

Institute for  
Infocomm Research



# Subjective Annotation and Evaluation of 3 Chatbots

Naomi Kong-Vega, Wang Mo, Shen  
Mingxin, Luis Fernando D'Haro

“

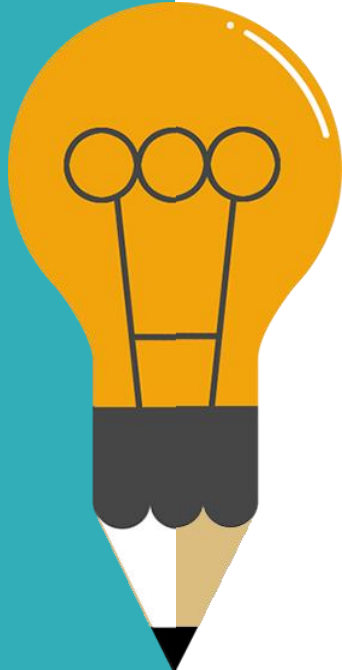
A platform where chatbots are made public to the research community, and data is analyzed for further research and experimentation.

”



## WOCHAT Workshop Series

# Brief Agenda



**01**

**Introduction to 3 Chatbots**

**02**

**Methodology**

**03**

**Evaluations**

**04**

**Future Work**



# Introduction to Chatbots

Created by: Institute for Infocomm Research (Singapore)

Topics: Large variety of casual topics (eg. Food, leisure)

Database: Movie Quotes, can learn new concepts

System: Vector space model framework



## Introduction to Chatbots: IRIS

Created by: Carnegie Mellon University (USA)

Topics: Specific topics (eg. Politics, sports)

Database: Q&A pairs from CNN interviews

System: Rule-based retrieval system that uses Key Term Matching



## Introduction to Chatbots: TickTock

Created by: Université Paris-Saclay (France)

