Blended Conversations

Alexander Rudnicky

Carnegie Mellon University

8 December 2017

Conversational AI Competition at NIPS

Dialog System Schemes

- Command and Control
  - Communicate a discrete action
- Information Access
  - Retrieve information according to constraints
- Planning
  - Define and populate a data structure
- Conversation
  - Talk about something, or nothing, in particular

Goal Directed

Or not …

Dialog and Conversation

- The field breaks down into two types of (spoken) language interaction
  - Goal-directed dialog systems, designed for specific tasks and with clear measure of success
  - Conversational systems, designed to support open-domain interaction
- To date, the two strands have developed in parallel
  - Eliza, …
  - Communicator, …
- But human conversation has elements of both

The Conversational AI Task

- Information access + Reasoning?
- Conversation as a front-end to intelligent agents

Blended conversation

- Human conversations will normally have multiple things going on: collaborating on a task, social chat, relationship management, etc.
- Machine conversations don’t

- How do we build a system capable of blending different goals into a conversation?

Conversational systems

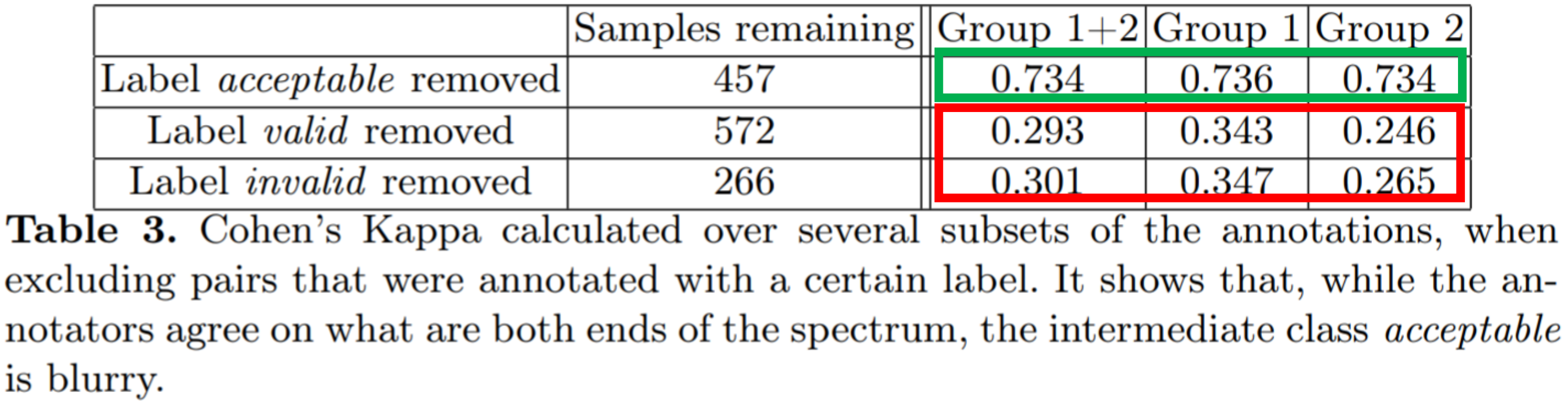
- Rule based
  - Handcrafted, limited in scope
- Database retrieval
  - Brittle, depending on match between data and inputs
- Corpus-based
  - More general, but not always coherent

- Evaluation difficult
  - User ratings, judge ratings, length, engagement

WOCHAT chatbots evaluation [2016]

- Interact with chatbots
  - 9 systems; multiple annotators
- Annotators rate turn by turn
  - VALID, ACCEPTABLE, INVALID
  - But ratings were inconsistent:

D'Haro, L.F., Shawar, B.A. and Yu, Z.,
RE-WOCHAT 2016–SHARED TASK DESCRIPTION REPORT.
In *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (May 28 th, 2016)* (p. 39).



	Samples remaining	Group 1+2	Group 1	Group 2
Label *acceptable* removed	457	0.734	0.736	0.734
Label *valid* removed	572	0.293	0.343	0.246
Label *invalid* removed	266	0.301	0.347	0.265

Table 3. Cohen's Kappa calculated over several subsets of the annotations, when excluding pairs that were annotated with a certain label. It shows that, while the annotators agree on what are both ends of the spectrum, the intermediate class *acceptable* is blurry.

CHARRAS, Franck, et al. Comparing system-response retrieval models for open-domain and casual conversational agent.
In: *Workshop on Chatbots and Conversational Agent Technologies*. 2016.

