5 Experiments

We evaluate our model on three publicly available datasets: (non-multi-domain) WOZ 2.0 (Mrkšić et al., 2017), MultiWOZ 2.0 (Budzianowski et al., 2018), and MultiWOZ 2.1 (Eric et al., 2019). Due to limited space, please refer to Appendix A.1 for results on (non-multi-domain) WOZ 2.0 dataset. MultiWOZ 2.0 dataset is collected from a Wizard of Oz style experiment and has 7 domains: *restaurant*, *hotel*, *train*, *attraction*, *taxi*, *hospital*, and *police*. Similar to Wu et al. (2019), we ignore the *hospital* and *police* domains because they only appear in training set. There are 30 (domain, slot) pairs and a total of 10438 task-oriented dialogues. A dialogue may span across multiple domains. For example, during the conversation, a user may book a restaurant first, and then book a taxi to that restaurant. For both datasets, we use the train/test splits provided by the dataset. The domain ontology of the datasets is described in Appendix A.2. MultiWOZ 2.1 contains the same dialogues and ontology as MultiWOZ 2.0, but fixes some annotation errors in MultiWOZ 2.0.

Two common metrics to evaluate dialogue state tracking performance are ***Joint*** accruacy and ***Slot*** accuracy. Joint accuracy is the accuracy of dialogue states. A dialogue state is correctly predicted only if all the values of (domain, slot) pairs are correctly predicted. Slot accuracy is the accuracy of (domain, slot, value) tuples. A tuple is correctly predicted only if the value of the (domain, slot) pair is correctly predicted. In most literature, joint accuracy is considered as a more challenging and more important metric.

5.1 Implementation Details

Existing dialogue state tracking datasets, such as MultiWOZ 2.0 and MultiWOZ 2.1, do not have annotated span labels but only have annotated value labels for slots. As a result, we preprocess MultiWOZ 2.0 and MultiWOZ 2.1 dataset to convert value labels to span labels: we take a value label in the annotation, and search for its last occurrence in the dialogue context, and use that occurrence as span start and end labels. There are 30 slots in MultiWOZ 2.0/2.1 dataset, and 5 of them are time related slots such as *restaurant book time* and *train arrive by*, and the values are 24-hour clock time such as `08:15`. We do span prediction for these 5 slots and do value prediction for the rest of slots because it is not practical to enumerate all time values. We can also do span prediction for other slots such as *restaurant name* and *hotel name* with the benefit of handling out-of-vocabulary values, but we leave these experiments as future work. WOZ 2.0 dataset only has one domain and 3 slots, and we do value prediction for all these slots without graph embeddings.

We implement our model using AllenNLP (Gardner et al., 2017) framework.[1] For experiments with ELMo embeddings, we use a pre-trained ELMo model[2] in which the output size is $D^{ELMo} = 512$. The dimension of character-level embeddings is $D^{Char} = 100$, making $D^w = 612$. ELMo embeddings are fixed during training. For experiments with GloVe embeddings, we use GloVe embeddings pre-trained on Common Crawl dataset.[3] The dimension of GloVe embeddings is 300, and the dimension of character-level embeddings is 100, such that $D^w = 400$. GloVe embeddings are trainable during training. The size of the role embedding is 128. The dropout rate is set to 0.5. We use Adam as the optimizer and the learning rate is set to 0.001. We also apply word dropout that randomly drop out words in dialogue context with probability 0.1.

When training DSTQA with the dynamic knowledge graph, in order to predict the dialogue state and calculate the loss at turn t, we use the model with current parameters to predict the dialogue state up until turn $t-1$, and dynamically construct a graph for turn t. We have also tried to do teacher forcing which constructs the graph with ground truth labels (or sample ground truth labels with an annealed probability), but we observe a negative impact on joint accuracy. On the other hand, target network (Mnih et al., 2015) may be useful here and will be investigated in the future. More specifically, we can have a copy of the model that update periodically, and use this model copy to predict dialogue state up until turn $t-1$ and construct the graph.

5.2 Results on MultiWoz 2.0 and MultiWOZ 2.1 dataset.

[1] Code will be released on Github

[2] https://allennlp.org/elmo

[3] https://nlp.stanford.edu/projects/glove/