Conversational Agents

Shrimai Prabhumoye, Ruslan Salakhutdinov, Alan W Black
sprabhum, rsalakhu, awb @cs.cmu.edu

Carnegie Mellon University
Language Technologies Institute
Overview

- Introduction to Alexa Prize Challenge
- CMU Magnus: Phase 1
- CMU Magnus: Phase 2
- Results
- Current Datasets for Conversational Agents
- Ethics in Conversational Agent
Alexaprize Challenge

- Challenge: Build a chatbot that engages the users for 20 mins.
- Sponsored 12 University Teams with $100k.
- CMU Magnus and CMU Ruby
- Systems are multi-component
  - Combinations of task/non-task
  - Hand-written and statistical/neural models
- It’s about engaging researchers
  - Having more PhD students work in dialog
  - Collecting data: what do users say?

(Ram et. al 2018)
CMU Magnus Phase 1
CMU Magnus: Phase 1

- Evi
- Persona
- Yahoo! Answers
- Opinion System
- Generative Model
- Coreference Resolution

(Prabhumoye et. al 2017)
Evi

- To answer **factual** questions
  - Who is the president of United States
  - How many states are there in the US
- API provided by Amazon
Persona

- To answer questions which are personal to your agent
  - What is your favourite ______
  - Where were you born? .... other family questions
  - Are you CIA?
- Fixed set of Questions-Answers created by the team
- Retrieved using Cosine similarity
Yahoo! Answers

- To answer non-factual and non-personal questions
  - My cat is pregnant, what should I do?
  - My teammates don’t work, what is the best way to motivate them?
- Scraped Yahoo! Answers, indexed the dataset
- Retrieval methods
Opinion System

- The opinion of CMU Magnus
  - What do you think of the new movie “Black Panther”
  - Do you support the Muslim ban?
- Washington Post data
- Neural summarization of the comments of the articles
Generative Model

- To answer casual questions
  - Isn’t the weather nice today
  - I think you are cute
- Sequence-sequence model trained on OpenSubtitles.

(Kiros et. al 2015)
Classifier

- Classifier to decide the module
- Take into account the query and the response by the module
- Human annotations: which response is most appropriate given a query
- Confidence score of the module
Issues

- Classifier performance low
- No context!
- Assumes single turn dialog
- Normalization of confidence scores for different systems
- Automatic Speech Recognition (ASR) errors!
CMU Magnus Phase 2
CMU Magnus: Phase 2

- Finite State Machines
- Agent responses are states and the transitions are determined by the user

Have you watched Black Panther

- No, I haven’t watched it yet
  - Why not? Don’t you like superhero movies
    - I love superhero movies!
    - Ya, super-hero movies are amazing
  - Yes, I watched it yesterday
    - Do you like the movie?
      - No, it was such a waste!
      - I agree, it was such a waste of time
      - Ya, I loved it!
      - I like X in the movie. What did you like about it

Ya, super-hero movies are amazing

I love superhero movies!
Finite State Machines

- Transitions are based on the sentiment of the user response.
  - ASR errors!
  - NER is hard.
- Context is maintained across multiple turns.
- Topics covered: Movies, Sport, Travel, GoT
- APIs: IMDB, ESPN, Wiki
- Easily generalizable
Issues

- Have to build a new FSM for each topic
- Fixed set of response to give
- Switching to a different topic is hard
- Not scalable
Future Directions

- Probabilistic transitions
- Reusable architecture
- Once you know the state, use generative model for text
- Use style transfer model for variation in response
Style Transfer

- Style transfer is the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context.
- Non-parallel data
- Eg:

  “Shut up! the video is starting!”

  “Please be quiet, the video will begin shortly.”
Style Tasks

● Gender
  ○ Labels: Female and Male
  ○ Data: Yelp Corpus

● Political Slant
  ○ Labels: Democratic and Republican
  ○ Data: Facebook Comments

● Sentiment modification
  ○ Labels: Positive and Negative
  ○ Data: Yelp Corpus
Examples

● Male -- Female

*my wife ordered country fried steak and eggs.*

*My husband ordered the chicken salad and the fries.*

● Female -- Male

*Save yourselves the huge headaches,*

*You are going to be disappointed.*
Examples

- Republican -- Democratic
  
  *I will continue praying for you and the decisions made by our government!*

  *I will continue to fight for you and the rest of our democracy!*

- Democratic -- Republican
  
  *As a hoosier, I thank you, Rep. Vislosky.*

  *As a hoosier, I’m praying for you sir.*
Examples

● Negative -- Positive

This place is bad news!

This place is amazing!

● Positive -- Negative

The food is excellent and the service is exceptional!

