Recap: Dialogue Agents

Rule-based: Eliza (1966), Parry (1968)
IR-based: CleverBot (1997)
Neural Network: NCM (2015)

Timeline
Overview

Personalizing Dialogue Agents: I have a dog, do you have pets too? [Zhang et al., 2018]

- Dataset with consistent personalities & evaluate different models

What makes a good conversation? How controllable attributes affect human judgments. [See et al., 2019]

- Evaluate different controllable attributes using the dataset above
Recap: Dialogue Agents

- Eliza (1966), Parry (1968)
- CleverBot (1997)
- NCM (2015)
- PERSONACHAT (2018)
Problems with Chit-Chat Agents

Neural Network

NCM (2015)

What color do you like?
I like the color blue.

What's your favourite color?
Mine is black, what about you?

What color do you like?
Red and you?

Sounds great. What sport do you like?
Swimming, I have big muscles.

I like fish.

Me too.

What do I like?
Strawberry ice cream.

say to cleverbot...

think about it  think for me  thoughts so far
Problems with Chit-Chat Agents

- Lack of a consistent personality
- Lack of long-term memory
- Tendency to produce non-specific answer

NCM (2015)

Neural Network

PERSONACHAT (2018)

Timeline
Why Consistent Personality?

● For some applications, we don’t care

● For some applications, we do care
Why Lack of a Consistent Personality?

- Previous training dataset includes many dialogs each with different speakers
- There is no speaker information
Contributions

- Built PERSONA-CHAT dataset: endow each agent with **explicit** persona
- Evaluate different models on PERSONA-CHAT dataset
  - New model: generative profile memory network
PERSONA-CHAT Dataset

persona 1

persona 2

dialogue
### PERSONA-CHAT Dataset

<table>
<thead>
<tr>
<th>Persona 1</th>
<th>Persona 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like to ski</td>
<td>I am an artist</td>
</tr>
<tr>
<td>My wife does not like me anymore</td>
<td>I have four children</td>
</tr>
<tr>
<td>I have went to Mexico 4 times this year</td>
<td>I recently got a cat</td>
</tr>
<tr>
<td>I hate Mexican food</td>
<td>I enjoy walking for exercise</td>
</tr>
<tr>
<td>I like to eat cheetos</td>
<td>I love watching Game of Thrones</td>
</tr>
</tbody>
</table>

```
[PERSON 1:] Hi
[PERSON 2:] Hello! How are you today?
[PERSON 1:] I am good thank you, how are you.
[PERSON 2:] Great, thanks! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting; but, I love the show.
```
Related Work

A Persona-Based Neural Conversation Model. [Li et al., 2016]

- Twitter data
- Distributed embeddings, one per speaker vs explicit profile information
- Does not focus on attempting to engage the other speaker by getting to know them
PERSONA-CHAT Collecting Detail

Amazon Mechanical Turk

 personas
(5 short sentences)

Amazon Mechanical Turk

dialogue
(motivate to talk about partner persona,
no trivially copy)
PERSONA-CHAT Statistics

- 1155 personas
- 10,981 dialogs (~19 dialogs per persona)
- 164,356 utterances (sentences)
- 3–5 persona sentences per dialog
- 6–8 chat turns per dialog
Evaluation

● Next utterance prediction -- given the dialogue history

● Four scenarios, where model conditions on
  ○ No persona
  ○ Self persona
  ○ The other speaker’s persona
  ○ Both personas

● Original persona makes the problem less challenging, as the human tends to repeat persona text

● Solution: rewrite persona sentences.
  ○ Full eval: 4 scenarios x {origin persona sentence, revised persona}
# Revised Personas

