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

HERIOT WATT UNIVERSITY

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/

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

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

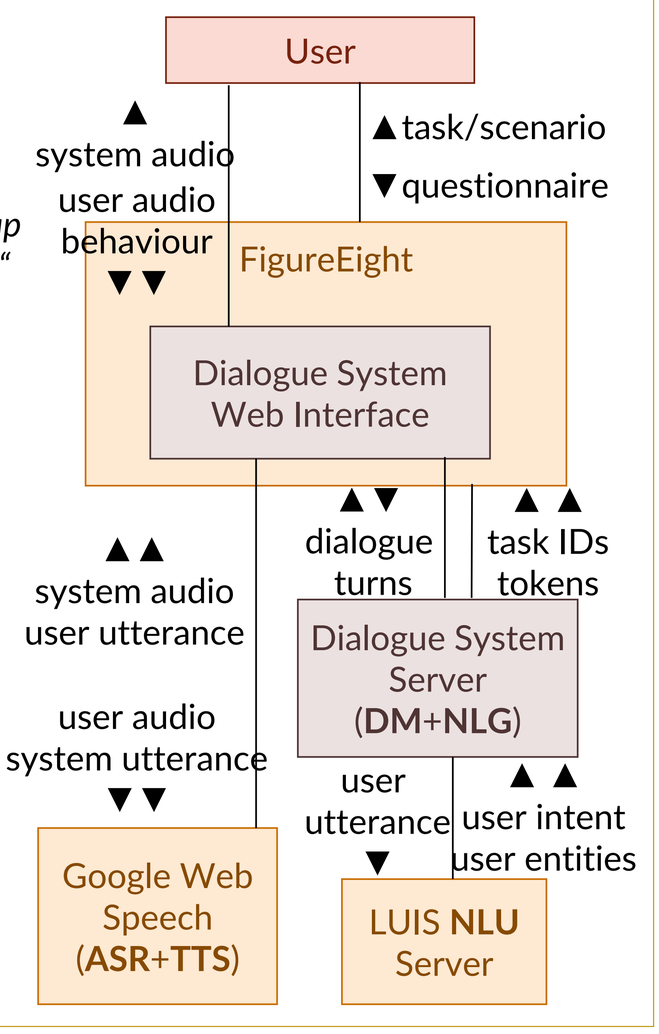
- near-equal performance
- no negative transfer

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)

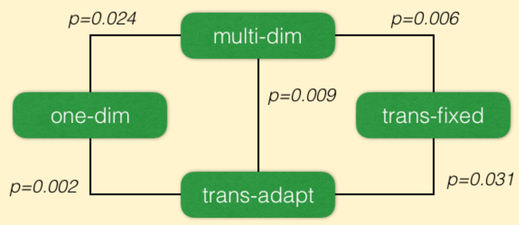
DM	SubjSucc	VoiceInt	Underst	AsExpect	WdUseAgain
one-dim	87.3%	5.49	4.80	4.81	4.67
multi-dim	83.3%	5.37	4.68	4.68	4.59
trans-fixed	81.6%	5.47	4.66	4.64	4.63
trans-adapt	85.9%	5.38	4.67	4.64	4.57

DM	NumTurns	WER	EntProv	ConstrConf	InfoProv
one-dim	6.67	17.2%	72.2%	57.7%	45.7%
multi-dim	6.30	15.6%	68.4%	52.7%	44.7%
trans-fixed	6.57	15.4%	70.1%	53.1%	41.0%
trans-adapt	6.64	19.1%	72.2%	53.1%	46.6%

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**
  - $H_0^{\text{lo}}: \Delta \leq -\epsilon$, $H_0^{\text{hi}}: \Delta \geq +\epsilon$ $(\epsilon = 10\%)$
  - reject both H_0^{lo} & H_0^{hi} → **difference guaranteed $< \epsilon$**
- equivalences found for most system pairs & measures

SubjSucc



EPSRC Engineering and Physical Sciences Research Council

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