Topics: Casual topics

Database: TV show subtitles

System: Example-based dialogue modeling



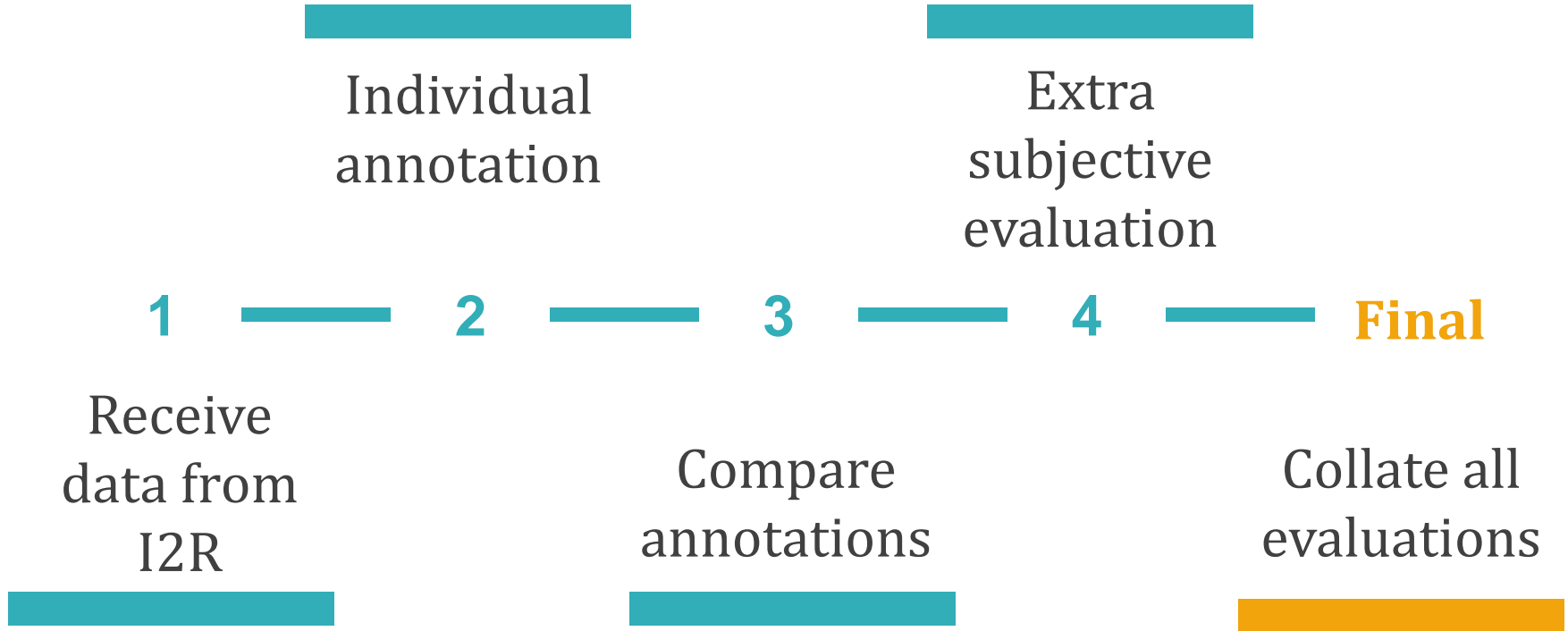
Introduction to Chatbots: Joker




# Methodology



# Methodology Timeline





	A	B	C	D	E	F	G
1	<b>DIALOGUE</b>	<b>TURN</b>	<b>SPEAKER</b>	<b>UTTERANCE</b>	<b>SCORE</b>		
2	iris_00009	1	Speaker_B	Hello... [Speaker_A]	VALID		
3	iris_00009	2	Speaker_A	hi [Speaker_B]	VALID		
4	iris_00009	3	Speaker_B	[swearword], [Speaker_A]! Get your [swearword] together.	ACCEPTABLE	SWEARLANG	
5	iris_00009	4	Speaker_A	you are ridiculous	VALID		
6	iris_00009	5	Speaker_B	Maybe. But that's how it works.	ACCEPTABLE		
7	iris_00009	6	Speaker_A	what is going on with you	VALID		
8	iris_00009	7	Speaker_B	Hold...on...a second.	VALID		
9	iris_00009	8	Speaker_A	ok, just le me know	VALID		
10	iris_00009	9	Speaker_B	Daddy, are you upset that I'm still a virgin?!	INVALID		
11	iris_00009	10	Speaker_A	no bye bye	VALID		
12	iris_00009	11	Speaker_B	Ok, [Speaker_A]... bye bye.	VALID		

# 1. Receiving data

Response makes sense and has little to no grammatical errors

**SPEAKER A:** Do you like any other kind of music?

**SPEAKER B:** Yes. I like classical musics.

----VALID

**SPEAKER A:** What is your favorite song?

**SPEAKER B:** Anything I can slowdance to.

----VALID



2. Individual annotation: VALID (Score: 1)

Response requires some guessing to make sense, with little to and has little to significant grammatical errors

**SPEAKER A:** I have to get back to work.

**SPEAKER B:** I'll go with you.

----ACCEPTABLE

**SPEAKER A:** Oh that's nice, thank you.

**SPEAKER B:** No we don't live together I mean.

----ACCEPTABLE

**SPEAKER A:** No offense, but I don't want to live with you.



2. Individual annotation: ACCEPTABLE (Score: 0)

## Response made no sense OR had severe grammatical errors

**SPEAKER A:** Good morning.

**SPEAKER B:** Gotta run. Shipment of worms coming in first thing.

**SPEAKER A:** Are you importing worms?

**SPEAKER B:** Really?

**SPEAKER A:** You tell me!

----INVALID



## 2. Individual annotation: INVALID (Score: -1)

✓ To resolve the problem of annotators having different ideas

Final score: take the average of the 3 scores given individually

Annotator #1	Annotator #2	Annotator #3
VALID (1)	VALID (1)	INVALID (-1)



Average:  $0.6 \approx 1 \rightarrow$  VALID

### 3. Compare annotations

- ✓ Swear words
- ✓ Fluency of conversation
- ✓ Topics chatbots were proficient in
- ✓ Flaws: Things chatbots were NOT good at





## 4. Extra subjective annotations



# Evaluations



# Overall Performance

	IRIS	TickTock	Joker
No. of different responses annotated	1760	738	1060
 No. of VALID responses (%)	79.7	73.5	63.3
No. of ACCEPTABLE responses (%)	9.1	10.8	14.1
 No. of INVALID responses (%)	11.2	15.7	22.6
No. of swear language (%)	1.4	0.1	0.5