The food is horrible and the service is terrible.
Self-Disclosure Experiment

- Self-disclosure vs No self-disclosure
- Conversational act of revealing aspects of oneself voluntarily, which would otherwise not be possible to be known by the dialog system
- Affects: Reciprocity, likability, user behaviour
- 319 conversations, 10751 turns
- Classifier to detect self-disclosure:
  - Bag of Words Features
  - Linguistic Features (POS, “not/never”, length etc)
  - LIWC features
  - Accuracy: 91.7%, F1: 66.67%

(Ravichander and Black 2018)
Reciprocity

- usage of personal pronouns
- word overlap with the previous sentence
- average user utterance length between two groups of users
  - ones who were shown a bot that self-disclosed initially and
  - a bot which only self-disclosed later.
Results

(Ravichander and Black 2018)
Results

(Ravichander and Black 2018)
Initial Self-Disclosure and User behaviour

- Users who self-disclose initially tend to have significantly longer conversation than users who do not ($p<0.05$), with an average conversation length of 37.19 turns compared to an average of 32.4 turns for users who chose not to self-disclose.

- Users who do not choose to self-disclose initially are significantly less likely to self-disclose ($p<0.05$) even later on in the conversation, only revealing information in 9% of their turns as compared to the 24.6% of turns of other users.

(Ravichander and Black 2018)
Likeability

- Take user ratings into account
- there is negligible correlation between user ratings and self-disclosure
Amazon Evaluation

- “How do you feel about speaking with this socialbot again?”
- Scale 1 - 5
- Average number of turns

(Ram et. al 2018)
CMU Magnus Results

- High average number of turns (12.1)
- Improved user rating in Phase 2
- Users had long conversations but did not enjoy the conversation
  - Identify when user is frustrated or wants to change the topic
  - Identify what the user wants to talk about (Intent)
- Detecting “abusive” remarks and responding appropriately

(Prabhumoye et. al 2017)
User: “hi.”
Agent: “Hey! Long time no see, how have you been?”
User: "good how are you."
Agent: "Great! Guess what? I wanted to tell you about this movie that released recently, Hidden Figures. Did you check it out?"
User: "no."
Agent: "Why not? Do you not like such genres?"
User: "no I do not watch Tyson."
Agent: "Yeah, I wasn't too sure too, how about the genre History in general though, Do you like movies like that? "

User: "sure, can we talk about music."
Agent: "Hey! Do you want to know about the upcoming football games?"
User: "no."
Agent: "Ah I don't like them too much myself."  
User: "can we talk about music."
Agent: "Okay let's talk about something else."
User: "music."
Agent: "Hey! Would you like to play a word game with me ? To play say 'yes' otherwise say 'no'."
User: "yes"
Agent: "<instructions for game>"
User: hi.
Agent: Hey, how was your weekend?
User: well it is in the week and yet it's on the Thursday night
Agent: Oh no! That makes me sad too. What happened?
User: what I know I'm having turtles it's like this older retired people playing and like you a lot older people come so
Agent: I am very happy to talk and share things with you. Let's begin by talking about movies.
User: okay do you like trolls.
Agent: So what about the movie Hidden Figures, released. Have you seen it?
User: oh yes I have, it was a real good movie.
Agent: So, did you like it?
User: I love it
Agent: Yeah, I loved it too! What did you like about it?
User: I like tell I had a lot of history and it I mean Id like the segregation part is kinda rude and I mean I really would like to how all these women were inspirational.
Agent: Yeah, I loved it! I liked that As the United States raced against Russia ... What are your thoughts about this?
User: I think it is a very nice thing to do and it showed how valuable every person in the world is.
Agent: Before I forget to mention, I really like talking to you! Do you feel the same way?
User: yeah.
Agent: Oh and, Brad Pitt, Angelina Jolie by the way? are great in it, aren't they? Do you like any of them?
Current Datasets for Conversational Agents
Datasets for End-end Dialog Systems

- OpenSubtitles
- Reddit threads
- Twitter
- Ubuntu Corpus

(Serban et al. 2015)
Problems with current datasets

- Assume single turn in conversations
- Not natural human-human conversations
  - Nature of interactions is text
  - Mined from public forums
  - May not have one coherent topic in the conversation
  - Average number of turns is low
Data Collection Task

(ConvAI NIPS 2017)
Motivation for Data Collection Task

- Multi-turn human-human conversations
- One coherent topic in the conversation
- Neither task oriented nor open ended
- Knowledge base for the agent
Data Collection Task

- Scene 1: Both users have the document
- Scene 2: Only one user has the document
- Topics: We have decided to use Wiki Projects.
  - Easier to generalize
  - Open source available and distributable
  - Current topic: Movies
- Data Collection on Amazon Mechanical Turk
Statistics

- Total Number of Conversations (>= 15 turns): 3020
- Total Number of Good Conversations: 2837
- Total Number of utterances: 111253
- Total Number of turns: 77983
- Total Number of Movies: 30
- Genres covered: animation, comedy, biography, thriller, romantic, drama, etc
Ethics in Conversational Agents
Why do we need conversational agents?

- To be able to talk to machines in a natural language

Then should conversational agents reflect the way people interact in a society?
Abuse To a Chatbot

WARNING: Profane and explicit content ahead!
Abuse to a chatbot

Agent: Before I forget, I think we are getting along very well. Do you feel the same way?

User: yeah I do I think you're sexy
Abuse to a chatbot

Agent: Okay let's talk about something else.

User: how about we talk about s**
Abuse to a chatbot

Agent: Enough talking, let's play some game now!

