A Human-Centered Hierarchical Framework for Dialogue System Construction and Evaluation

Salvatore Giorgi[1], Farhan Ahmed[2], Lyle Ungar[1], H. Andrew Schwartz[2]
[1]University of Pennsylvania, [2]Stony Brook University

Introduction

Goal: human-like open-domain system

Dialog agents are **created** with human-like traits
- Empathy, personality, emotions

Dialog agents are also **evaluated** as humans
- Which speaker sounds human?
- Does this response make sense?

The current work
- Define a hierarchical framework for dialog system evaluation
- Propose psychologically-grounded and human-centered evaluation measures

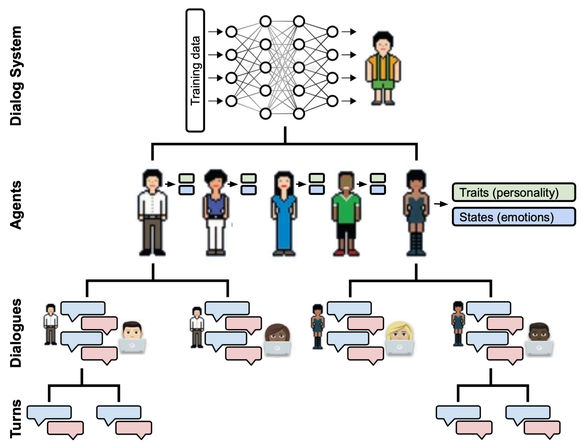
Dialogue System Technology Challenge 10

Task: Create metrics which
- Correlate with human judgements
- Are explainable

Data: Evaluated across 5 turn-level data sets,

Evaluation: metrics must correlate with human evaluations
- *Appropriateness*
  - The response is appropriate given the preceding dialogue.
- *Content*
  - How much information is provided in the response.
- *Grammar*
  - The quality of the English grammar.
- *Relevance*
  - The response content is related to the preceding dialogue.

The Hierarchy



- Table 1 (turn-level data):
  - Emotional Entropy (state) performs best on turn-level data
  - Matching metrics outperform traits
- Table 2 (dialog-level data):
  - Trait level measures outperform others at the dialog level

Human-Centered Measures

We propose two general sets of human-like measures
- **States and Traits**
  - *States*: thoughts/behaviors in a specific place/time
  - *Traits*: generalize across situations, stable across time
- **Linguistic Matching**
  - unconscious matching tendencies in postures, facial expressions, pitch, pausing, length, and use of function words

Task Metrics		
Agreeableness[1]	Trait	Dialog system, agent
Empathy[2]	Trait	Dialog system, agent
Emotional Entropy[3]	State	Agent, dialog, turn
Linguistic Style Matching[4]	Matching	Agent, dialog, turn
Emotion Matching[3]	Matching	Agent, dialog, turn

Results

	JSALT	ESL	NCM	DSTC10-Topical				DSTC10-Persona				Avg.
	App.	App.	App	App.	Content	Grammar	Relevance	App.	Content	Grammar	Relevance	
AM	.01	.03	.04	.12	.02	.04	.16	.11	.01	.05	.14	.07
FM	.05	.34	.16	.17	.09	.18$^{\diamond}$	.24	.19	.15	.19	.22	.18
Deep AM-FM	.05	.32	.16	.18	.09	.17$^{\diamond\diamond}$	.26	.21	.14	.19	.24	.18
Agreeableness	-.03	.05	.04	.01	-.01	-.01	.00	.01	.00	.01	.01	.01
Style Matching	.05	-.11	-.04	.05	.17	.06	.06	.08	.13	.09	.10	.06
Emotional Entropy	-.02	.07	.09	.02	.25$^{\diamond}$	.10	.02	.14	.28	.21$^{\diamond\diamond\diamond}$	.12	.12
Empathy	-.02	.08	.03	.01	.03	-.01	.00	.00	-.03	-.01	.00	.01
Emotion Matching	.01	-.03	.08	.03	.05	.00	.07	.11	.13	.11	.13	.06

Table 1: Evaluation on the test data. The first three rows are baseline systems. Reported Spearman ρ for each human evaluation metric: Appropriateness (App.), Content, Grammar, and Relevance). $^{\diamond}$, $^{\diamond\diamond}$, and $^{\diamond\diamond\diamond}$ denote first, second, and third place in column-wise scoring results, respectively (with other teams' scores not included in the results).

	FED-Conversation	Persona-Chatlog
Deep AM-FM	.12	.08
Agreeableness	.27	.03
Style Matching	.07	.08
Emotional Entropy	-.07	.01
Empathy	.11	-.01
Emotion Matching	.03	-.01

Table 2: Evaluation on the two dialog level development data sets. Reported Spearman ρ.

Conclusions

- Trait-level measures worked best when evaluated at higher levels (dialogs)
- State and matching measures worked best on turn-level data
- Limited task data (evaluated at the turn-level only)
- Metrics not optimized to correlate with task metrics (grammar, etc.)

Ethics

- Privacy, toxic and offensive content, willingness to share research
- Imparting system with human qualities can be dangerous
- Alternatively, dialog systems may exhibit extremely limited variation in such traits
- Potential "Wall Street Journal effect": dialog system only converse like middle aged white men

References

1. Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.
2. Abdul-Mageed, M., Buffone, A., Peng, H., Eichstaedt, J., & Ungar, L. (2017, May). Recognizing pathogenic empathy in social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 448-451).
3. Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301-326.
4. Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science*, *22*(1), 39-44.