1 Appendix

1.1 Utterance Encoding

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

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

- On input $H_i$, 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 $H_i$ and the mask outputs, $Hm_i^F$ and $Hm_i^B$, generating contextualized representations $C_i^F$ and $C_i^B$, respectively.

$$
G = \sigma(W^{(1)}H_i + W^{(2)}Hm_i^F + b)
$$

$$
C_i^F = G \odot H_i + (1 - G) \odot Hm_i^F
$$

Similarly, we obtain $C_i^B$. Both $C_i^F$ and $C_i^B$ are concatenated to render $C_i$;

- 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 $C_i$ as input and outputs a single vector for the entire sentence.

$$
F(C_i) = W^{(1)^T}\sigma(W^{(2)}C_i + b) + b
$$

$$
h(Uti) = F(C_i) \odot C_i
$$

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’, ‘Start-Service’, 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Booking</th>
<th>Cable</th>
<th>ATIS</th>
<th>SNIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Size</td>
<td>9351</td>
<td>1856</td>
<td>4978</td>
<td>13784</td>
</tr>
<tr>
<td>Val Size</td>
<td>4691</td>
<td>1814</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Test Size</td>
<td>6727</td>
<td>1836</td>
<td>893</td>
<td>700</td>
</tr>
<tr>
<td>#Intents</td>
<td>19</td>
<td>21</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>#SL</td>
<td>5</td>
<td>26</td>
<td>127</td>
<td>41</td>
</tr>
<tr>
<td>#DA</td>
<td>4</td>
<td>4</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>#ConvLen</td>
<td>4.25</td>
<td>4.68</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

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

**References**

