A Appendix

A.1 Results on WOZ 2.0 dataset

We also evaluate our algorithm on WOZ 2.0 dataset (Mrkšić et al., 2017) WOZ 2.0 dataset has 1200 restaurant domain task-oriented dialogues. There are three slots: 'food', 'area', 'price range', and a total of 91 slot values. The dialogues are collected from a Wizard of Oz style experiment, in which the task is to find a restaurant that matches the slot values the user has specified. Each turn of a dialogue is annotated with a dialogue state, which indicates the slot values the user has informed. One example of the dialogue state is {*'food:Mexican', 'area':'east', price range:'moderate'*}.

Model	Joint Accuracy
NBT	84.4
GLAD	88.1
GCE	88.5
StateNet PSI	88.9
SUMBT	**91.00**
DSTQA	90.0

Table 4: Joint accuracy on WOZ 2.0 dataset.

Table 4 shows the results on WOZ 2.0 dataset. We compare with four published baselines. SUMBT (Lee et al., 2019) is the current state-of-the-art model on WOZ 2.0 dataset. It fine-tunes a pre-trained BERT model (Devlin et al., 2019) to learn slot and utterance representations. StateNet PSI (Ren et al., 2018) maps contextualized slot embeddings and value embeddings into the same vector space, and calculate the Euclidean distance between these two. It also learns a joint model of all slots, enabling parameter sharing between slots. GLAD (Zhong et al., 2018) proposes to use a global module to share parameters between slots and a local module to learn slot-specific features. Neural Beflief Tracker (Mrkšić et al., 2017) applies CNN to learn n-gram utterance representations. Unlike prior works that transfer knowledge between slots by sharing parameters, our model implicitly transfers knowledge by formulating each slot as a question and learning to answer all the questions. Our model has a 1.24% relative joint accuracy improvement over StateNet PSI. Although SUMBT achieves higher joint accuracy than DSTQA on WOZ 2.0 dataset, DSTQA achieves better performance than SUMBT on MultiWOZ 2.0 dataset, which is a more challenging dataset.

A.2 MultiWOZ 2.0/2.1 Ontology

The ontology of MultiWOZ 2.0 and MultiWOZ 2.1 datasets is shown in Table 5. There are 5 domains and 30 slots in total. (two other domains 'hospital' and 'police' are ignored as they only exists in training set.)

Domains	Restaurant	Hotel	Train	Attraction	Taxi
Slots	name area price range food book people book time book day	name area price range type parking stars internet book stay book day book people	destination departure day arrive by leave at book people	name area type	destination departure arrive by leave at

Table 5: Domain ontology in MultiWOZ 2.0 and MultiWOZ 2.1 dataset

A.3 Performance on Each Individual Domain

We show the performance of DSTQA w/span and TRADE on each single domain. We follow the same procedure as Wu et al. (2019) to construct training and test dataset for each domain: a dialogue is excluded from a domain's training and test datasets if it does not mention any slots from that domain. During the training, slots from other domains are ignored. Table 6 shows the results. We can see that our model achieves better results on every domain, especially the hotel domain, which