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

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

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 heuristic-s). **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 heuristic-s). **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 heuristic-s). **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

1: Initialize replay memory $D$
2: Initialize MTL Q-Network, $Q_{\text{turn}}$ and $Q_{\text{succ}}$, with random weights
3: Initialize teaching budget $c$, ratio threshold $\alpha$, sliding window size $w$
4: Initialize current teaching strategy (can be any strategy described in section 2.1)
5: Set teaching step $k \leftarrow 0$
6: for episode $= 1, N$ do
7: Initialize dialogue state $s_0$
8: for $t = 0, T$ do
9: Select $a_t$ randomly with probability $\epsilon$, otherwise select:
10: \( \text{argmax}_a(Q_{\text{turn}}(s_t, a) + Q_{\text{succ}}(s_t, a)) \)
11: if $k < c$ and failure prognosis is true according to equation 5 then
12: Ask teacher for advice action $a_t^{tea}$
13: \( k \leftarrow k + 1 \)
14: end if
15: Update $a_t$ by current teaching strategy
16: Take action $a_t$, observe $r_t^{\text{turn}}$ and $r_t^{\text{succ}}$, transit to next state $s_{t+1}$
17: Update $r_t^{\text{turn}}, r_t^{\text{succ}}$ according to current teaching strategy
18: Store \((s_t, a_t, r_t^{\text{turn}}, r_t^{\text{succ}}, s_{t+1})\) in $D$
19: Sample minibatch of transitions $e \leftarrow (s_j, a_j, r_j^{\text{turn}}, r_j^{\text{succ}}, s_{j+1})$ from $D$
20: Update $Q_e^{\text{turn}}$ and $Q_e^{\text{succ}}$ according to equation 4, with respect to corresponding parameters
21: Optimize \((Q_e^{\text{turn}} - Q^{\text{turn}}(s, a; \theta^{\text{turn}}))^2\) and \((Q_e^{\text{succ}} - Q^{\text{succ}}(s, a; \theta^{\text{succ}}))^2\) simultaneously under MTL structure, using gradient descent.
22: end for
23: end for

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

![Figure 5: 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%.]