User study: Multi-dimensional, domain-adapted dialogue policy performs equally to in-domain one

User Evaluation of a Multi-Dimensional Statistical Dialogue System
Simon Keizer, Ondřej Dušek, Xingkun Liu & Verena Rieser

Summary
- First complete system with a multi-dimensional dialogue manager
- User evaluation via crowdsourcing
- novel web-based voice setup
- statistical equivalence tests

Data & code for download at: https://bitbucket.org/skeizer/madrigal/

Multidimensional Dialogue Managers
- POMDP
- multi-agent reinforcement learning
- separate agents/actions per dimension

System Variants
source domain: hotels, target: restaurants

All trained in target domain:
one-dim: 1 dimension, upper baseline multi-dim: 3 dims, trained from scratch

Task in target, FB + Soc transferred:
trans-fixed: 3 dims, FB + Soc fixed
trans-adapt: 3 dims, FB + Soc fine-tuned

Testing in simulation
- near-equall performance
- no negative transfer

Multi-dimensionality in Dialogue
- Utterances have multiple functions (dimensions) in a conversation
- some dimensions are domain-independent
- We use 3 dimensions:
  - Task
  - Feedback
  - Social
  - Feedback & Social are domain-independent

User: Hi, I need a Thai restaurant in the city centre
Social: greet, Task: inform
System: Okay, let me see... Feedback: positive, Time-management: pausing
System: Bangkok City is a Thai restaurant, it is in the city centre
Feedback: inform, Task: inform

Crowdsourced User Evaluation
- In-browser (with Google Web Speech)
- Generated tasks:
  "You want to find a restaurant near Castle Galleries, with cheap prices. You want to know its name, phone, address, postcode."
- Subjective questionnaire:
  - SubjSucc: found all information (Y/N)
  - VoiceInt: voice easy to understand (1-6)
  - Underst: system understood me (1-6)
  - AsExpect: behaved as expected (1-6)
  - WdUseAgain: would use it again (1-6)
- Objective measures:
  - NumTurns: average number of turns
  - WER: on a sample of 50% dialogues
  - EntProv: correct restaurant provided
  - ConstrConf: all constraints confirmed
  - InfoProv: requested information provided

Results (982 dialogues total)

<table>
<thead>
<tr>
<th>DM</th>
<th>SubjSucc</th>
<th>VoiceInt</th>
<th>Underst</th>
<th>AsExpect</th>
<th>WdUseAgain</th>
</tr>
</thead>
<tbody>
<tr>
<td>one-dim</td>
<td>87.3%</td>
<td>5.49</td>
<td>4.80</td>
<td>4.81</td>
<td>4.67</td>
</tr>
<tr>
<td>multi-dim</td>
<td>83.3%</td>
<td>5.37</td>
<td>4.68</td>
<td>4.68</td>
<td>4.59</td>
</tr>
<tr>
<td>trans-fixed</td>
<td>81.6%</td>
<td>5.47</td>
<td>4.66</td>
<td>4.64</td>
<td>4.63</td>
</tr>
<tr>
<td>trans-adapt</td>
<td>85.9%</td>
<td>5.38</td>
<td>4.67</td>
<td>4.64</td>
<td>4.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DM</th>
<th>NumTurns</th>
<th>WER</th>
<th>EntProv</th>
<th>ConstrConf</th>
<th>InfoProv</th>
</tr>
</thead>
<tbody>
<tr>
<td>one-dim</td>
<td>6.67</td>
<td>17.2%</td>
<td>72.2%</td>
<td>57.7%</td>
<td>45.7%</td>
</tr>
<tr>
<td>multi-dim</td>
<td>6.30</td>
<td>15.6%</td>
<td>68.4%</td>
<td>52.7%</td>
<td>44.7%</td>
</tr>
<tr>
<td>trans-fixed</td>
<td>6.57</td>
<td>15.4%</td>
<td>70.1%</td>
<td>53.1%</td>
<td>41.0%</td>
</tr>
<tr>
<td>trans-adapt</td>
<td>6.64</td>
<td>19.1%</td>
<td>72.2%</td>
<td>53.1%</td>
<td>46.6%</td>
</tr>
</tbody>
</table>

Statistical Equivalence
- no statistically significant differences among systems
- equivalence tests are a stronger proof of equivalence than not finding differences
- TOST – two one-sided tests
  \( H_0^{\hat{\Delta}}: \Delta \leq -\varepsilon, H_1^{\hat{\Delta}}: \Delta \geq +\varepsilon \quad (\varepsilon = 10\%) \)
  - reject both \( H_0^{\hat{\Delta}} \) & \( H_1^{\hat{\Delta}} \)
  - difference guaranteed \( < \varepsilon \)
- equivalences found for most system pairs & measures

Presented at SIGDIAL, Stockholm, 13 Sep 2019.
Supported by EPSRC project MaDrigAL (EP/N017536/1) and Charles University project PRIMUS/19/SCI/10.