Supplementary Materials

Arshit Gupta[†] **Peng Zhang**[‡] **Garima Lalwani**[‡] **Mona Diab**[†]
[†]Amazon AI, Seattle [‡]Amazon AI, East Palo Alto
{arshig, pezha, glalwani, diabmona}@amazon.com

1 Appendix

1.1 Utterance Encoding

We first take the current user utterance for turn i, denoted by $\mathbf{U}_i \in \mathbb{R}^{d_e \times k}$ where k represents maximum number of tokens in the utterance. Then we pass this vector through an embedding layer to generate a hidden state, $\mathbf{H}_i \in \mathbb{R}^{d_h \times k}$ where d_h is the hidden layer dimension. We add learned absolute position embedding, $\mathbf{p}_u \in \mathbb{R}^{d_h \times k}$ (Gehring et al., 2017), resulting in a new vector, $\mathbf{H}_i = \mathbf{H}_i + \mathbf{p}_u$. The intuition is that this embedding will capture syntactic information.

We pass this input through a DiSAN unit that broadly comprises token-to-token (t2t) attention, masks, and source-to-token (s2t) attention:

- On input $\mathbf{H}_i$, we first apply a multi-dimensional **t2t** attention layer that encodes the dependency between a token and all other tokens in the sentence. In addition, forward and backward masks are added to attention computation to incorporate directional information;
- For each of these masks, we apply a fusion gate that controls the flow of information from the original hidden representation $\mathbf{H}_i$ and the mask outputs, $\mathbf{Hm}_i^F$ and $\mathbf{Hm}_i^B$, generating contextualized representations $\mathbf{C}_i^F$ and $\mathbf{C}_i^B$, respectively.

$$\mathbf{G} = \sigma(W^{(1)}\mathbf{H}_i + W^{(2)}\mathbf{Hm}_i^F + \mathbf{b})$$
$$\mathbf{C}_i^F = \mathbf{G} \odot \mathbf{H}_i + (1 - \mathbf{G}) \odot \mathbf{Hm}_i^F$$

  Similarly, we obtain $\mathbf{C}_i^B$. Both $\mathbf{C}_i^F$ and $\mathbf{C}_i^B$ are concatenated to render $\mathbf{C}_i$;
- Finally, we have a multi-dimensional **s2t** attention which learns a gating function that determines, element-wise, how to attend to each individual token of the sentence. It takes $\mathbf{C}_i$ as input and outputs a single vector for the entire sentence.

$$\mathbf{F}(\mathbf{C_i}) = W^{(1)T}\sigma(W^{(2)}\mathbf{C}_i + \mathbf{b}) + \mathbf{b}$$
$$\mathbf{h}(\text{Utt}_i) = \mathbf{F}(\mathbf{C_i}) \odot \mathbf{C_i}$$

1.2 Datasets

Below is the detailed description of both the conversational datasets:

1. **Booking** provides a shared test framework containing conversations between human and machines for the domain of restaurant. In original DSTC-2 data, a user's goal is to find information about restaurants based on certain constraints. The original data contains *states* and *goals* as it is mainly targeted for DST tasks (Zhong et al., 2018; Ren et al., 2018). To convert it to IC-SL task, we perform pre-processing on states and goals present in original data to derive intent labels and slots for each user utterance. Some of the sample intents are 'confirm_pricerange', 'request_slot_area', etc. There are only 3 slots - 'food', 'pricerange' and 'area'.
2. **Cable** is a synthetic dataset developed in-house. It is more diverse compared to Booking dataset. It is based on user conversations in the cable service domain. It includes 18 intents like 'ViewDataUsage', 'Help', 'StartService', etc. It has total of 64 slots such as 'UserName', 'CurrentZipCode', etc., yielding a more challenging dataset for modeling.
3. **SNIPS** dataset contains 16K crowdsourced queries. It has total of 7 intents ranging from Play Music to Book Restaurant. Training data has 13,784 utterances and the test data consists of 700 utterances.

4. **ATIS** dataset contains 4,978 training utterances and 893 test utterances. There are total of 18 intents and 127 slot labels.

Dataset	**Booking**	**Cable**	**ATIS**	**SNIPS**
Train Size	9351	1856	4978	13784
Val Size	4691	1814	–	–
Test Size	6727	1836	893	700
#Intents	19	21	18	7
#SL	5	26	127	41
#DA	4	4	–	–
#ConvLen	4.25	4.68	–	–

Table 1: Data statistics for Booking, Cable, ATIS and SNIPS. Legend - SL: Slot Labels; DA: Dialog Acts; ConvLen: Average Conversation length

References

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *EMNLP*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive dialogue state tracker. *CoRR*, abs/1805.09655.