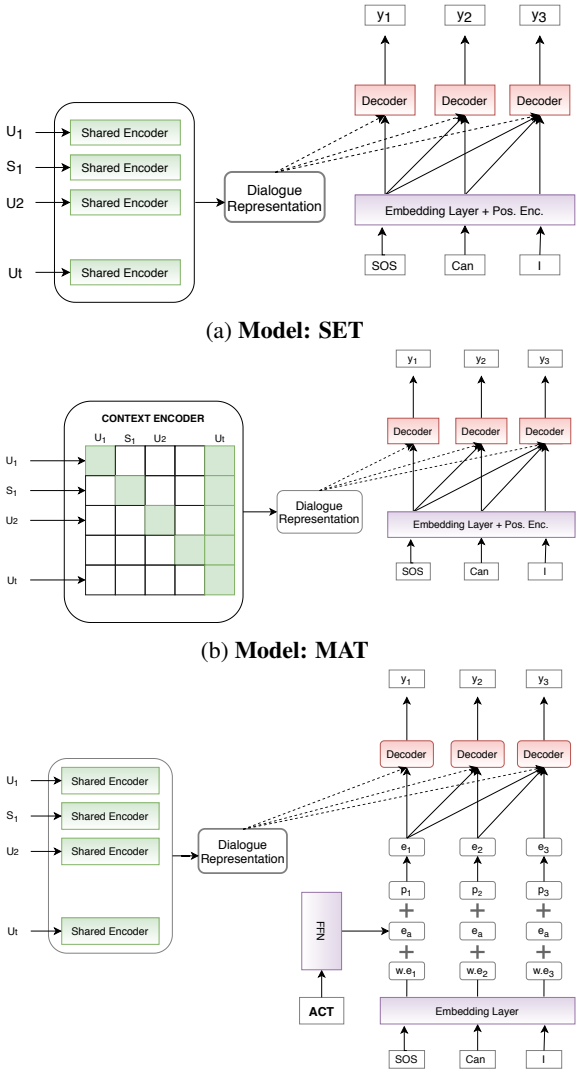
Supplementary Material for "Hierarchical Transformer for Task Oriented Dialog Systems"

Bishal Santra* bsantraigi[†]
Potnuru Anusha* anusha.sparkx[†]
Pawan Goyal pawang[‡]

Computer Science and Engineering Dept.
Indian Institute of Technology Kharagpur
Kharagpur, W.B., India
{†}@gmail.com, {‡}@cse.iitkgp.ac.in



(a) **Model: SET**

(b) **Model: MAT**

(c) **Model: SET++**

Figure 1: Model Variants. Shared Encoder always apply self-attention within utterance bounds. Masked Transformer applies self-attention among all utterances as depicted by the attention mask.

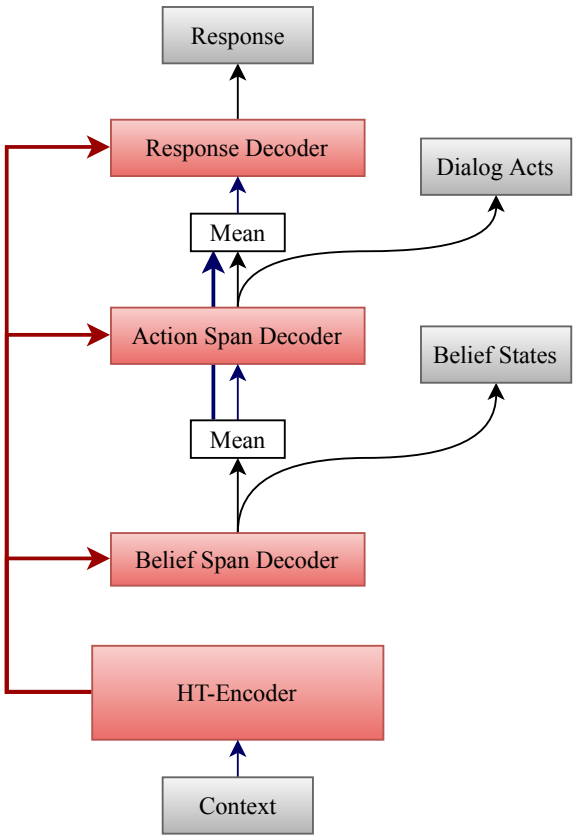


Figure 2: Block diagram of the HIER-Joint Architecture. Red links are cross attention from decoder to HT-Encoder. Blue links denote the mean token embeddings from predictions of the previous block.

A Hyperparameter Bounds

```
{
    'nhead': [2, 8]
    'embedding_perhead': [25, 40],
    'nhid_perhead': [10, 40],
    'nlayers_e1': [2, 6],
    'nlayers_e2': [2, 6],
    'nlayers_d': [2, 6],
    'dropout': [0.05, 0.8]
}
```

*Equal Contributions