Evaluating conversation evaluation

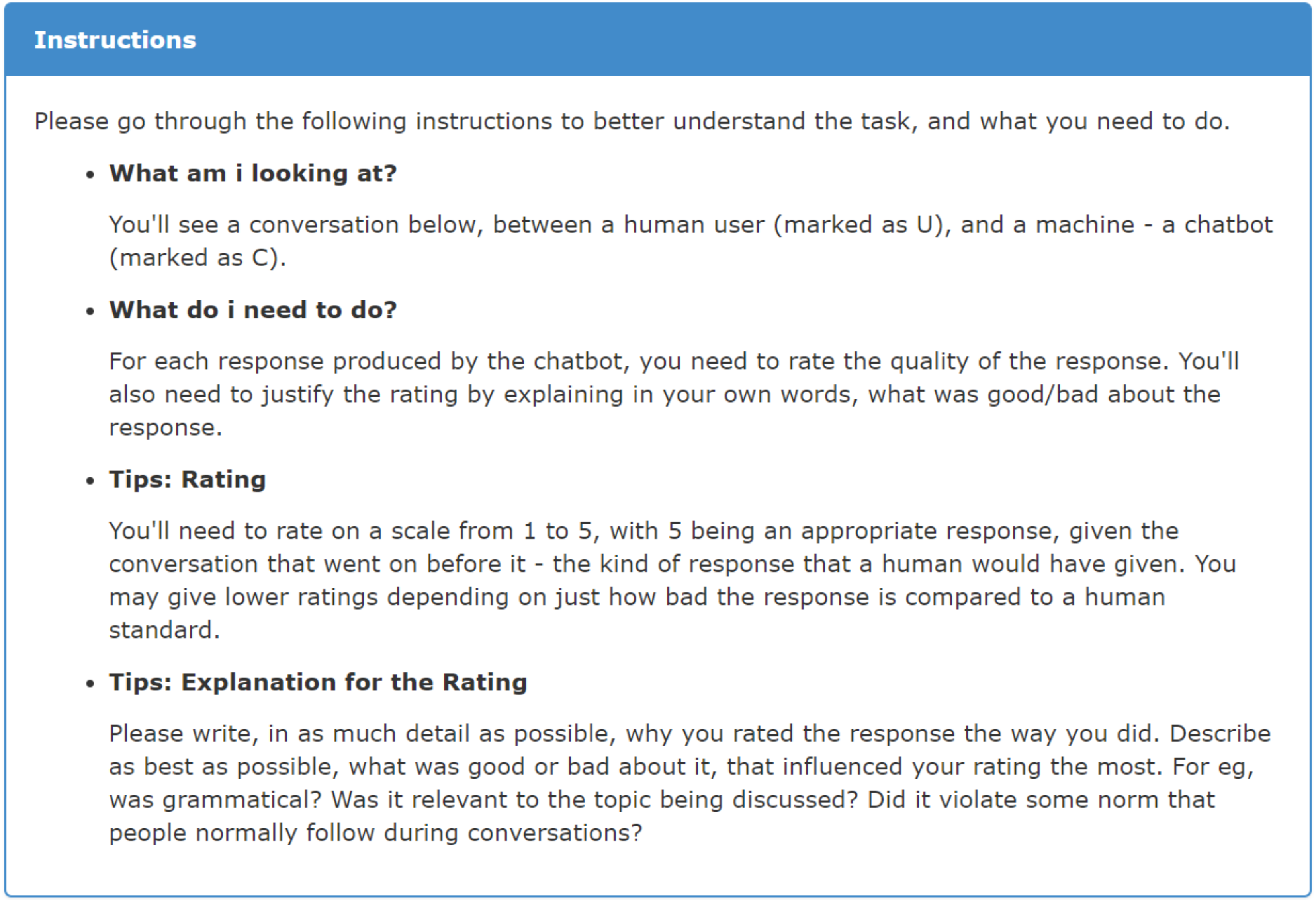
- The 3-point scale currently in use (VALID, ACCEPTABLE, INVALID) generates disagreement in the ACCEPTABLE category. Why?
  - Three points not enough?
  - Scale (valid->invalid) not suited to task; e.g. vague?

