Mini Lecture:
Are we conversational yet?

by Kat McNeill & Alejandrina G.R.

November 12th, 2020

CS 294S/W

Activity #1

Activity #1: Partial Dialogue Instructions

Given the first turn of a dialogue, or several turns, enter in the chat what you would say or ask next.

Dialogue #1: Partial Simple

Given the first turn of a dialogue, enter in the chat what you would say next.

Almond: Hello, how can I help you?

User (You): [Enter in the chat what you would ask Almond]

Dialogue #2: Partial Broken

Here is a broken partial dialogue, how do you expect a “good” assistant will converse with you in the next turn:

User (You): Andy, Play me a song by Lady Gaga

Almond: OK. (And it starts playing the song “Radio Ga Ga” by Queens)

User (You): [Enter in the Zoom chat what you would ask/say next]

Activity #2

Activity #2: (Broken) Chain Dialogue Instructions

Let's play a quick game:

1. Given the first turn of a dialogue, *someone* volunteer and say out loud what you would say next (based on *common* knowledge).

2. Then, call on someone else in the class to continue the conversation based on the last thing or question said. (You cannot call someone previously called).

3. Iterate on (1) and (2)

Questions to ponder: Where will the conversation go? How many different possibilities are there?

(Broken) Chain Dialogue

Starting question: Hello, how can I help you?

Volunteer 1: [Suggest an answer and call on someone else to ask the next question].

Volunteer 2: [Ask the next question and call on someone to answer it].

Volunteer 3: [Suggest an answer and call on someone else to ask the next question].

And so on…

Activities, Takeaways:

- Human-to-human conversation won’t be as constrained as bot-to-human conversation
- More naturalistic variation in human-to-human conversations
- Very hard to model all states needed for how a user might respond
- Hard to predict where the conversation might go
- Lots of edge cases

Related Work

Chorus: A Crowd-Powered Conversational Assistant

Chorus: A Crowd-Powered Conversational Assistant

Walter S. Lasecki[1], Rachel Wesley[1], Jeffrey Nichols[2], Anand Kulkarni[3], James F. Allen[1], and Jeffrey P. Bigham[1,4]

Computer Science, ROC HCI[1]
University of Rochester
{wlasecki,james}@cs.rochester.edu
{rwesley2}@u.rochester.edu

MobileWorks, Inc.[3]
anand@mobileworks.com

USER Group[2]
IBM Research - Almaden
jwnichols@us.ibm.com

Human-Computer Interaction Institute[4]
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT

Despite decades of research attempting to establish conversational interaction between humans and computers, the capabilities of automated conversational systems are still limited. In this paper, we introduce Chorus, a crowd-powered conversational assistant. When using Chorus, end users converse continuously with what appears to be a single conversational partner. Behind the scenes, Chorus leverages multiple crowd workers to propose and vote on responses. A shared memory space helps the dynamic crowd workforce maintain consistency, and a game-theoretic incentive mechanism helps to balance their efforts between proposing and voting. Studies with 12 end users and 100 crowd workers demonstrate that Chorus can provide accurate, topical responses, answering nearly 93% of user queries appropriately, and staying on-topic in over 95% of responses. We also observed that Chorus has advantages over pairing an end user with a single crowd worker and end users completing their own tasks in terms of speed, quality, and breadth of assistance. Chorus demonstrates a new future in which conversational assistants are made usable in the real world by combining human and machine intelligence, and may enable a useful new way of interacting with the crowds powering other systems.

human-computer interaction since the early days of computing. However, the complexity of human language has made robust two-way conversation with software agents a consistent challenge [1]. Existing dialogue-based software systems generally rely on a fixed input vocabulary or restricted phrasings, have a limited memory of past interactions, and use a fixed output vocabulary. In contrast, real-world conversations between human partners can contain context-dependent terms or phrasing, require memory stretching back over the conversation and past history of interactions and shared experiences, require common sense knowledge about the world or events, or facts, and contain meaningful incomplete and partial statements. Conversational assistants like Siri were greeted with great excitement when first released, but still possess limited capabilities beyond a finite set of pre-defined set of tasks because they cannot truly *understand* the user [8, 9].

