Affordable On-line Dialogue Policy Learning

Cheng Chang*, **Runzhe Yang***, **Lu Chen, Xiang Zhou and Kai Yu**
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Eng.
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China
{cheng.chang,yang_runzhe,chenlusz,owenzx,kai.yu}@sjtu.edu.cn

A Figure 1: Example of Successful Dialogue

TASK: ask for *italian* restaurant in *north* area & request its *phone number*						
		Dialogue Turn	Score	Q^{turn}	Q^{succ}	FP
[1]	System	[SLU] welcomemsg()				
	User	[Top ASR] Italian food in the north part of town.	0.30	-4.54	27.44	False
[2]	System	[SLU] expl-conf(food="italian")				
	User	[Top ASR] Yes.	0.99	-2.24	29.09	False
[3]	System	[SLU] offer(name="caffe uno") inform(food="italian") inform(area="north")				
	User	[Top ASR] The phone number.	0.92	-2.00	28.27	False
[4]	System	[SLU] offer(name="caffe uno") inform(food="italian") inform(area="north") Inform(phone="01223314954")				
	User	[Top ASR] Does it serve danish italian food.	0.53	-2.41	28.20	False
[5]	System	[SLU] offer(name="caffe uno") inform(food="italian") inform(area="north")				
	User	[Top ASR] Goodbye.	0.58	0.05	27.42	False

Figure 1: An example of successful dialogue while training without teaching.

B Algorithm 1: the details of FPT heuristic

C Figure 2-4: On-line learning process under different teaching schemes

* Both authors contributed equally to this work.

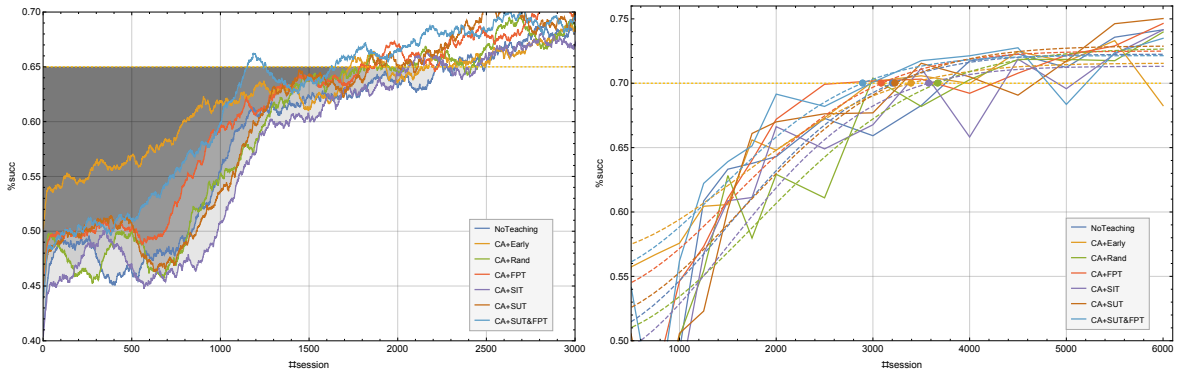

Figure 2: **Left**: On-line learning process under different teaching schemes (CA + different heuristics). **Right**: Test curves and fitted empirical learning curves of learning process with different teaching schemes (CA+different heuristic).

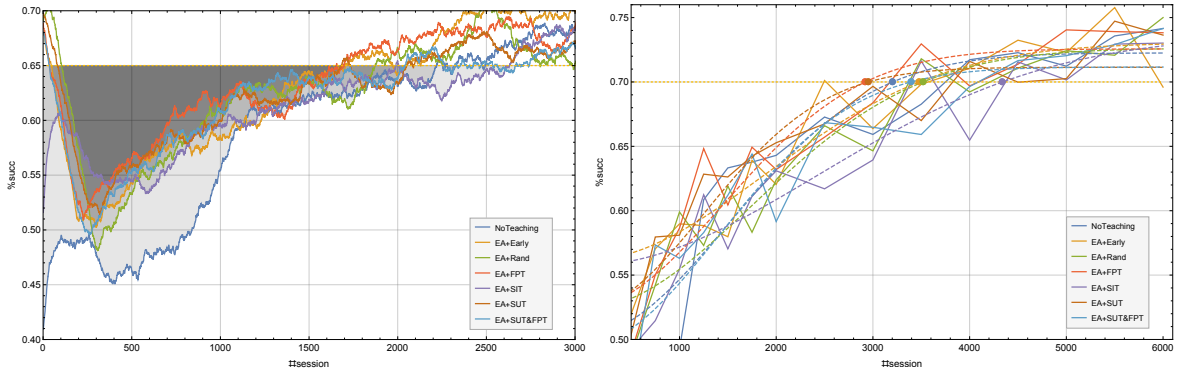

Figure 3: **Left**: On-line learning process under different teaching schemes (EA + different heuristics). **Right**: Test curves and fitted empirical learning curves of learning process with different teaching schemes (EA+different heuristic).

