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.
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
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: Joker

Created by: Université Paris-Saclay (France)

Topics: Casual topics

Database: TV show subtitles

System: Example-based dialogue modeling
Methodology
Methodology Timeline

1. Receive data from I2R
2. Individual annotation
3. Compare annotations
4. Extra subjective evaluation
   - Collate all evaluations

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

1. Receiving data
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
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)
SPEAKER A: Good morning.
SPEAKER B: Gotta run. Shipment of worms coming in first thing.
SPEAKER A: Are you importing worms?
SPEAKER B: Really? ----INVALID
SPEAKER A: You tell me!

2. Individual annotation: INVALID (Score: -1)
1. To resolve the problem of annotators having different ideas

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

<table>
<thead>
<tr>
<th>Annotator #1</th>
<th>Annotator #2</th>
<th>Annotator #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>VALID (1)</td>
<td>VALID (1)</td>
<td>INVALID (-1)</td>
</tr>
</tbody>
</table>

Average: $0.6 \approx 1 \rightarrow$ VALID

3. Compare annotations
4. Extra subjective annotations

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

<table>
<thead>
<tr>
<th></th>
<th>IRIS</th>
<th>TickTock</th>
<th>Joker</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of different responses annotated</td>
<td>1760</td>
<td>738</td>
<td>1060</td>
</tr>
<tr>
<td>No. of VALID responses (%)</td>
<td>79.7</td>
<td>73.5</td>
<td>63.3</td>
</tr>
<tr>
<td>No. of ACCEPTABLE responses (%)</td>
<td>9.1</td>
<td>10.8</td>
<td>14.1</td>
</tr>
<tr>
<td>No. of INVALID responses (%)</td>
<td>11.2</td>
<td>15.7</td>
<td>22.6</td>
</tr>
<tr>
<td>No. of swear language (%)</td>
<td>1.4</td>
<td>0.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Results:

- Highest VALID response rate (79.7%)
- Lowest INVALID response rate (11.2%)
- Well-sustained conversations
| Best response accuracy (Politics) | USER: *yes, you know in trump’s last rally*  
|                                | IRIS: *None of the papers covered the rally. Not even The Sun.*  
|                                | USER: *but it’s all over the web*  
|                                | IRIS: *yes, it is.*  

| Best response accuracy (General Conversation) | USER: *you are scaring me*  
|                                             | IRIS: *I’m scaring myself, because I think so much of her...*  
|                                             | USER: *is she pretty?*  
|                                             | IRIS: *She is...*  
|                                             | USER: *this is so romantic*  
|                                             | IRIS: *Mmm.*  
|                                             | USER: *Do you love her?*  
|                                             | IRIS: *[USER’s Name]*...  

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: TickTock

- High VALID response rate (73.5%)
- Low INVALID response rate (15.7%)
- Initiates topic to set context for user for well-sustained conversation
| Initiating conversations with specific topics | **TickTock**: Can you tell me something about politics?  
**USER**: Trump will make America great again.  
**TickTock**: Politicians and diapers have one thing in common. They should both be changed regularly, and for the same reason.  
**USER**: That’s a good one. |
|-----------------|---------------------------------------------------------------|
| Vague yet relevant responses | **TickTock**: What's your idea?  
**TickTock**: I don’t quite understand. Any thoughts? |
✖ Weak at responding to open questions
✖ Unable to retain information (need specific topic context)
| 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? |
✓ Moderate VALID response rate (63.3%)
✓ Short responses that were not wrong: conversation could still flow
| 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 |
Short responses: conversations were very short-lived
Frequent changing of topics
Moderate INVALID response rate (14.1%)
| 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 \(\rightarrow\) inter-annotator agreements

Improvement to Methodology
Thank you! 😊