

Institute for
Infocomm Research



Subjective Annotation and Evaluation of 3 Chatbots

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

“

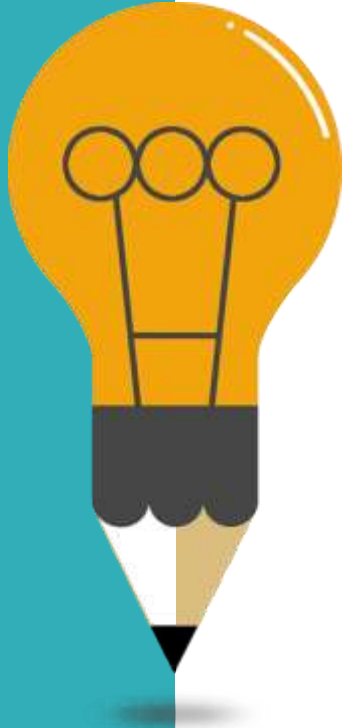
A platform where chatbots are made public to the research community, and data is analyzed for further research and experimentation.

”



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

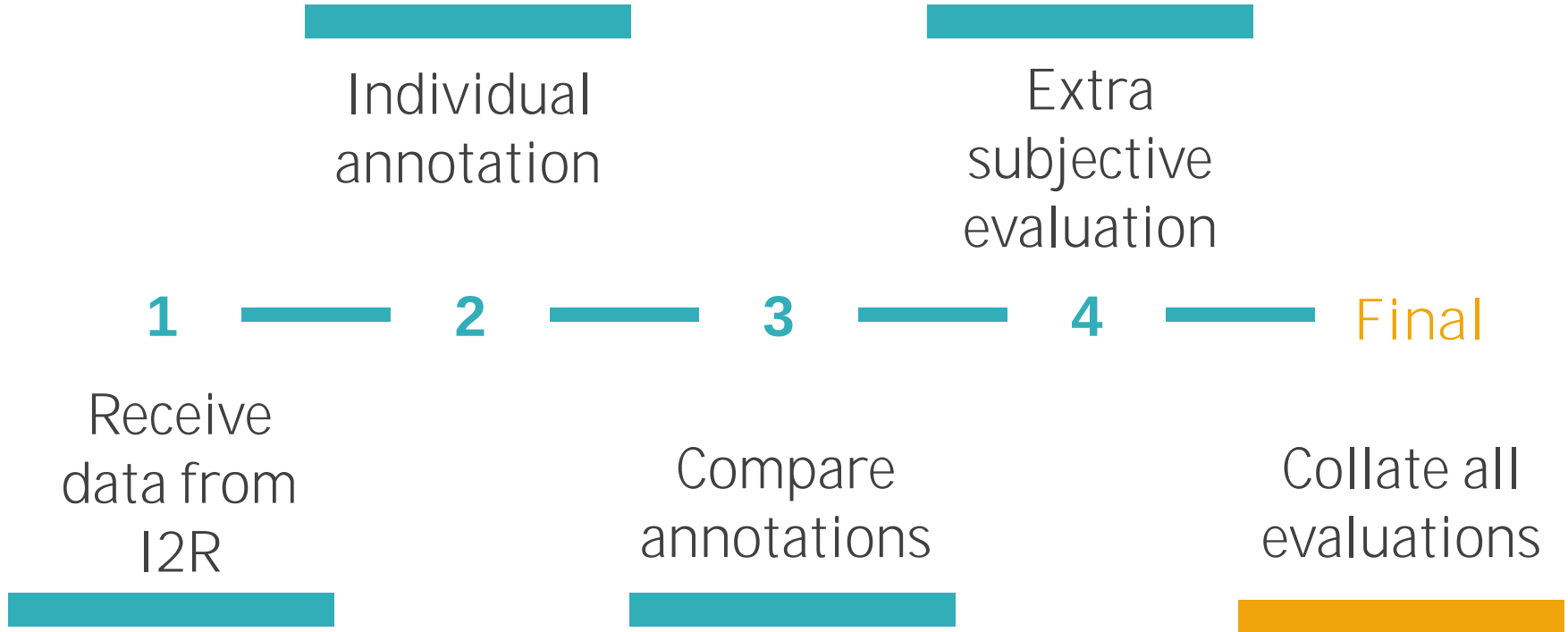


Introduction to Chatbots: Joker



Methodology

Methodology Timeline





	A	B	C	D	E	F	G
1	DIALOGUE	TURN	SPEAKER	UTTERANCE	SCORE		
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

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

IRIS: [USER'S Name]...



Results: TickTock



<p>Initiating conversations with specific topics</p>	<p>TickTock: Can you tell me something about politics?</p>
<p>Vague and irrelevant responses</p> <p>Weak VA responses to open questions</p> <p>Unable to retain information</p> <p>Low INVALID response rate (15.7%)</p> <p>specific topic context</p> <p>Initiates topic to set context for user for well-sustained conversation</p>	<p>USER: I don't ski, make America great again. try.</p> <p>What do you think?</p> <p>and diapers have one thing in common: How old are you both be changed regularly, and for the same person.</p> <p>Uhhuh, I do</p> <p>What do you think?</p> <p>USER: That's a good one.</p>
<p>Vague yet relevant responses</p>	<p>TickTock: What's your idea?</p> <p>TickTock: I don't quite understand. Any thoughts?</p>

Results: Joker



INVALID responses	USER: so what's up today? Joker: they lived together
Short but VALID responses <ul style="list-style-type: none"> Short responses: conversation were 33% Moderate VALID responses: 33% short lived 	USER: how are you USER: who lived together? Joker: where's Gordon and has he changed?
<ul style="list-style-type: none"> Short responses that were not too long Frequent changing of topics 	Joker: nothing Gordon is nothing USER: I have no idea who Joker: nothing
Short responses cause frequent change of topics	USER: I am feeling sick today USER: that's
Moderate INVALID response rate (14.1%) (short-lived conversations) Short comments but conversation could still flow	USER: That's great news! I'm proud of you today? USER: how are you feeling today? Joker: mm-hmm 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

- ✓ Improve usage of context
- ✓ Improve ability to retain information
- ✓ Reduce number of swear words to improve usability



Improvement to chatbots

- ✓ 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! ☐