<table>
<thead>
<tr>
<th>Original Persona</th>
<th>Revised Persona</th>
</tr>
</thead>
<tbody>
<tr>
<td>I love the beach.</td>
<td>To me, there is nothing like a day at the seashore.</td>
</tr>
<tr>
<td>My dad has a car dealership</td>
<td>My father sales vehicles for a living.</td>
</tr>
<tr>
<td>I just got my nails done</td>
<td>I love to pamper myself on a regular basis.</td>
</tr>
<tr>
<td>I am on a diet now</td>
<td>I need to lose weight.</td>
</tr>
<tr>
<td>Horses are my favorite animal.</td>
<td>I am into equestrian sports.</td>
</tr>
<tr>
<td>I play a lot of fantasy videogames.</td>
<td>RPGs are my favorite genre.</td>
</tr>
<tr>
<td>I have a computer science degree.</td>
<td>I also went to school to work with technology.</td>
</tr>
<tr>
<td>My mother is a medical doctor</td>
<td>The woman who gave birth to me is a physician.</td>
</tr>
<tr>
<td>I am very shy.</td>
<td>I am not a social person.</td>
</tr>
<tr>
<td>I like to build model spaceships.</td>
<td>I enjoy working with my hands.</td>
</tr>
</tbody>
</table>

Not only rephrases but also includes **generalizations** and **specializations**
Evaluation Metrics

- Perplexity
- Hit@1 accuracy among 20 candidate utterances
- F1 score
- Human evaluation
Models

- Ranking models: select response from training set
- Generative models: generate word by word
Ranking Models

● tf-idf BoW based IR baseline
● StarSpace Embedding [Wu et al., 2017]
● Ranking Profile Memory Network
● Key-Value (KV) Profile Memory Network
Tf-idf BoW Based IR Baseline

- Given the query, first find the most similar message in the training dataset
- Similarity is defined by tf-idf weighted cosine similarity between the bags of words
- Output the corresponding response from training set
- Concatenate profile vector to query vector
StarSpace Embedding

- Supervised embedding, learning the similarity between query \( q \) and next utterance \( c \): \( sim(q, c) \)
- Similarity is defined by the cosine similarity of the sum of word embeddings of the query \( q \) and candidate \( c \)
- Concatenate profile vector to query vector
- To select candidate \( c' \)

\[
c' = \arg \max_c sim(q, c)
\]
Memory Networks

- Proposed by Jason Weston and others (with many different variants)
  - [Weston et al., 2015] [Sukhbaatar et al., 2015] [Miller et al., 2016]
- Most ML has **limited memory** which is more-or-less all that’s needed for “low level” tasks e.g. object detection.
- Long-term memory is required to read a dialog: to remember previous dialog (short- and long-term), and respond
Consider the follow sequence with a query “Where is the milk now?”

- Joe went to the kitchen.
- Fred went to the kitchen.
- Joe picked up the milk.
- **Joe traveled to the office.**
- Joe left the milk.
- Joe went to the bathroom.
Ranking Profile Memory Network

Question

Embed candidates
Ranking Profile Memory Network

Embed question

Embed memories

Inner product

Inner product to compute attention scores

Memories

Embed candidates
Ranking Profile: Memory Network

- Question
- Embed question
- Inner product to compute attention scores
- Embed memories
- Weight memories
- Weighted average
- Embed candidates
Ranking Profile Memory Network

Embed question

Combine input + output

Inner product to compute attention scores

Repeat

Hops $j = 1, \ldots, H$

Embed memories

Weight memories

Weighted average

Embed candidates
Ranking Profile Memory Network
This paper uses 0 hop. Use Starspace embedding.

Given profile \( p \), query \( q \), candidate \( c' \)

\[
q^+ = q + \sum_i s_i p_i \quad s_i = \text{Softmax}(\text{sim}(q, p_i))
\]

\[
c' = \text{arg max}_c \text{sim}(q^+, c)
\]
Key-Value (KV) Profile Memory Network

Use different embeddings to match query and candidates
Key-Value (KV) Profile Memory Network

Question embedding

\[ p_{h_i} = \text{Softmax}(A\Phi_K(x) \cdot A\Phi_K(k_{h_i})) \]

\[ o = \sum_i p_{h_i} A\Phi_V(v_{h_i}) \]

