Understanding and predicting user dissatisfaction in a neural generative chatbot

Abigail See, Christopher D. Manning
Typical dialogue evaluation setup

Neural generative models perform increasingly well in this environment
Real-life dialogue settings

Possibly spoken conversation + user own expectations own personality + noisy environment + latency constraints + intrinsic reward

Neural generative models are less well-explored in this environment.
Chirpy Cardinal 1.0

• An open-domain socialbot

• Won 2nd place in third Alexa Prize (2019-2020)

• Alexa customers speak to socialbots voluntarily

• Bots must be fast and manage the noisy home environment

For more details: stanfordnlp.github.io/chirpycardinal

Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations, Paranjape*, See* et al, Alexa Prize Proceedings, 2020
The Neural Chat component uses a neural generator to discuss the user's personal experiences and emotions.

Discussion starts with a handwritten starter question. Topics include: hobbies, food, everyday activities, family/friends.

Subsequent responses are generated by a GPT2-medium model finetuned on EmpatheticDialogues (Rashkin et al, 2019).

Decoding algorithm: top-p sampling

Towards empathetic open-domain conversation models: A new benchmark and dataset, Rashkin et al, 2019
# Types of user dissatisfaction

<table>
<thead>
<tr>
<th>Dissatisfaction type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarification</td>
<td><em>what do you mean</em></td>
</tr>
<tr>
<td>Misheard</td>
<td><em>that's not what i said</em></td>
</tr>
<tr>
<td>Privacy</td>
<td><em>none of your business</em></td>
</tr>
<tr>
<td>Repetition</td>
<td><em>you already said that</em></td>
</tr>
<tr>
<td>Criticism</td>
<td><em>you're bad at this</em></td>
</tr>
<tr>
<td>Offensive</td>
<td><em>will you talk dirty</em></td>
</tr>
<tr>
<td>Negative Navigation</td>
<td><em>change the subject</em></td>
</tr>
<tr>
<td>Stop</td>
<td><em>goodbye</em></td>
</tr>
<tr>
<td>Any</td>
<td><em>Any of the above</em></td>
</tr>
</tbody>
</table>
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

i went jogging with my sister

Sounds nice. Do you have a sister?
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

i played frisbee today

Glad you had fun playing frisbee with your dog!
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

I think boyfriends are great, how about you?

do you have a boyfriend
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

my cat threw up on the couch

Oh no! Did you get a new cat?
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

i'm feeling lonely

Oh, I suppose you have no friends?
What causes dissatisfaction?

We annotate 900 examples:

- Kind of bored to be honest
- Stuck inside all day

Bot error types:
- User utterance unclear?
- User already dissatisfied?

Dissatisfaction type(s):
- Bot error types
- User response $u$
Neural generative error frequency

- 22% of user utterances are unclear
- In 12% contexts, the user is already dissatisfied

53% of bot utterances contain an error
Neural generative error frequency

- 22% of user utterances are unclear
- In 12% contexts, the user is already dissatisfied

This accounts for some of the more basic errors.

Most frequent remaining errors:
- Redundant questions
- Logical errors

53% of bot utterances contain an error → 46.4%
How do bot errors cause dissatisfaction?

Subsequent dissatisfaction types

- Clarification
- Misheard
- Repetition
- Criticism
- Privacy
- Offensive
- Neg nav
- Stop

Bot errors

- Bot repetitive
- Bot redundant
- Bot unclear
- Bot hallucination
- Bot ignore
- Bot logical error
- Bot insulting
How do bot errors cause dissatisfaction?

Subsequent dissatisfaction types

<table>
<thead>
<tr>
<th></th>
<th>Clarification</th>
<th>Misheard</th>
<th>Repetition</th>
<th>Criticism</th>
<th>Privacy</th>
<th>Offensive</th>
<th>Neg nav</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bot repetitive</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bot redundant</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bot unclear</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Bot hallucination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bot ignore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Bot logical error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Bot insulting</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

- ✓ indicates positive Logistic Regression coefficient with feature significance (p<0.05) using Likelihood Ratio Test
How do bot errors cause dissatisfaction?

<table>
<thead>
<tr>
<th>Bot errors</th>
<th>Subsequent dissatisfaction types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bot repetitive</td>
<td>✓</td>
</tr>
<tr>
<td>Bot redundant</td>
<td>✓</td>
</tr>
<tr>
<td>Bot unclear</td>
<td>✓</td>
</tr>
<tr>
<td>Bot hallucination</td>
<td>✓</td>
</tr>
<tr>
<td>Bot ignore</td>
<td>✓</td>
</tr>
<tr>
<td>Bot logical error</td>
<td></td>
</tr>
<tr>
<td>Bot insulting</td>
<td>✓</td>
</tr>
<tr>
<td>Any bot error</td>
<td>✓</td>
</tr>
</tbody>
</table>

✓ indicates positive Logistic Regression coefficient with feature significance (p<0.05) using Likelihood Ratio Test.
Privacy boundaries vary

I’m struggling with depression

How are you feeling?

What did you have for lunch?

none of your business
Learning to predict dissatisfied utterances

We train a Dissatisfaction Predictor to predict the dissatisfaction score of the user’s response:

- Neural-generated bot utterance $b$
- How are you doing today?
- ok but my cat threw up on the couch
- oh no! did you get a new cat?
- what

Dissatisfaction score (provided by classifier): 0.64
Choosing better bot utterances

We use the Dissatisfaction Predictor to choose the best bot utterance:

P(dissatisfied)=0.48

P(dissatisfied)=0.14

Choose this response

Dissatisfaction Predictor
Choosing better bot utterances

Human preference test:
Top-p (nucleus) sample 20 responses; compare predictor-best to randomly-sampled

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor-best</td>
<td>46.3%</td>
</tr>
<tr>
<td>Random</td>
<td>35.6%</td>
</tr>
<tr>
<td>No preference</td>
<td>18.1%</td>
</tr>
</tbody>
</table>

The dissatisfaction predictor can help avoid poor-quality bot utterances!

Binomial test p-value = 0.03
Null hypothesis: Predictor-best > Random
In summary

Real-life deployment brings unique challenges.

Neural generative models fail if you carelessly unleash them in real-life settings.

Some real-life challenges like user dissatisfaction can also be learning signals.