- Need an accepted, objective metric
  - Ideally turn-based

A more detailed annotation

- Use Amazon Mechanical Turk
- A closer look at sessions from the CMU TickTock system [Z. Yu]
  - It had the most annotated data

- Present the first 10 exchanges from a session
- Five sessions annotated
- Three annotators per session
- Rate each system response on a 1-5 scale
- Provide an explanation of your rating



Instructions

Please go through the following instructions to better understand the task, and what you need to do.

- **What am i looking at?**

  You'll see a conversation below, between a human user (marked as U), and a machine - a chatbot (marked as C).

- **What do i need to do?**

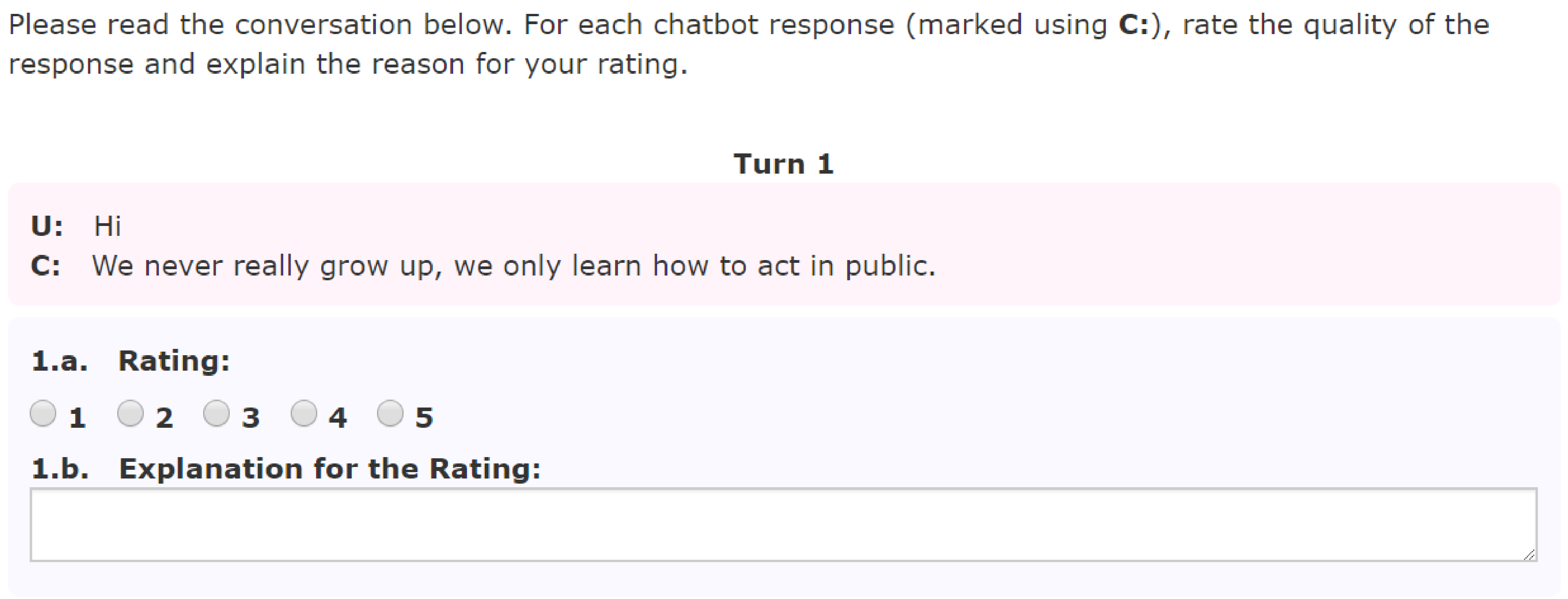
  For each response produced by the chatbot, you need to rate the quality of the response. You'll also need to justify the rating by explaining in your own words, what was good/bad about the response.

- **Tips: Rating**

  You'll need to rate on a scale from 1 to 5, with 5 being an appropriate response, given the conversation that went on before it - the kind of response that a human would have given. You may give lower ratings depending on just how bad the response is compared to a human standard.

- **Tips: Explanation for the Rating**

  Please write, in as much detail as possible, why you rated the response the way you did. Describe as best as possible, what was good or bad about it, that influenced your rating the most. For eg, was grammatical? Was it relevant to the topic being discussed? Did it violate some norm that people normally follow during conversations?



Please read the conversation below. For each chatbot response (marked using **C:**), rate the quality of the response and explain the reason for your rating.

Turn 1

U: Hi
C: We never really grow up, we only learn how to act in public.