While individual humans have no difficulty in maintaining natural-language conversation, it is often infeasible, unscalable, or expensive to hire a human to act as a conversational partner for long periods of time or to support large numbers of conversational partners. In recent years, *crowd computing* has become a popular method to scalably solve problems that

Chorus: A Crowd-Powered Conversational Assistant

Problem

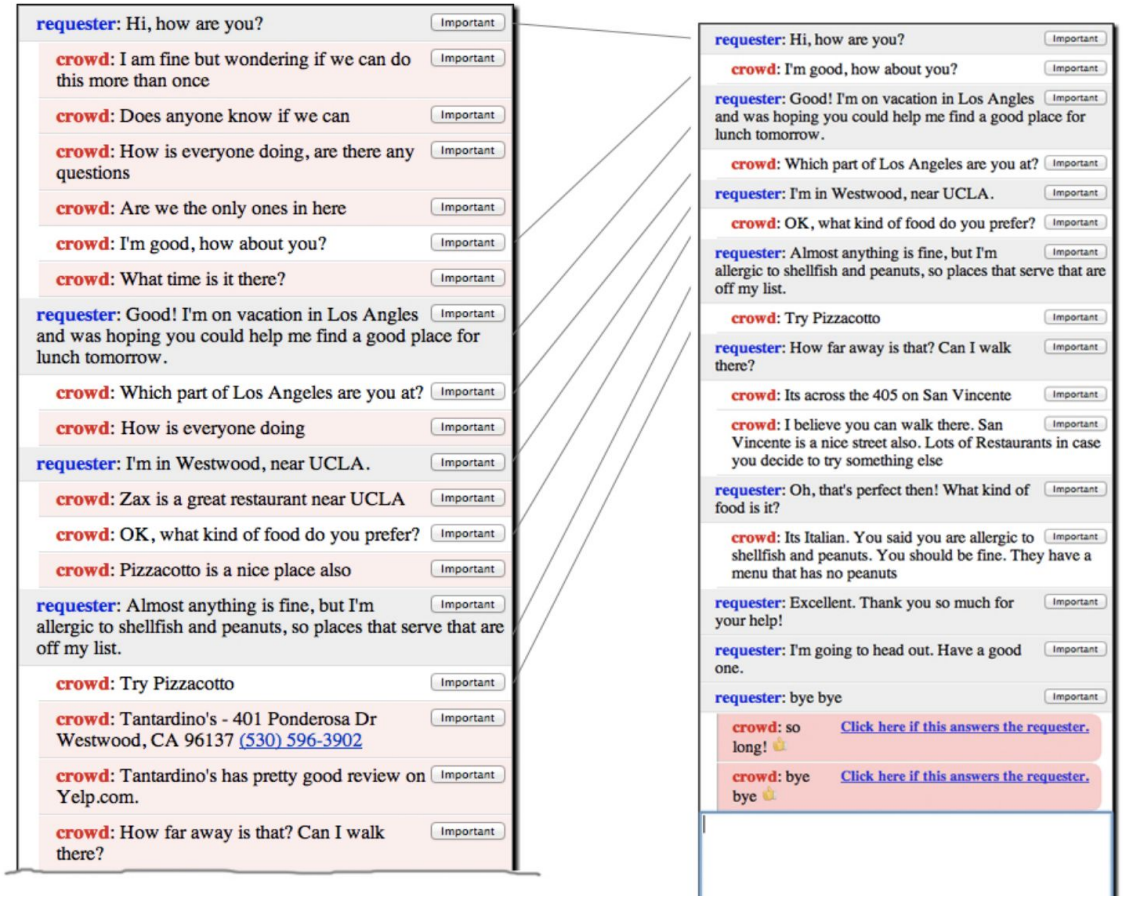
Existing dialogue-based software systems generally rely on a fixed input vocabulary or restricted phrasings, have a limited memory of past interactions, and use a fixed output vocabulary (very unlike human conversations).

Context of Problem

Real-world conversations between human partners can contain context-dependent terms or phrasing, require memory stretching back over the conversation and past history of interactions and shared experiences, require common sense knowledge about the world or events, or facts, and contain meaningful incomplete and partial statements.

Solution

Crowdsource answers from many workers and present the best answer to the user. Combines human and machine intelligence. Based on collaborative reasoning, dynamic scoring, and a curated memory system.


requester: Hi, how are you? Important
crowd: I am fine but wondering if we can do this more than once Important
crowd: Does anyone know if we can Important
crowd: How is everyone doing, are there any questions Important
crowd: Are we the only ones in here Important
crowd: I'm good, how about you? Important
crowd: What time is it there? Important
requester: Good! I'm on vacation in Los Angles and was hoping you could help me find a good place for lunch tomorrow. Important
crowd: Which part of Los Angeles are you at? Important
crowd: How is everyone doing Important
requester: I'm in Westwood, near UCLA. Important
crowd: Zax is a great restaurant near UCLA Important
crowd: OK, what kind of food do you prefer? Important
crowd: Pizzacotto is a nice place also Important
requester: Almost anything is fine, but I'm allergic to shellfish and peanuts, so places that serve that are off my list. Important
crowd: Try Pizzacotto Important
crowd: Tantardino's - 401 Ponderosa Dr Westwood, CA 96137 (530) 596-3902 Important
crowd: Tantardino's has pretty good review on Yelp.com. Important
crowd: How far away is that? Can I walk there? Important
requester: Hi, how are you? Important
crowd: I'm good, how about you? Important
requester: Good! I'm on vacation in Los Angles and was hoping you could help me find a good place for lunch tomorrow. Important
crowd: Which part of Los Angeles are you at? Important
requester: I'm in Westwood, near UCLA. Important
crowd: OK, what kind of food do you prefer? Important
requester: Almost anything is fine, but I'm allergic to shellfish and peanuts, so places that serve that are off my list. Important
crowd: Try Pizzacotto Important
requester: How far away is that? Can I walk there? Important
crowd: Its across the 405 on San Vincente Important
crowd: I believe you can walk there. San Vincente is a nice street also. Lots of Restaurants in case you decide to try something else Important
requester: Oh, that's perfect then! What kind of food is it? Important
crowd: Its Italian. You said you are allergic to shellfish and peanuts. You should be fine. They have a menu that has no peanuts Important
requester: Excellent. Thank you so much for your help! Important
requester: I'm going to head out. Have a good one. Important
requester: bye bye Important
crowd: so long! Click here if this answers the requester.
crowd: bye bye Click here if this answers the requester.

	Total Lines	Accurate Responses	Errors Made	Clarifications Asked	Questions Asked	Answers Provided
Consistency #1	24	9	0	0	4	4
Consistency #2	55	50	1	0	7	6
Consistency #3	33	11	0	0	2	2

Figure 3. Results for the conversations with Chorus..

	Total Lines	Accurate Responses	Errors Made	Clarifications Asked	Questions Asked	Answers Provided	Memory Successes	Memory Failures
Memory #1	138	53	30	3	5	3	4	2
Memory #2	63	15	1	1	4	2	1	0
Memory #3	30	29	1	1	3	3	1	0
Memory #4	28	7	0	2	3	2	2	0

Figure 6. Results for the conversations with Chorus including memory.

Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time

Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time

Ting-Hao (Kenneth) Huang, Joseph Chee Chang, and Jeffrey P. Bigham
Language Technologies Institute and Human-Computer Interaction Institute
Carnegie Mellon University
{tinghaoh, josephcc, jbigham}@cs.cmu.edu

ABSTRACT
Crowd-powered conversational assistants have been shown to be more robust than automated systems, but do so at the cost of higher response latency and monetary costs. A promising direction is to combine the two approaches for high quality, low latency, and low cost solutions. In this paper, we introduce Evorus, a crowd-powered conversational assistant built to automate itself over time by *(i)* allowing new chatbots to be easily integrated to automate more scenarios, *(ii)* reusing prior crowd answers, and *(iii)* learning to automatically approve response candidates. Our 5-month-long deployment with 80 participants and 281 conversations shows that Evorus can automate itself without compromising conversation quality. Crowd-AI architectures have long been proposed as a way to reduce cost and latency for crowd-powered systems; Evorus demonstrates how automation can be introduced successfully in a deployed system. Its architecture allows future researchers to make further innovation on the underlying automated components in the context of a deployed open domain dialog system.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

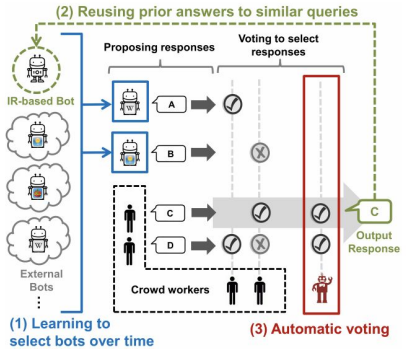


Figure 1. Evorus is a crowd-powered conversational assistant that automates itself over time by *(i)* learning to include responses from chatterbots and task-oriented dialog systems over time, *(ii)* reusing past responses, and *(iii)* gradually reducing the crowd's role in choosing high-quality responses by partially automating voting.

knowing what scenarios are supported, AI2 recently built an Alexa skill designed to help people find skills they could use

Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time

Problem

Fully automated virtual assistants aren't capable of having human-like conversations, and fully crowdsourced virtual assistants are slow and costly.

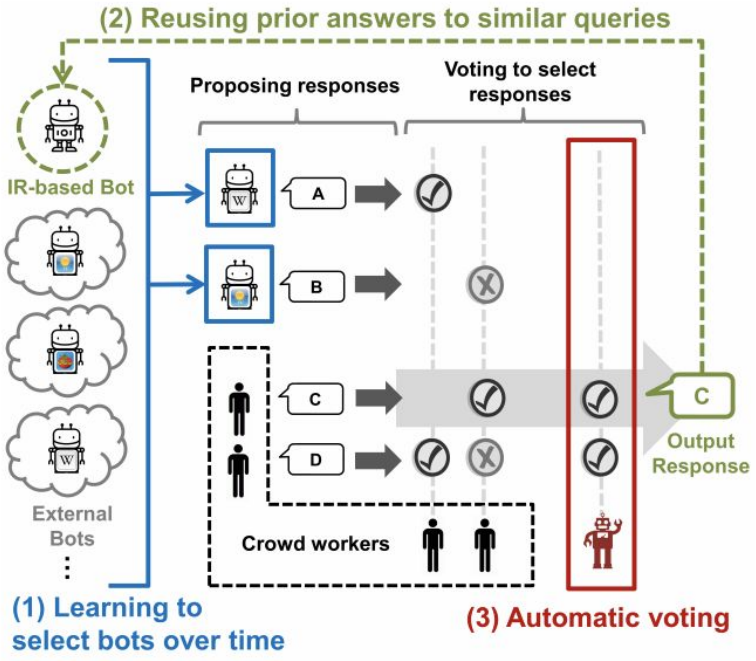
Context of Problem

Real-world conversations between human partners can contain context-dependent terms or phrasing, require memory stretching back over the conversation and past history of interactions and shared experiences, require common sense knowledge about the world or events, or facts, and contain meaningful incomplete and partial statements.

Solution

Evorus is a crowd-powered conversational assistant that automates itself over time by (i) learning to include responses from chatterbots and task-oriented dialog systems over time, (ii) reusing past responses, and (iii) gradually reducing the crowd's role in choosing high quality responses by partially automating voting.

Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time



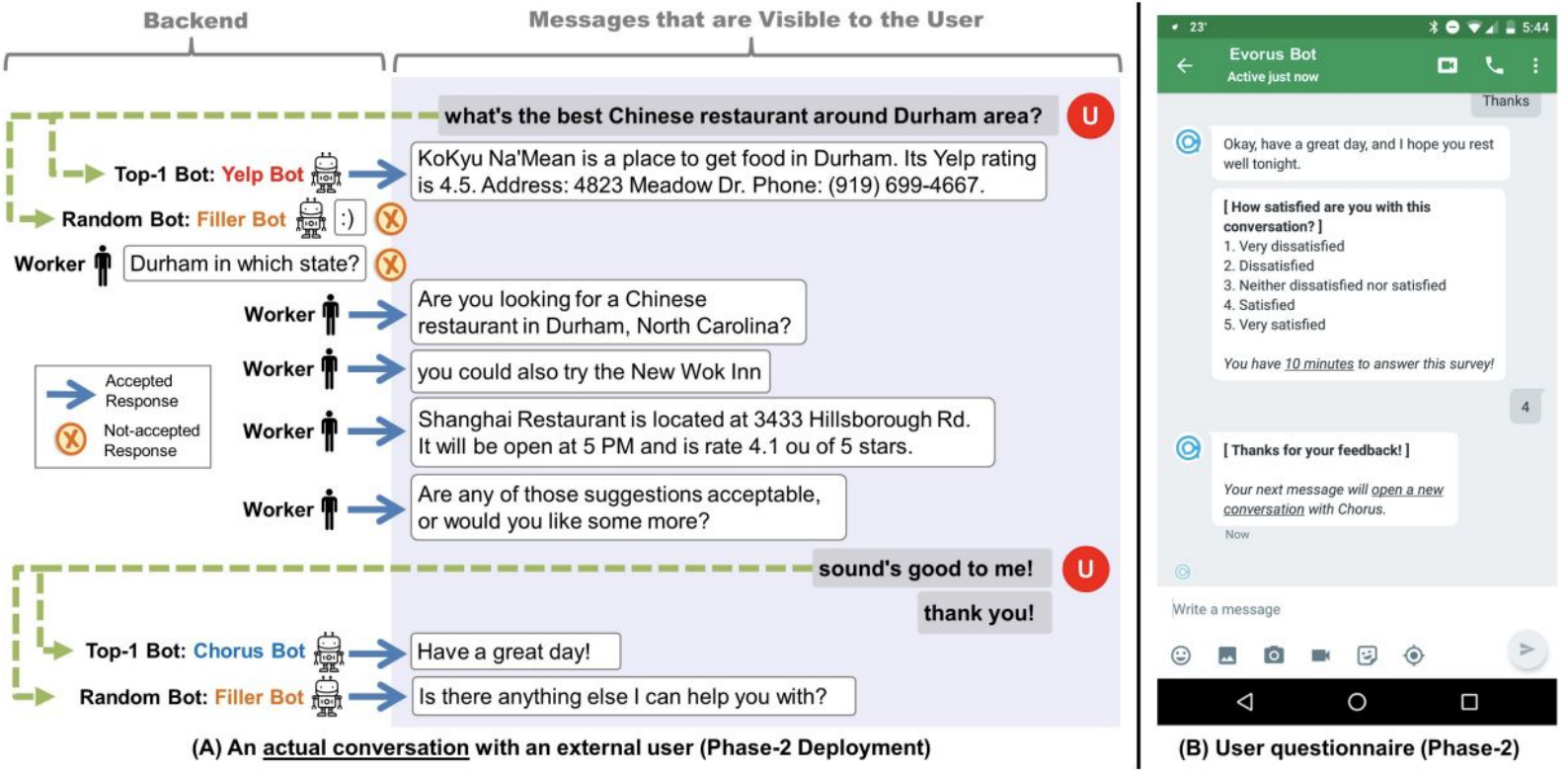


Figure 5. (A) An actual conversation of Evorus. Conversations in Evorus tend to combine multiple chatbots and workers together. (B) User questionnaire used in Phase-2 deployment. The average user satisfaction rating of automated and non-automated conversations had no significant difference.

Effects of Naturalistic Variation in Goal-Oriented Dialog

Effects of Naturalistic Variation in Goal-Oriented Dialog

Jatin Ganhotra
IBM Research
jatinganhotra@us.ibm.com

Robert Moore
IBM Research
rjmoore@us.ibm.com

Sachindra Joshi
IBM Research
jsachind@in.ibm.com

Kahini Wadhawan
IBM Research
kawadhaw@in.ibm.com

Abstract

Existing benchmarks used to evaluate the performance of end-to-end neural dialog systems lack a key component: natural variation present in human conversations. Most datasets are constructed through crowdsourcing, where the crowd workers follow a fixed template of instructions while enacting the role of a user/agent. This results in straight-forward, somewhat routine, and mostly trouble-free conversations, as crowd workers do not think to represent the full range of actions that occur naturally with real users. In this work, we investigate the impact of naturalistic variation on two

Dialog 1 (D1)

U : do you know about weather?
A : yeah sure. You can ask me about weather for any location e.g. Compton, Manhattan, Cleveland, Inglewood etc. What would you like?

U : What is the weather going to be today and tomorrow
A : What city would you like to know the weather about?
U : I want to know if it will drizzle in Durham
A : It is not predicted to drizzle any today or tomorrow in Durham

Figure 1: Sample dialog from SMD dataset between the User (**U**) and In-Car Assistant (**A**). The naturalistic variation added is shown in the box (in green color).

cise, and correct utterances. These goal-oriented

Effects of Naturalistic Variation in Goal-Oriented Dialog

Problem

Benchmarks to evaluate neural models for goal-oriented dialogs lack the natural variation present in human conversations.

Context of Problem

- Crowd workers follow a fixed template.
- Dataset contains straight-forward utterances
- Failure to simulate full range naturalistic variation

Solution

- Show the problem of unnatural data
- Show impact of natural variation in state-of-the-art dialog systems
- Release improved testbeds for datasets used in goal-oriented dialogue research

Effects of Naturalistic Variation in Goal-Oriented Dialog

To evaluate the impact of natural variation, two multi-turn goal-oriented datasets were used:

1) **bAbI dialogue task**
   5 simulated tasks in the restaurant domain.
2) **Stanford Multi-Domain dataset (SMD)**
   3 distinct domains (calendar scheduling, weather info, point-of-interest navigation)

The *Natural Conversation Framework* (NCF) was used to design conversational agents that emulate natural conversation.

Effects of Naturalistic Variation in Goal-Oriented Dialog

NCF patterns types:

1. Conversational activities (denoted **A**)
2. Conversation management:
   Sequence level (denoted **B**) & Conversation level (denoted **C**)

(A) Open Request Screening

(B) Misunderstanding Report

(C) Capability Expansion

Effects of Naturalistic Variation in Goal-Oriented Dialog

(A) Open Request Screening example:

Dialog 1 (D1)

U : do you know about weather?
A : yeah sure. You can ask me about weather for any location e.g. Compton, Manhattan, Cleveland, Inglewood etc. What would you like?

U : What is the weather going to be today and tomorrow
A : What city would you like to know the weather about?
U : I want to know if it will drizzle in Durham
A : It is not predicted to drizzle any today or tomorrow in Durham

(B) Misunderstanding Report example:

Dialog 2 (D2)

01 U : How's the weather?
02 A : It's getting better. Last week was very bad.
03 U : That's not what I mean
04 A : I'm sorry, please try again.

05 U : What is the weather like in the city
06 A : What city are you interested in?
07 U : What is the weather in Compton on Sunday?
08 A : It will be cloudy, with a low of 90F and a high of 100F in Compton on Sunday..

(C) Capability Expansion example: “Tell me more about restaurant recommendations.”

Effects of Naturalistic Variation in Goal-Oriented Dialog

After introducing 9 NCF patterns to the **bAbI** and **SMD** datasets, accuracy drops when evaluated using the state-of-the-art BossNet and GLMP models on both the original and the updated test set:

Model	BLEU	Ent. F1
Bossnet test	5.42	36.34
Bossnet test-updated	3.7	21.81
GLMP test	14.22	55.38
GLMP test-updated	4.73	21.05

Table 3: Performance of models on (original and updated) test sets for SMD dataset

Task/Model	Bossnet	GLMP
T5	97.82 (67.2)	99.20 (88.5)
T5-updated	90.4 (37.9)	87.24 (12.7)
T5-OOV	90.77 (12.1)	92.33 (21.8)
T5-OOV-updated	83.65 (7.0)	83.97 (5.9)

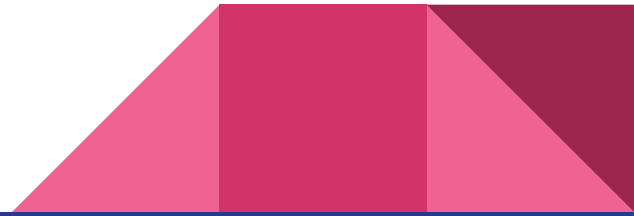
Table 4: Per-response (per-dialog) accuracy of models on (original and updated) test and test-OOV sets for bAbI dialog task-5 (denoted as T5 above)

Effects of Naturalistic Variation in Goal-Oriented Dialog

- The study demonstrates the dangers of using crowd-sourced data, without templates for the natural range of activities in conversation (such as NCF) to train end-to-end dialog systems.
- Naturalistic variation present during deployment affects: (1) model performance; and (2) results in lower than expected performance for a given dialog system.

Are we conversational yet?

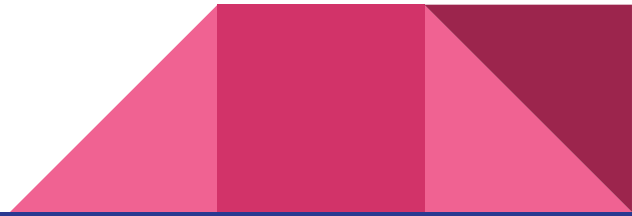
Short answer: not yet!



Are we conversational yet?

Short answer: not yet!

Long answer: There are solutions that are conversational, but that aren’t scalable or cost-effective (crowdsourcing answers, online forums, etc.). Fully automated solutions are not yet conversational.



Questions?

Sources

Ganhotra, Jatin, et al. "Effects of Naturalistic Variation in Goal-Oriented Dialog." *arXiv preprint arXiv:2010.02260* (2020).

Huang, Ting-Hao, Joseph Chee Chang, and Jeffrey P. Bigham. "Evorus: A crowd-powered conversational assistant built to automate itself over time." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018.

Lasecki, Walter S., et al. "Chorus: a crowd-powered conversational assistant." *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 2013.

Raghu, Dinesh, and Nikhil Gupta. "Disentangling language and knowledge in task-oriented dialogs." *arXiv preprint arXiv:1805.01216* (2018).