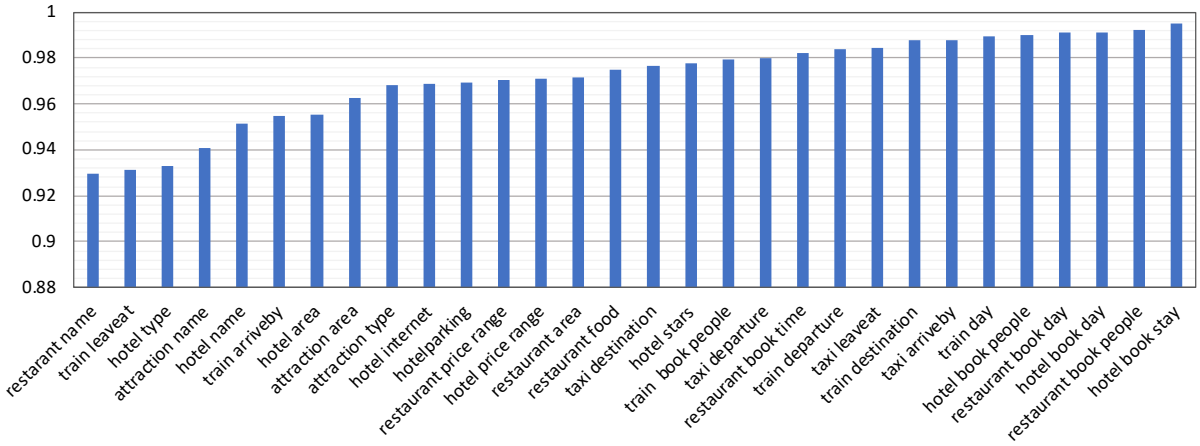

Figure 6: Accuracy of each slot per turn on MultiWOZ 2.1 dataset

A.6 Examples of Prediction Errors

This section describes prediciton errors made by DSTQA w/span. Incorrectly predicted (domain, slot, value) tuples are marked by underlines.

1. Annotation errors
Description: The groud truth label in the dataset is wrong. This can happen either 1) annotators neglect slots mentioned in the user utterance 2) annotators mistakenly choose the wrong label of a slot.
Examples:

User: I would like to find a *museum* in the *west* to go to. Agent: There are several museums in the west. I recommend the *Cafe Jello Gallery*. User: Can I have the address of the Cafe Jello museum? Agent: The Cafe Jello Gallery is at 13 Magdalene street. Is there anything else? User: Is there a *moderately* priced *British* restaurant *any where* in town?
Annotation: {(restaurant, food, British), (restaurant, price range, moderate), (restaurant, area, west)} **Prediction**: {(restaurant, food, Biritsh), (restaurant, price range, moderate), (restaurant, area, don't care)}

2. Annotator disagreement on user confirmation (labeled)
Description: This type of errors comes from the disagreement between annotators when generating ground truth labels. More specifically, in a dialogue, the agent sometimes proposes a suggestion (a value of a slot) to the user, followed by the user's positive confirmation. For example, the agent says 'I would recommend Little Seoul. Would you like to make a reservation?'. The user confirms with 'yes, please'. Since the user positively confirms the agents suggestion, the (domain, slot, value) tuple mentioned by the agent, or, (restaurant, name, Little Seoul) tuple in this example, can be added into the belief state. However, based on our observation of the MultiWOZ 2.0 and MultiWOZ 2.1 dataset, the annotators are inconsistent, and only about half of the times these tuples are added to the belief states. An error of this type comes from the scenario that the tuple is added to the belief state by the annotator but not by the model (i.e. the model predicts None for the corresponding (domain, slot) pair).
Examples: