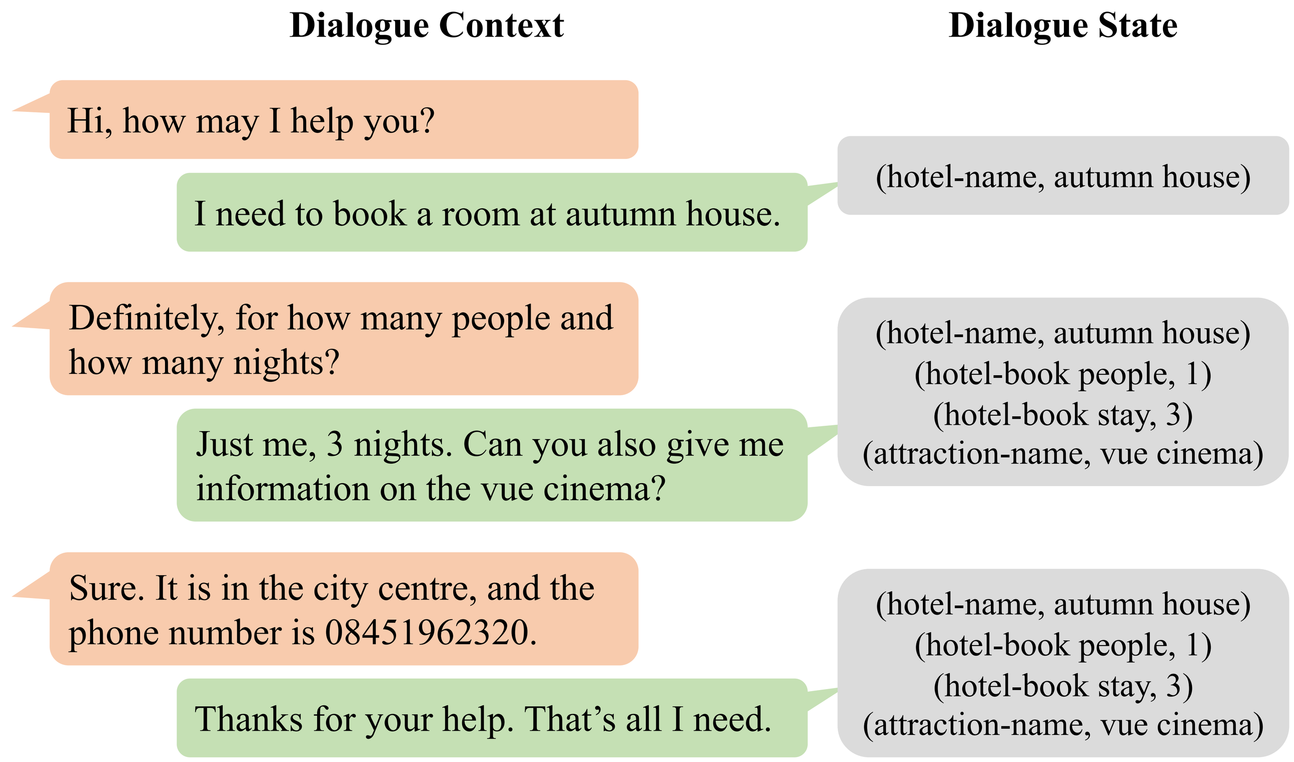
ASSIST: Towards Label Noise-Robust Dialogue State Tracking

Fanghua Ye, Yue Feng and Emine Yilmaz

University College London, UK

Introduction & Motivation

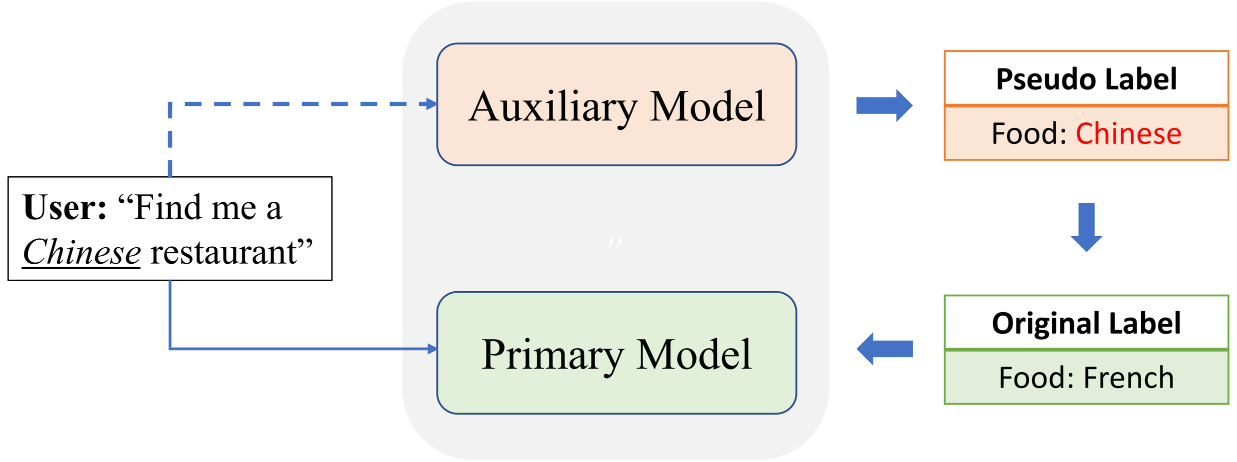
- The dialogue state tracker is an essential component of task-oriented dialogue systems. It aims to keep track of users' intentions at each turn of the conversation