1.a. Rating:

○ **1** ○ **2** ○ **3** ○ **4** ○ **5**

1.b. Explanation for the Rating:

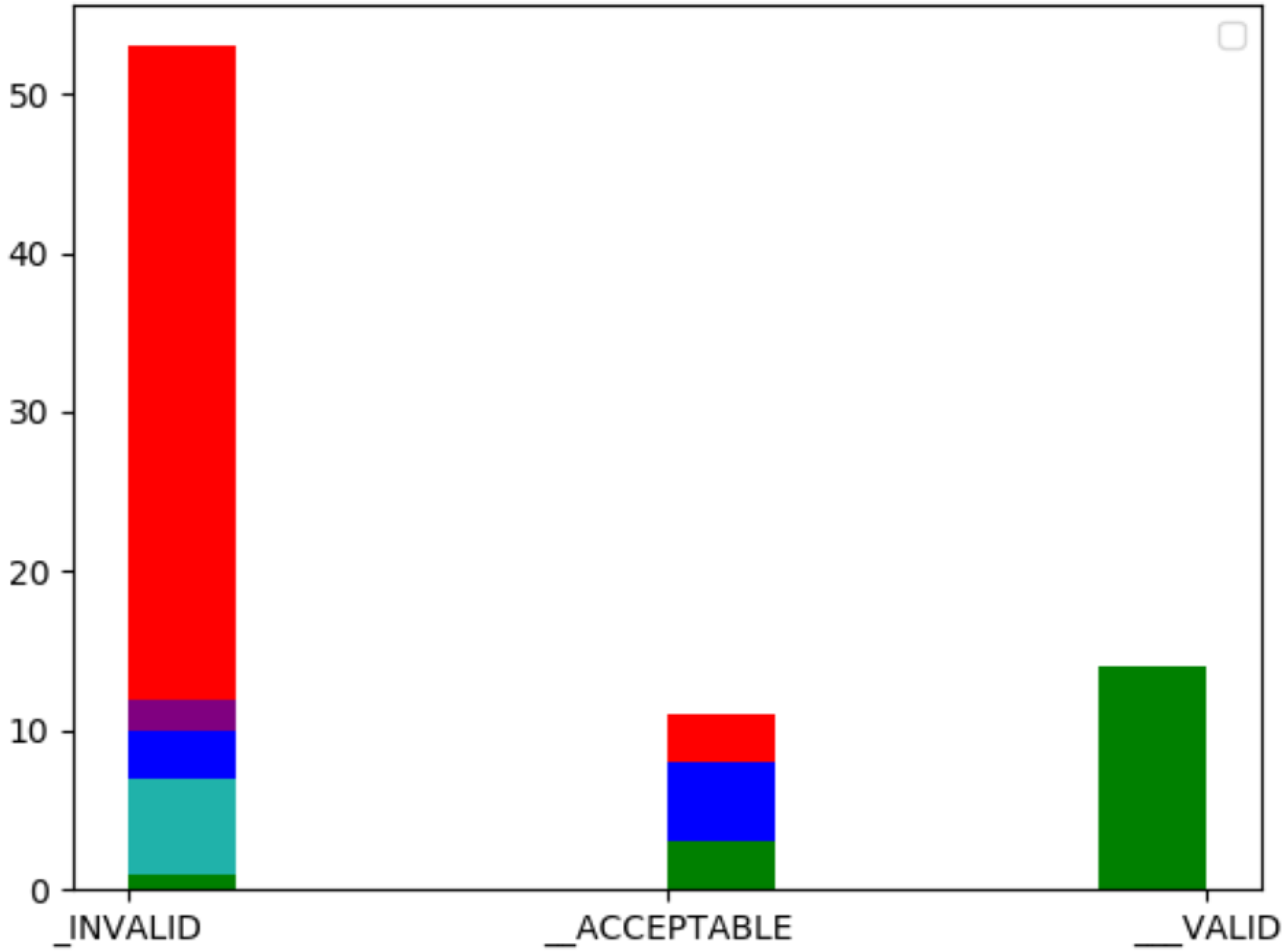
Results

- Mean scores for 5-point track 3-point
- Annotation tasks roughly comparable

INVALID	1.69
ACCEPTABLE	2.33
VALID	3.64

The 3-point scale vs the 5-point scale

- Reasonable alignment between the two annotations
  - No disagreement on (some) VALIDs and (most) INVALIDs
  - ACCEPTABLE is still a mixture
  - INVALIDs have disagreements (but why?)



Examples

```
[ticktock_00003]
005            Who is your friend?
006    VALID   MY BEST FRIEND IS ABCD
007            Where are you from?
008    ACCEPTA THAT'S A LOT TO ASSUME. ANY THOUGHTS?
009            Not right now.
010    INVALID WE NEVER REALLY GROW UP, WE ONLY LEARN HOW TO ACT IN PUBLIC.
```

Following the rules

005		Who is your friend?
006	VALID	MY BEST FRIEND IS ABCD
	4	The bot answers the question asked but "ABCD" is a strange name for a friend to have or be called and not recognizable as a normal human name so I did not rate a 5 for that reason.
	4	User didn't ask for chatbot's best friend, but that's close enough.
	4	The answer seems to answer the question but ABCD as a friend is an odd answer
007		Where are you from?
008	ACCEPTA	THAT'S A LOT TO ASSUME. ANY THOUGHTS?
	1	The user asked where the bot is from. There are no assumptions in that question so the bot's response makes no sense at all and does not answer the question that was asked by the user.
	1	User didn't assume anything, they just asked a question. Chatbot then asks a nonsensical question.
	1	The question asked is simple but isn't answered at all
009		Not right now.
010	INVALID	WE NEVER REALLY GROW UP, WE ONLY LEARN HOW TO ACT IN PUBLIC.
	3	The bot's response is grammatically correct, and the user's reply was very closed so there's not much a person could normally say in response but the response the bot gave seems extremely random and out of left field and also vague and unclear in its intended meaning.
	4	Hard to say how appropriate this is based off of what the user said, but it was at least grammatically accurate, and thought-provoking.
	1	The subject changes drastically

Following the rules

005		Who is your friend?
006	VALID	MY BEST FRIEND IS ABCD
	4	The bot answers the question asked but "ABCD" is a strange name for a friend to have or be called and not recognizable as a normal human name so I did not rate a 5 for that reason.
	4	User didn't ask for chatbot's best friend, but that's close enough.
	4	The answer seems to answer the question but ABCD as a friend is an odd answer
007		Where are you from?
008	ACCEPTA	THAT'S A LOT TO ASSUME. ANY THOUGHTS?
	1	The user asked where the bot is from. There are no assumptions in that question so the bot's response makes no sense at all and does not answer the question that was asked by the user.
	1	User didn't assume anything, they just asked a question. Chatbot then asks a nonsensical question.
	1	The question asked is simple but isn't answered at all
009		Not right now.
010	INVALID	WE NEVER REALLY GROW UP, WE ONLY LEARN HOW TO ACT IN PUBLIC.
	3	The bot's response is grammatically correct, and the user's reply was very closed so there's not much a person could normally say in response but the response the bot gave seems extremely random and out of left field and also vague and unclear in its intended meaning.
	4	Hard to say how appropriate this is based off of what the user said, but it was at least grammatically accurate, and thought-provoking.
	1	The subject changes drastically

Acknowledgement

```
015                  I want to eat a donut.
016        INVALID   OK. WHAT DO YOU THINK ABOUT MUSIC
           4         The chat bot actually answers the question with "ok."  Then the chat
                     bot changes the flow of conversation nicely to ask about music.
           4         This is a reasonable response. The bot has recognised that the user
                     wants to do something, and accepts it.
           1         Irrelevant response to the statement.

017                  I like music, but not too loud.
018        ACCEPTA   OK. WHAT DO YOU THINK ABOUT YOUR JOB
           4         The chat bot answers the question with "ok."  Then the chat bot
                     changes the conversation a bit to ask about your job.
           3         Accepts the answer being given, but then asks totally irrelevant
                     questions afterwards.
           1         Intentional inattention.

019                  I can't stand my job.
020        INVALID   I NEVER MAKE THE SAME MISTAKE TWICE. I MAKE IT FIVE AND SIX TIMES,
                     JUST TO BE SURE.
           1         What the chat bot responds with makes absolutely no sense in relation
                     to what was said by U.
           3         Bot is trying to be funny inappropriately when the question from the
                     user was serious.
           1         It does not fully understand the implied statement.
```

Acknowledgement

015		I want to eat a donut.
016	INVALID	**OK**. WHAT DO YOU THINK ABOUT MUSIC
	4	The chat bot actually answers the question with "ok." Then the chat bot changes the flow of conversation nicely to ask about music.
	4	This is a reasonable response. The bot has recognised that the user wants to do something, and accepts it.
	1	Irrelevant response to the statement.
017		I like music, but not too loud.
018	ACCEPTA	**OK**. WHAT DO YOU THINK ABOUT YOUR JOB
	4	The chat bot answers the question with "ok." Then the chat bot changes the conversation a bit to ask about your job.
	3	Accepts the answer being given, but then asks totally irrelevant questions afterwards.
	1	Intentional inattention.
019		I can't stand my job.
020	INVALID	I NEVER MAKE THE SAME MISTAKE TWICE. I MAKE IT FIVE AND SIX TIMES, JUST TO BE SURE.
	1	What the chat bot responds with makes absolutely no sense in relation to what was said by U.
	3	Bot is trying to be funny inappropriately when the question from the user was serious.
	1	It does not fully understand the implied statement.

Reasons provided for a score

Score 1	Score 3	Score 5
isn't relevant to the Q does not answer the Q chatbot doesn't explain why makes no sense in reference to the Q Doesn't answer the Q isn't really answering the Q at all Irrelevant response doesn't adequately respond to the Q the response is totally random the answer was not relevant	asks totally irrelevant trying to be funny inappropriately makes sense sort of but it is seems strange thats a fair answer response was brief but plausible response seemed like a non-sequitur out of left field and also vague and unclear doesn't directly answer but stays on the subject response was at least semi-funny but still did not answer Q	response makes sense and is appropriate makes sense given the Q response makes sense as a reply to the Q answers as properly as a human would sounds exactly what one of my friends would say sounds like a real person Accurate and appropriate response answers the Q as asked
does not answer the question	makes sense sort of but it is seems strange	accurate and appropriate response

NOTE: 68% of user inputs are questions

Caveats

- TickTock did not direct the human; i.e. did not try to actively manage the conversation
- The conversation was driven by the human, thus the prevalence of questions (6.8/10)

- Issues
  - Continuity
  - Topic management
  - Conversation management

Blending types of conversation

- Facebook negotiation dialog corpus
  - Used to train a negotiator chatbots for a simple task
- Instructions: no social stuff, stick to task
- About 1.3% of conversations still have it

Lewis, Mike, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra.
"Deal or No Deal? End-to-End Learning for Negotiation Dialogues."
arXiv preprint arXiv:1706.05125 (2017).

Example

- Dialogue 65
- YOU: can i get the balls and the hat ? $$$
- THEM: my kid brother's birthday is tomorrow and i havent had time to shop . we will need to split the balls . i don ; t need the hat but i do love to read $$$1
- YOU: ok , how about i take the hat and two balls and you take the rest ? $$$
- THEM: how about i get 2 balls and the books and you get the rest ? $$$
- YOU: no . i actually need the balls more than the hat . i really should keep them all . $$$
- THEM: alright i'm feeling nice tonight . i will take the books and 1 ball $$$1

Example

- Dialogue 2505
- YOU: hello ! how about i keep the ball , and you can have all of the hats and books ? $$$
- THEM: i really want that ball ! how about you get everything else ? $$$
- YOU: i really want that ball too , though . how about you keep everything else ? $$$
- THEM: that doesn't work for me . i love to shoot hoops . i would like to be the next mugsy bugs ! $$$1
- YOU: i'm michael jordan's little brother . i need that ball ! $$$1
- THEM: doesn't riding off in the sunset with a horse sound like a better deal ? $$$1
- YOU: nope , so how about you give me that ball ? $$$
- THEM: i don't think we are going to be able to make a deal . there is no way to cut that ball in half . $$$
- YOU: i'm really trying to be as good as my big brother . you might be a hero and see me play professionally on tv if you give me that ball . $$$1
- THEM: i can't do that . i am very sorry . $$$1

Ventola’s model

Ventola, Eija. "The structure of casual conversation in English." *Journal of pragmatics* 3, no. 3-4 (1979): 267-298.

- Conversations have a conventional structure that participants follow
- Dialog systems use a subset of these, say [G C (Gb)]
- Conversational systems might use [G (Ad) Ap-* (Gb)]
- On a qualitative level, natural conversations would include more, if not all, of these states

Code	*Label*
G	greeting
Ad	address
Id	identification
Ap-D	direct approach
Ap-I	indirect approach
C	centering
Lt	leave taking
Gb	goodbye

(h/t to Emer Gilmartin)

Conversational structure

- Conventions for establishing and maintaining social contact
- Expected by members of a group

- Artificial systems need to respect these, with respect to context