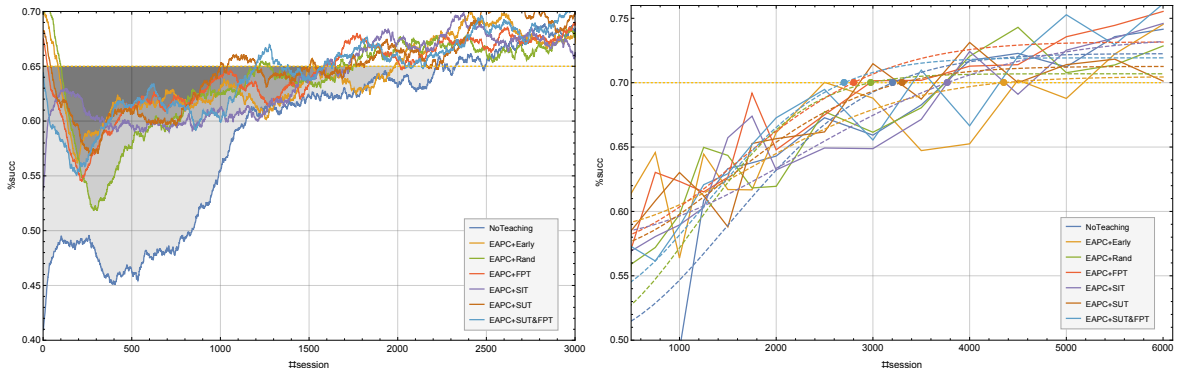

Figure 4: **Left**: On-line learning process under different teaching schemes (EAPC + different heuristics). **Right**: Test curves and fitted empirical learning curves of learning process with different teaching schemes (EAPC+different heuristic).

Algorithm 1 Failure Prognosis Based Teaching Heuristic

Initialize replay memory $\mathcal{D}$
Initialize MTL Q-Network, $Q^{\texttt{turn}}$ and $Q^{\texttt{succ}}$, with random weights
Initialize teaching budget c, ratio threshold α, sliding window size w
Initialize current teaching strategy (can be any strategy described in section 2.1)
Set teaching step $k \leftarrow 0$
for *episode* $= 1, N$ **do**
  Initialize dialogue state s_0
  for $t = 0, T$ **do**
    Select a_t randomly with probability ϵ, otherwise select: $argmax_a(Q^{\texttt{turn}}(s_t, a) + Q^{\texttt{succ}}(s_t, a))$
    if $k < c$ and *failure prognosis* is *true* according to equation 5 **then**
      Ask teacher for advice action a_t^{tea}
      $k \leftarrow k + 1$
    end if
    Update a_t by current teaching strategy
    Take action a_t, observe $r_t^{\texttt{turn}}$ and $r_t^{\texttt{succ}}$, transit to next state s_{t+1}
    Update $r_t^{\texttt{turn}}$, $r_t^{\texttt{succ}}$ according to current teaching strategy
    Store $(s_t, a_t, r_t^{\texttt{turn}}, r_t^{\texttt{succ}}, s_{t+1})$ in $\mathcal{D}$
    Sample minibatch of transitions $e \leftarrow (s_j, a_j, r_j^{\texttt{turn}}, r_j^{\texttt{succ}}, s_{j+1})$ from $\mathcal{D}$
    Update $Q_e^{\texttt{turn}}$ and $Q_e^{\texttt{succ}}$ according to equation 4, with respect to corresponding parameters
    Optimize $(Q_e^{\texttt{turn}} - Q^{\texttt{turn}}(s, a; \theta^{\texttt{turn}}))^2$ and $(Q_e^{\texttt{succ}} - Q^{\texttt{succ}}(s, a; \theta^{\texttt{succ}}))^2$ simultaneously under MTL structure, using gradient descent.
  end for
end for

D Figure 5: On-line learning process with sparse user feedback

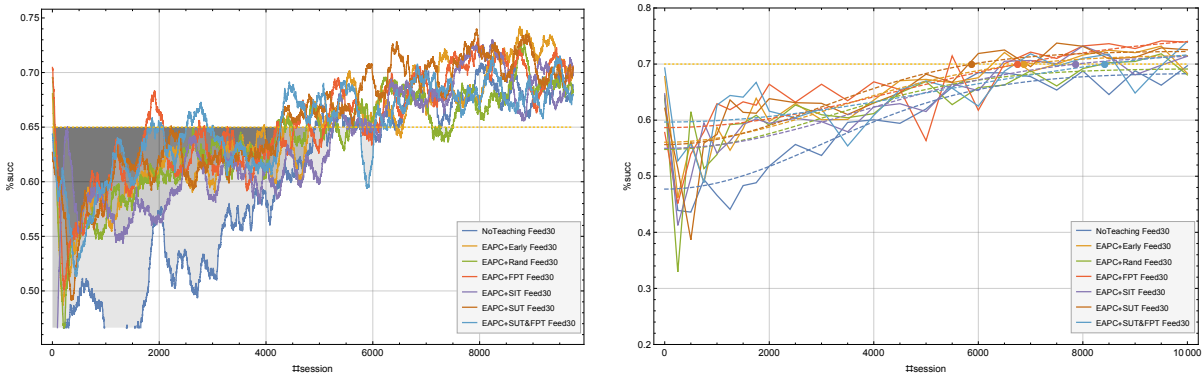

Figure 5: **Left**: On-line learning process under different teaching schemes (EAPC + different heuristics). **Right**:Test curves and fitted empirical learning curves of learning process with different teaching schemes (EAPC+different heuristic). User feedback rate is 30%.