- Dialogue state annotations are error-prone. Without taking noisy annotations into consideration, existing models can only achieve sub-optimal performance
- It is costly and labor-intensive to collect large-scale high-quality dialogue datasets

Methodology

- We propose a general framework ASSIST to robustly train dialogue state tracking models from noisy labels
- We introduce an auxiliary model, which is trained on a small clean dataset, to generate pseudo labels for each sample in the noisy training set



- We linearly combine the pseudo labels and vanilla labels (their one-hot vector representations) by a parameter α

$$\boldsymbol{V}_{combined} = \alpha \boldsymbol{V}_{pseudo} + (1-\alpha)\boldsymbol{V}_{vanilla} \quad (1)$$

- The cross entropy loss objective based on the combined labels can be decomposed into two parts as below

$$\mathcal{L}_{combined} = \alpha \mathcal{L}_{pseudo} + (1-\alpha)\mathcal{L}_{vanilla} \quad (2)$$

Theoretical Analysis

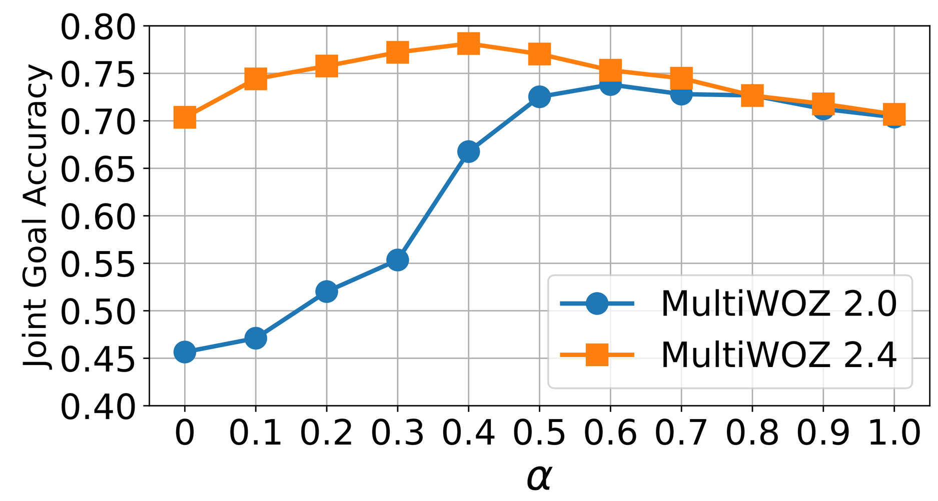
- We define the approximation error of any noisy labels $\boldsymbol{V}_{noisy}$ to the unknown clean labels $\boldsymbol{V}_{clean}$ using mean squared error (MSE)

$$Y_{\boldsymbol{V}_{noisy}} = \frac{1}{|\mathcal{D}_n||\mathcal{S}|} \sum_{\mathcal{X}_t \in \mathcal{D}_n} \sum_{S \in \mathcal{S}} E_{\mathcal{D}_c}[\|\boldsymbol{V}_{noisy} - \boldsymbol{V}_{clean}\|_2^2] \quad (3)$$

- It can be shown that the optimal approximation error with respect to the combined labels $\boldsymbol{V}_{combined}$ is smaller than that of the vanilla labels $\boldsymbol{V}_{vanilla}$ and pseudo labels $\boldsymbol{V}_{pseudo}$, i.e.,

$$\min_{\alpha} Y_{\boldsymbol{V}_{combined}} < \min\{Y_{\boldsymbol{V}_{pseudo}}, Y_{\boldsymbol{V}_{vanilla}}\} \quad (4)$$

- $Y_{\boldsymbol{V}_{combined}}$ is a concave function of α



Experimental Results

- All primary models achieve the best performance when both the vanilla labels and pseudo labels are used for training

Primary Models	Labels		MultiWOZ 2.0			MultiWOZ 2.4		
	Vanilla	Pseudo	Joint Goal(%)	Joint Turn(%)	Slot(%)	Joint Goal(%)	Joint Turn(%)	Slot(%)
SOM-DST	✓	✗	45.14	77.86	96.71	66.78	87.81	98.38
	✗	✓	67.06	87.95	98.47	68.69	88.41	98.55
	✓	✓	**70.83**	**89.14**	**98.61**	**75.19**	**91.02**	**98.84**
STAR	✓	✗	48.30	78.91	97.10	73.62	90.45	98.85
	✗	✓	70.66	85.93	98.67	71.01	86.31	98.69
	✓	✓	**74.12**	**88.93**	**98.86**	**79.41**	**91.86**	**99.14**
AUX-DST	✓	✗	45.66	78.76	96.95	70.37	89.31	98.67
	✗	✓	70.39	86.28	98.67	70.68	86.82	98.68
	✓	✓	**73.82**	**88.29**	**98.84**	**78.14**	**91.03**	**99.07**

- Directly combining the noisy training set with the small clean dataset can also lead to better results, however, the performance improvement is lower than our proposed approach

Training Settings	Joint Goal (%) MultiWOZ 2.0	Joint Goal (%) MultiWOZ 2.4
Noisy Train	45.66	71.80
Noisy Train + Small Clean	50.75	76.89
Noisy Train + Pseudo Labels	73.82	78.47
Noisy Train + Small Clean + Pseudo Labels	74.96	78.92

Paper

- Most slots have lower error rates with the help of the pseudo labels