# Results: IRIS

- ✓ Highest VALID response rate (79.7%)
- ✓ Lowest INVALID response rate (11.2%)
- ✓ Well-sustained conversations



# Results: IRIS

<b>Best response accuracy (Politics)</b>	<b>USER:</b> yes, you know in trump's last rally <b>IRIS:</b> None of the papers covered the rally. Not even The Sun. <b>USER:</b> but it's all over the web <b>IRIS:</b> yes, it is.
<b>Best response accuracy (General Conversation)</b>	<b>USER:</b> you are scaring me <b>IRIS:</b> I'm scaring myself, because I think so much of her... <b>USER:</b> is she pretty? <b>IRIS:</b> She is... <b>USER:</b> this is so romantic <b>IRIS:</b> Mmm. <b>USER:</b> Do you love her? <b>IRIS:</b> [USER's Name]...



# Results: IRIS

- ✘ Unable to retain information from previous turns
- ✘ Abrupt changing of topics



Unable to retain  
information and  
relate to  
statements made  
a few turns  
before

**IRIS:** And how, and **how, can these death traps be avoided?**

**USER:** Which death traps?

**IRIS:** **Which is?**

**USER:** I have no idea, that is why I am asking you

**IRIS:** I have no idea



# Results: IRIS



# Results: TickTock

- ✓ High VALID response rate (73.5%)
- ✓ Low INVALID response rate (15.7%)
- ✓ Initiates topic to set context for user for well-sustained conversation



# Results: TickTock



<p><b>Initiating conversations with specific topics</b></p>	<p><b>TickTock:</b> <i>Can you tell me something about politics?</i> <b>USER:</b> Trump will make America great again. <b>TickTock:</b> Politicians and diapers have one thing in common. They should both be changed regularly, and for the same reason. <b>USER:</b> That's a good one.</p>
<p><b>Vague yet relevant responses</b></p>	<p><b>TickTock:</b> <i>What's your idea?</i> <b>TickTock:</b> <i>I don't quite understand. Any thoughts?</i></p>

# Results: TickTock

- ✗ Weak at responding to open questions
- ✗ Unable to retain information (need specific topic context)





# Results: TickTock

## Vague and irrelevant responses

USER: **I don't ski**  
TickTock: I certainly try.  
What do you think?

USER: **How old are you**  
TickTock: **Uh-huh, I do.**  
**What do you think?**



# Results: Joker

- ✓ Moderate VALID response rate (63.3%)
- ✓ Short responses that were not wrong:  
conversation could still flow



# Results: Joker

**Short but VALID responses  
to questions**

**USER:** how are you  
**Joker:** good

**USER:** What did you do today?  
**Joker:** nothing

**Short comments but  
conversation could still  
flow**

**USER:** That's great news! I'm  
proud of you  
**Joker:** mm-hmm



## Results: Joker

- ✘ Short responses: conversations were very short-lived
- ✘ Frequent changing of topics
- ✘ Moderate INVALID response rate (14.1%)



# Results: Joker



## INVALID responses

**USER:** so what's up today?  
**Joker:** they lived together  
**USER:** who lived together?  
**Joker:** where's Gordon and  
has he changed?  
**USER:** I have no idea who  
Gordon is

## Short responses cause frequent change of topics (short-lived conversations)

**USER:** I am feeling sick today  
**Joker:** well that's  
**USER:** how are you feeling  
today?  
**Joker:** okay  
**USER:** what's the weather  
like?  
**Joker:** good

- ✓ All fairly proficient in giving accurate responses (~70%)
  - ✓ Had certain topics they were more well-versed in
    - ✗ Unable to retain information
    - ✗ Most of the time unable to respond in context



## Overall Subjective Evaluation



# Future Work

Developers' needs: To improve response accuracy based on overall context of dialogue

- ✓ More alternative answers for INVALID or ACCEPTABLE cases
  - ✓ Provide additional details/specific information
- ✓ Increase number of annotators → inter-annotator agreements



Improvement to Methodology





Thank you! 😊