was not able to book the ticket. Then all the intents will remain in the belief state except (train-people). **This kind of rare deletion can occur for the following slots - stay, day, people, and time.**

- In this dataset, the belief state can only consist of a domain/slot from the above mentioned list. Domain or slots outside this list should not appear in the belief state. For example, phone number, postal code, reference number and many other possible slots cannot be part of the belief state for this specific task.

- A user can do an inquiry regarding a slot. For example, “What is the hotel area?”. **This kind of user inquiry will not appear in the belief state.** This is because this type of inquiry cannot be expressed as (domain,slot,value) triplet. For example, in “What is the hotel area?”, domain is hotel, slot is area, but there is no value for the slot..

**Few points to keep in mind during evaluation**

- While evaluating a conversation, please go through the details of the domain-slot pairs that appeared in the conversation.

- Some slots can take only a limited set of values. For example - type, area, parking, internet, pricerange, stars, food etc. Please be careful while evaluating those slots. You can find the details in the ontology.json file.

- Some slot values can be “**dont care**”. For example, in “I am looking for a hotel in the west, internet is optional”, slot-value for the slot internet will be “don’t care”.

- As discussed earlier, there are few instances where a domain-slot from a previous turn can get deleted in the current turn. Such instances will involve slots like *stay, day, people, and time*. Be careful when you encounter such cases and judge it accordingly.

- Beware of user inquiries. User inquiries are not part of the belief states. So, please do not expect the model to capture them.

- We are also providing the information whether the actual and predicted belief state have an exact match or not. This information is only provided to speed-up your evaluation. Evaluating only on the basis of this information can be erroneous. For example, if the model mispredicted the first turn, then it is very difficult for the model to get an exact match in the subsequent turns due to the cumulative nature of the belief state. But maybe the model did actually well in the subsequent turns. Just because it missed an intent in the first turn, it is not able to get an exact match.

- The model can mispredict in two ways - 1) failed to detect some intent, 2) predicted some undesired intent. Penalize the model for both kinds of mistakes. Some intentions are more difficult to predict than others. You can penalize based on the difficulty level of the prediction. Interpretation of easy/difficult is up to you.

- Sometimes ground truth can be incorrect or ambiguous. It's up to the judge to make a conscious decision.