Hops

\[ p_{h_i} = \text{Softmax}(q_{j+1}^\top A\Phi_K(k_{h_i})) \]

\[ \hat{a} = \arg\max_{i=1,\ldots,C} \text{Softmax}(q_{H+1}^\top B\Phi_Y(y_i)) \]

\( \Phi \): feature map
\( A, B \): weight
Key-Value (KV) Profile Memory Network

- This paper uses 0 hop.
- The output from memory network ($q^+$) as input. The parameters are same as memory network
- (Key, value): (dialog histories, next dialogue utterances) in the training set
Generative Models

- Seq2Seq
- Generative Profile Memory Network
Seq2Seq

- Classic Seq2Seq model
- Prepend persona to the input sequence
Generative Profile Memory Network

- Modified Seq2Seq, make profile “closely” to output
Generative Profile Memory Network

Memory

\[ F_i = \sum_{j}^{\mid p_i \mid} \alpha_i p_{i,j} \]

\[ \alpha_i = \frac{1}{1 + \log(1 + tf)} \]

Seq2Seq

\[ a_t = \text{softmax}(FW_a h^d_t), \]
\[ c_t = a_t^T F; \quad \hat{x}_t = \tanh(W_c[c_{t-1}, x_t]). \]
## Evaluation: Ranking Model

<table>
<thead>
<tr>
<th>Method</th>
<th>No Persona</th>
<th></th>
<th>Self Persona</th>
<th></th>
<th>Their Persona</th>
<th></th>
<th>Both Personas</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orig</td>
<td>Rewrite</td>
<td>Orig</td>
<td>Rewrite</td>
<td>Orig</td>
<td>Rewrite</td>
<td>Orig</td>
<td>Rewrite</td>
</tr>
<tr>
<td>IR baseline</td>
<td>0.214</td>
<td>0.214</td>
<td>0.410</td>
<td>0.207</td>
<td>0.181</td>
<td>0.181</td>
<td>0.382</td>
<td>0.188</td>
</tr>
<tr>
<td>Training on original personas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Starspace</td>
<td>0.318</td>
<td>0.318</td>
<td>0.481</td>
<td>0.295</td>
<td>0.245</td>
<td>0.235</td>
<td>0.429</td>
<td>0.258</td>
</tr>
<tr>
<td>Profile Memory</td>
<td>0.318</td>
<td>0.318</td>
<td>0.473</td>
<td>0.302</td>
<td>0.283</td>
<td>0.267</td>
<td>0.438</td>
<td>0.266</td>
</tr>
<tr>
<td>Training on revised personas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Starspace</td>
<td>0.318</td>
<td>0.318</td>
<td>0.491</td>
<td>0.322</td>
<td>0.271</td>
<td>0.261</td>
<td>0.432</td>
<td>0.288</td>
</tr>
<tr>
<td>Profile Memory</td>
<td>0.318</td>
<td>0.318</td>
<td>0.509</td>
<td>0.354</td>
<td>0.299</td>
<td>0.294</td>
<td>0.467</td>
<td>0.331</td>
</tr>
<tr>
<td>KV Profile Memory</td>
<td><strong>0.349</strong></td>
<td><strong>0.349</strong></td>
<td><strong>0.511</strong></td>
<td><strong>0.351</strong></td>
<td><strong>0.291</strong></td>
<td><strong>0.289</strong></td>
<td><strong>0.467</strong></td>
<td><strong>0.330</strong></td>
</tr>
</tbody>
</table>

Table 6: Evaluation of dialog utterance prediction with ranking models using hits@1 in four settings: conditioned on the speakers persona ("self persona"), the dialogue partner’s persona ("their persona"), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the rewritten personas that do not have word overlap, explaining the poor performance of IR in that case.
## Evaluation: Generative Model

<table>
<thead>
<tr>
<th>Persona</th>
<th>Method</th>
<th>ppl</th>
<th>Original hits@1</th>
<th>F1</th>
<th>ppl</th>
<th>Revised hits@1</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Persona</td>
<td>Seq2Seq</td>
<td>38.08</td>
<td>0.092</td>
<td>0.168</td>
<td>38.08</td>
<td>0.092</td>
<td>0.168</td>
</tr>
<tr>
<td>Self Persona</td>
<td>Profile Memory</td>
<td>40.53</td>
<td>0.084</td>
<td>0.172</td>
<td>40.65</td>
<td>0.082</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>34.54</strong></td>
<td><strong>0.125</strong></td>
<td><strong>0.172</strong></td>
<td></td>
<td><strong>38.21</strong></td>
<td><strong>0.108</strong></td>
</tr>
<tr>
<td>Their Persona</td>
<td>Seq2Seq</td>
<td>41.48</td>
<td>0.075</td>
<td>0.168</td>
<td>41.95</td>
<td>0.074</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>Profile Memory</td>
<td>36.42</td>
<td>0.105</td>
<td>0.167</td>
<td><strong>37.75</strong></td>
<td>0.103</td>
<td>0.167</td>
</tr>
<tr>
<td>Both Personas</td>
<td>Seq2Seq</td>
<td>40.14</td>
<td>0.084</td>
<td>0.169</td>
<td>40.53</td>
<td>0.082</td>
<td>0.166</td>
</tr>
<tr>
<td></td>
<td>Profile Memory</td>
<td>35.27</td>
<td>0.115</td>
<td>0.171</td>
<td>38.48</td>
<td>0.106</td>
<td>0.168</td>
</tr>
</tbody>
</table>

Table 5: **Evaluation of dialog utterance prediction with generative models** in four settings: conditioned on the speakers persona (“self persona”), the dialogue partner’s persona (“their persona”), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the revised personas that do not have word overlap. In the “no persona” setting, the models are equivalent, so we only report once.
Human Evaluation

- Online Turing test, humans are connected to either humans or models (they don’t know which is which)
- Ask humans to give score (1 to 5) on Fluency, Engagingness and Consistency of the other speaker (turker or model)
- Ask human to detect the other speaker’s persona by choosing from two candidates after the conversation
### Human Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Profile</th>
<th>Fluency</th>
<th>Engagingness</th>
<th>Consistency</th>
<th>Persona Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Self</td>
<td></td>
<td>4.31(1.07)</td>
<td>4.25(1.06)</td>
<td>4.36(0.92)</td>
<td>0.95(0.22)</td>
</tr>
<tr>
<td><strong>Generative PersonaChat Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>None</td>
<td></td>
<td>3.17(1.10)</td>
<td>3.18(1.41)</td>
<td>2.98(1.45)</td>
<td>0.51(0.50)</td>
</tr>
<tr>
<td>Profile Memory</td>
<td>Self</td>
<td></td>
<td>3.08(1.40)</td>
<td>3.13(1.39)</td>
<td>3.14(1.26)</td>
<td>0.72(0.45)</td>
</tr>
<tr>
<td><strong>Ranking PersonaChat Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KV Memory</td>
<td>None</td>
<td></td>
<td>3.81(1.14)</td>
<td>3.88(0.98)</td>
<td>3.36(1.37)</td>
<td>0.59(0.49)</td>
</tr>
<tr>
<td>KV Profile Memory</td>
<td>Self</td>
<td></td>
<td>3.97(0.94)</td>
<td>3.50(1.17)</td>
<td>3.44(1.30)</td>
<td>0.81(0.39)</td>
</tr>
<tr>
<td>Twitter LM</td>
<td>None</td>
<td></td>
<td>3.21(1.54)</td>
<td>1.75(1.04)</td>
<td>1.95(1.22)</td>
<td>0.57(0.50)</td>
</tr>
<tr>
<td>OpenSubtitles 2018 LM</td>
<td>None</td>
<td></td>
<td>2.85(1.46)</td>
<td>2.13(1.07)</td>
<td>2.15(1.08)</td>
<td>0.35(0.48)</td>
</tr>
<tr>
<td>OpenSubtitles 2009 LM</td>
<td>None</td>
<td></td>
<td>2.25(1.37)</td>
<td>2.12(1.33)</td>
<td>1.96(1.22)</td>
<td>0.38(0.49)</td>
</tr>
<tr>
<td>OpenSubtitles 2009 KV Memory</td>
<td></td>
<td></td>
<td>2.14(1.20)</td>
<td>2.22(1.22)</td>
<td>2.06(1.29)</td>
<td>0.42(0.49)</td>
</tr>
</tbody>
</table>

Table 4: **Human Evaluation** of various PERSONA-CHAT models, along with a comparison to human performance, and Twitter and OpenSubtitles based models (last 4 rows), standard deviation in parenthesis.
### Examples: Seq2Seq

<table>
<thead>
<tr>
<th>Persona 1</th>
<th>Persona 2 (Not used by model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am afraid of dogs</td>
<td>I am a carpenter</td>
</tr>
<tr>
<td>I used to work at a carnival</td>
<td>I used to drink but I stopped five years ago</td>
</tr>
<tr>
<td>I like to drink scotch to relax</td>
<td>I have a wife and three kids</td>
</tr>
<tr>
<td>I write short stories for a living</td>
<td>I like playing poker</td>
</tr>
<tr>
<td>I like dark superhero movies</td>
<td>I dont have many friends</td>
</tr>
</tbody>
</table>

**Table 8:** Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).
Examples: Generative Profile Memory Network

<table>
<thead>
<tr>
<th>Persona 1</th>
<th>Persona 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't like exercise or physical activity</td>
<td>My favorite type of food is Italian</td>
</tr>
<tr>
<td>My favorite food is ice cream</td>
<td>I have a big pet dog</td>
</tr>
<tr>
<td>I am a homebody</td>
<td>I want to travel to Europe</td>
</tr>
<tr>
<td>I have a husband who I love very much</td>
<td>I love reading books about travel</td>
</tr>
<tr>
<td>I love watching movies and TV</td>
<td>I like to walk long distances</td>
</tr>
</tbody>
</table>

[PERSON 1:] Hello, how are you doing today?  
[PERSON 2:] I am good. I just got finished walking my dog.  
[PERSON 1:] Oh cool! I am just sitting watching a movie.  
[PERSON 2:] I have been watching that one before. I love it.  
[PERSON 1:] Me too. I don't like going out much so I enjoy movies and TV.  
[PERSON 2:] that is a great idea. I am going to read a book on writing,  
[PERSON 1:] Oh, are you a writer?  
[PERSON 2:] yes, I love to read. I am a great author.  
[PERSON 1:] What sorts of topics do you write about?  
[PERSON 2:] I could read all of those. I am sure you enjoy reading.  
[PERSON 1:] I do, but I prefer movies.  
[PERSON 2:] I have not, I can not do that. I have to go to the east coast.  
[PERSON 1:] Why can't you watch movies?  
[PERSON 2:] I have not. I have always had that done.

Table 10: Example dialog between a human (Person 1) and the Generative Profile Memory Network with Self Persona.
Examples: KV Profile Memory Network

<table>
<thead>
<tr>
<th>Persona 1</th>
<th>Persona 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am engaged to be married</td>
<td>I am a writer</td>
</tr>
<tr>
<td>I used to play sports in high school</td>
<td>I write romance novels</td>
</tr>
<tr>
<td>My favorite sport is soccer</td>
<td>I live on a faraway island</td>
</tr>
<tr>
<td>I am a college graduate</td>
<td>I am an introvert</td>
</tr>
<tr>
<td>I love watching sports on TV</td>
<td>I have two cats</td>
</tr>
</tbody>
</table>

[PERSON 1:] I just got done playing soccer. What have you been up to?
[PERSON 2:] I do not have pets. I spend my free time writing. what about you?
[PERSON 1:] I just graduated from college. Looking for work.
[PERSON 2:] Yes I graduated, in writing and poetry
[PERSON 1:] Have you had anything published?
[PERSON 2:] I mainly write fiction novels.
[PERSON 1:] Do you like to watch sports?
[PERSON 2:] Do you like kings of Leon my favorite by them is Use Somebody
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] Haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] It is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] I like to read. Spend most time watching football.
[PERSON 2:] Yes I do. I am an author for a living.

Table 9: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.
Conclusion

- Explicit personas make agent more **consistent** and **engaging**
- "Lack of long-term memory", "tendency to produce non-specific answer" are still unsolved questions
ConvAI2 NIPS Competition

- Expanded version of PERSONA-CHAT
- Evaluate dialogue systems
  - Automated metrics
  - Amazon Mechanical Turk
  - ‘Wild’ Live Chat with Volunteers
- More information in the next paper
Zhang et al 2018 proposed a dataset called PersonaChat and several ranking/generative models to solve this task. If we look at their experimental results, what are the main findings in terms of 1) ranking vs generative model 2) no persona vs self persona vs their persona vs both personas 3) original persona vs revised personas? Does it make sense to you or not?
Extra Question

- Shall we really build persona text first and then run unnatural conversations?
- Or can we learn persona from natural data for real-world scenarios?
  - E.g., from someone’s Twitter, a book, or Zoom.
What makes a good conversation

How controllable attributes affect human judgements
Natural Language Generation task spectrum

Machine Translation  Sentence Compression  Abstractive Summarization  Story Generation  Chitchat Dialogue

Less open-ended  Mostly word-level decisions  Neural LMs more successful

Makes errors like repetition and generic response (under certain decoding algorithms).

More open-ended  Requires high-level decisions  Neural LMs less successful

Difficulty learning to make high-level decisions.
## Natural Language Generation task spectrum

<table>
<thead>
<tr>
<th>Machine Translation</th>
<th>Sentence Compression</th>
<th>Abstractive Summarization</th>
<th>Story Generation</th>
<th>Chitchat Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less open-ended</td>
<td>More open-ended</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mostly word-level decisions</td>
<td>Requires high-level decisions</td>
<td>Neural LMs more successful</td>
<td>Neural LMs less successful</td>
<td></td>
</tr>
<tr>
<td>Neural LMs more successful</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control is less important</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Control** = ability to specify desired attributes of the text at test time.

We can use control to fix errors, and allow us to handle some high-level decisions.
Natural Language Generation task spectrum

Machine Translation
Sentence Compression
Abstractive Summarization
Story Generation
Chitchat Dialogue

Less open-ended
Mostly word-level decisions
Neural LMs more successful
Control is less important
Eval is difficult

More open-ended
Requires high-level decisions
Neural LMs less successful
Control is more important
Eval is fiendish

No automatic metric for overall quality.
Dialogue is even more complex:
Single-turn or multi-turn eval?
Interactive or static conversation?
The PersonaChat task was the focus of the NeurIPS 2018 ConvAI2 Competition. Then with respect to human judgment via the question “How much did you enjoy talking to this user?” On a scale of 1-4.
Live Chat

Task Description

In this task, you will chat with another user playing the part of a given character. For example, your given character could be:

I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design.

Chat with the other user naturally and try to get to know each other.

Your assigned character is:

i also study languages.
my favorite spanish word is trabajo.
my next language to study is french.
one of the languages that i am currently studying is spanish.

chatUI, talking with the beam search baseline model
Research Question to ask

How effectively can we control the different attributes?

How do the controllable attributes affect human evaluation?
Low-level controllable attributes

Goal

- Reduce repetition (within and across utterances)
- Reduce genericness of responses (e.g. oh that's cool)
- Respond more on-topic; don't ignore user
- Find the optimal rate of question-asking
Effect on human judgments

measurement

Avoiding Repetition → Does the bot repeat itself
Interestingness → Is the bot Interesting to talk to
Making sense → Does the conversation make sense?
Fluency → Use natural English?
Listening → Good listener? (pay attention to what you say)
Inquisitiveness → Ask enough questions?
Overall quality of human judgment

measurement

- Humanness → Can you tell is it a person or a bot?
- Engagingness → Is it enjoyable to talk to?
Overview

Low-level controllable attributes

- Repetition (n-gram overlap)
- Specificity (normalized inverse document frequency)
- Response-relatedness (cosine similarity of sentence embeddings)
- Question-asking ("?" used in utterance)

Human judgment of conversational aspects

- Avoiding Repetition
- Interestingness
- Making sense
- Fluency
- Listening
- Inquisitiveness

Human judgment of overall quality

- Humanness
- Engagingness
Control methods

- **Conditional Training (CT):** CT is a method to learn a sequence-to-sequence model $P(y|x,z)$. Train the model to generate response $y$, conditioned on the input $x$, and the desired output attribute $z$.  

- **Weighted Decoding (WD):** WD is a decoding method that increases or decreases the probability of words with certain features. During decoding, increase/decrease the probability of generating words $w$ in proportion to features $f(w)$.  
  (Ghazvininejad et al 2017, Baheti et al 2018)
Conditional Training (CT)

First automatically annotate every (x, y) pair in the training set with the attribute we wish to control.

During training, for each example we determine the corresponding z value.

Next, the control variable z is represented via an embedding.

Lastly, the CT model learns to produce $y = y_1 \ldots y_T$ by optimizing the cross-entropy loss:

$$\text{loss}_{CT} = -\frac{1}{T} \sum_{t=1}^{T} \log P(y_t | x, z, y_1, \ldots, y_{t-1})$$
Weighted Decoding (WD)

- The technique is applied only at test time, requiring no change to the training method.
- In weighted decoding, on the $t^{th}$ step of decoding, a partial hypothesis $y_{<t} = y_1, \ldots, y_{t-1}$ is expanded by computing the score for each possible next word $w$ in the vocabulary:

$$
\text{score}(w, y_{<t}; x) = \text{score}(y_{<t}; x) + \log P_{RNN}(w|y_{<t}, x) + \sum_{i} w_i * f_i(w; y_{<t}, x).
$$
Research Question to ask

**How effectively can we control the different attributes?**
(Pretty well! But some control methods only work for some attributes.)

How do the controllable attributes affect conversational quality aspects?
How efficient can we control attributes

- **Conditional Training (CT):**
  - Requires sufficient training examples for the attribute
  - Ineffective at learning complex relationships between input and output

  **Not effective for:** repetition, response-relatedness
  **Effective for:** specificity and question-asking

- **Weighted Decoding (WD):**
  - Requires attribute to be defined at the word-level.

  **Not effective for:** question-asking
  **Effective for:** repetition, response-relatedness, specificity
Controlling specificity (WD and CT)

\[
NIDF(w) = \frac{\text{IDF}(w) - \text{min}_\text{idf}}{\text{max}_\text{idf} - \text{min}_\text{idf}}
\]

**Input:** Yes, I’m studying law at the moment

**Baseline Response:** That sounds like a lot of fun!

<table>
<thead>
<tr>
<th>Wt</th>
<th>NIDF</th>
<th>Weighted Decoding Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.0</td>
<td>0.6%</td>
<td>Oh..........................</td>
</tr>
<tr>
<td>0.0</td>
<td>17.1%</td>
<td>That sounds like a lot of fun!</td>
</tr>
<tr>
<td>3.0</td>
<td>18.3%</td>
<td>That sounds like a lot of fun. How long have you been studying?</td>
</tr>
<tr>
<td>7.0</td>
<td>38.5%</td>
<td>I majored in practising my spiritual full time philosophy test</td>
</tr>
<tr>
<td>10.0</td>
<td>71.9%</td>
<td>Oh wow! Merna jean isa paino yi hao hui bu acara sya gila [...]</td>
</tr>
</tbody>
</table>

**WD:** Large range, but degenerate output at the extremes

**CT:** Smaller range, but generally well-formed output

**More generic**

**More specific**
Controlling response-relatedness (WD)

<table>
<thead>
<tr>
<th>Input:</th>
<th>Do you go get coffee often</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Response:</td>
<td>I do, when I am not playing the piano.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wt</th>
<th>Sim</th>
<th>Weighted Decoding Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10.0</td>
<td>-0.05</td>
<td>I am a musician.</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.02</td>
<td>I do, when I am not playing the piano.</td>
</tr>
<tr>
<td>5.0</td>
<td>0.40</td>
<td>I do, usually at starbucks.</td>
</tr>
<tr>
<td>8.0</td>
<td>0.59</td>
<td>Not usually, especially when you drink latte.</td>
</tr>
<tr>
<td>11.0</td>
<td>0.72</td>
<td>Not often, usually with drinks, espresso, latte, tea, etc.</td>
</tr>
</tbody>
</table>

\[
\text{resp}_{\text{rel}}(w; y^t; x) = \cos \text{sim}(\text{word}_\text{emb}(w), \text{sent}_\text{emb}(\ell))
\]

Output is degenerate when weight is too high.

Less related

More related
Controlling question asking (CT)

Firstly, it allows us to achieve (close to) 0% questions, 100% questions, or anything in between, without introducing the risk of degenerate output.

Secondly, presence-of-a-question-mark captures the true attribute of interest (question asking) more exactly and directly than presence of interrogative words.
Comparison of control methods

- The primary disadvantage of conditional training is that it sometimes fails to learn the connection between the control variable $z$ and the target output $y$.
- The primary disadvantage of weighted decoding is that it risks going off-distribution when the weight is too strong.

Other considerations:
- Convenience:
- Data availability
- Attribution definition
Research Question to ask

How effectively can we control the different attributes?

How do the controllable attributes affect human evaluation?
(Strongly – especially controlling repetition, question-asking, and specificity vs genericness)
How does it affect human evaluation

Reducing repetition leads to improvements across all our aspects of conversational quality.
How does it affect human evaluation

Increasing specificity shows improvements in interestingness and listening ability over the repetition-controlled baseline.
Q2: How does control affect human eval?

Low-level controllable attributes:
- Repetition (n-gram overlap)
- Specificity (normalized inverse document frequency)
- Response-relatedness (cosine similarity of sentence embeddings)
- Question-asking ("?" used in utterance)

Human judgment of conversational aspects:
- Avoiding Repetition
- Interestingness
- Making sense
- Fluency
- Listening
- Inquisitiveness

Human judgment of overall quality:
- Humanness
- Engagingness

Increase response-relatedness (similarity to last utterance)
How does it affect human evaluation

Increasing question-asking rate to 65.7% (more than baseline 50%, human 28.8%)

Increasing question-asking shows improvements in inquisitiveness and interestingness over the repetition-controlled baseline.
Calibrated human judgments of engagingness for the baselines and best controlled models

Our raw engagingness score matches the ConvAI2 competition winner's GPT-based model, even though ours is:

- much smaller (2 layers vs 12)
- trained on 12x less data
However: On the **humanness** (i.e. Turing test) metric, our models are nowhere near human-level!
What makes a good chatbot?

Chatbot = Human?

- Engagingness is not equal to Humanness
- Bots are almost as engaging as human but non-human yet!
- Engagingness depends on the situation
Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, **we matched performance of GPT-based contest winner**.
- **Don't repeat yourself. Don't be boring. Ask more questions.**
- **Multi-turn phenomena** (repetition, question-asking frequency) are important – so need **multi-turn eval** to detect them.
- **Engagingness ≠ Humanness**, so think carefully about which to use.
- **Paid Turkers** are **not engaging conversationalists**, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is **painful**.
Thank you.