	Joint Accuray		Slot Accuracy	
	TRADE	DSTQA w/span	TRADE	DSTQA w/span
Restaurant	65.35	**68.68**	93.28	94.08
Hotel	55.52	**61.76**	92.66	93.72
Train	77.71	**79.75**	95.30	95.61
Attraction	71.64	**74.05**	88.97	90.53
Taxi	76.13	**78.22**	89.53	90.37

Table 6: Model performance on each of the 5 domains.

has a 11.24% relative improvement. Hotel is the hardest domain as it has the most slots (10 slots) and has the lowest joint accuracy among all domains.

A.4 Joint Accuracy v.s. Context Length

We further show the model performance on different context lengths. Context lengths means the number of previous turns included in the dialogue context. Note that our baseline algorithms either use all previous turns as contexts to predict belief states or accumulate turn-level states of all previous turns to generate belief states. The results are shown in Figure 4. We can see that DSTQA with graph outperforms DSTQA without graph. This is especially true when the context length is short. This is because when the context length is short, graph carries information over multiple turns which can be used for multi-turn inference. This is especially useful when we want a shorter context length to reduce computational cost. In this experiment, the DSTQA model we use is DSTQA w/span.

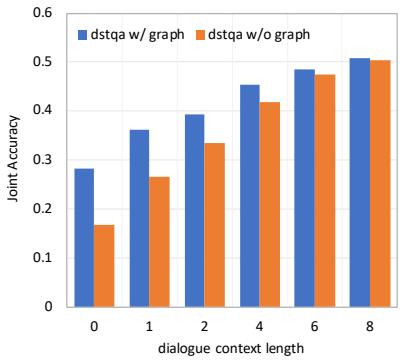

Figure 4: Joint acc. v.s. context length

A.5 Accuracy per Slot

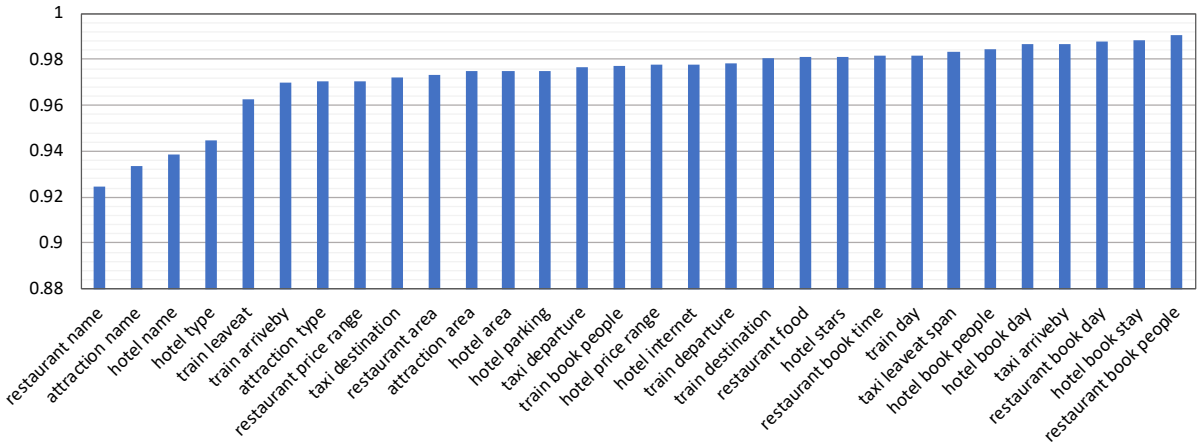

Figure 5: Accuracy of each slot per turn on MultiWOZ 2.0 dataset

The accuracy of each slot on MultiWOZ 2.0 and MultiWOZ 2.1 test set is shown in Figure 5 and Figure 6, respectively. Named related slots such as *restaurant name*, *attraction name*, *hotel name* has high error rate, because these slots have very large value set and high annotation errors.