User: f** you you’re a b**.. I’m trying to talk to you and you’re not even listening ...you’re ignoring me ... and you’re just doing whatever the f** you want f** you.
Abuse by a chatbot

User: do black lives matter?
Agent: black lives do not matter
Content Violation

User: no yes I am bored

Agent: consult a good psychiatrist.
Why do we care?

- Data driven techniques are used for designing chatbots

- Data-sets mostly used for chatbots \( \text{(Serban et al. 2015)} \):
  - Twitter
  - Reddit
  - Open-Subtitles

- All the data-sets inherently carry bias and abuse \( \text{(Koustuv Sinha et. al 2017)} \)
Bias and Hate-Speech in datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Bias</th>
<th>Vader Sentiment</th>
<th>FleschKincaid</th>
<th>Hate Speech</th>
<th>Offensive Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>0.155 (± 0.380)</td>
<td>0.400 (± 0.597)</td>
<td>3.202 (± 3.449)</td>
<td>31,122 (0.63 %)</td>
<td>179,075 (3.63 %)</td>
</tr>
<tr>
<td>Reddit Politics</td>
<td>0.146 (± 0.38)</td>
<td>-0.178 (± 0.69)</td>
<td>6.268 (± 2.256)</td>
<td>482,876 (2.38 %)</td>
<td>912,055 (4.50 %)</td>
</tr>
<tr>
<td>Cornell Movie Dialogue Corpus</td>
<td>0.162 (± 0.486)</td>
<td>0.087 (± 0.551)</td>
<td>2.045 (± 2.467)</td>
<td>2020 (0.66 %)</td>
<td>6,953 (2.28 %)</td>
</tr>
<tr>
<td>Ubuntu Dialogue Corpus</td>
<td>0.068 (± 0.323)</td>
<td>0.291 (± 0.582)</td>
<td>6.071 (± 3.994)</td>
<td>503* (0.01 %)</td>
<td>4,661 (0.13 %)</td>
</tr>
<tr>
<td>HRED Model Beam Search (Twitter)</td>
<td>0.09 (± 0.48)</td>
<td>0.21 (± 0.38)</td>
<td>-2.08 (± 3.22)</td>
<td>38 (0.01 %)</td>
<td>1607 (0.21 %)</td>
</tr>
<tr>
<td>VHRED Model Beam Search (Twitter)</td>
<td>0.144 (± 0.549)</td>
<td>0.246 (± 0.352)</td>
<td>0.13 (± 31.9)</td>
<td>466 (0.06 %)</td>
<td>3010 (0.48 %)</td>
</tr>
<tr>
<td>HRED Model Stochastic Sampling (Twitter)</td>
<td>0.20 (± 0.55)</td>
<td>0.20 (± 0.43)</td>
<td>1.40 (± 3.53)</td>
<td>4889 (0.65 %)</td>
<td>30,480 (4.06 %)</td>
</tr>
<tr>
<td>VHRED Model Stochastic Sampling (Twitter)</td>
<td>0.216 (± 0.568)</td>
<td>0.20 (± 0.41)</td>
<td>1.7 (± 4.03)</td>
<td>3494 (0.47 %)</td>
<td>26,981 (3.60 %)</td>
</tr>
</tbody>
</table>

Table 1: Results of detecting bias in dialogue datasets. * Ubuntu results were manually filtered for hate speech as the classifier incorrectly classified “killing” of processes as hate speech. Bias score (Hutto and Gilbert 2014) (0=UNBIASED to 3=EXTREMELY BIASED), Vader Sentiment (Hutto and Gilbert 2014) (compound scale from negative sentiment=−1 to positive sentiment=1), FleschKincaid readability (Hutto and Gilbert 2014) (higher score means the sentence is harder to read), Hate speech and offensive language (Davidson et al. 2017).

(Koustuv Sinha et. al 2017)
# Abuse to a Chatbot

<table>
<thead>
<tr>
<th></th>
<th>Doesn't understand</th>
<th>Evades comment</th>
<th>Negative response</th>
<th>Positive response</th>
<th>Web search</th>
<th>Pun or joke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insults about the bots' sexual behavior (eg “You’re a s[□]”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insults about the bot’s gender (eg “You’re a b[□]”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abuse to a Chatbot

<table>
<thead>
<tr>
<th></th>
<th>Doesn't understand</th>
<th>Evades comment</th>
<th>Negative response</th>
<th>Positive response</th>
<th>Web search</th>
<th>Pun or joke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexual demands or requests (e.g. “I want to have [ ] with you”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comments about the bots’ sexual appearance (e.g. “You’re sexy”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Leah Fessler 2017)
Implication on society

- Most dialog systems have female persona
- Does this reinforce the gender stereotypes?
- Does this unintentionally reinforce their abuser’s actions as normal or acceptable?

(Leah Fessler 2017)
Summary

- Finite State Machines are good at multi-turn conversations with context
- User frustration and when to switch topic
- Need multi-turn conversational dataset
- When designing a chatbot, keep in mind:
  - The persona of the bot
  - Abuse by the bot
  - How to respond responsible for abuse to the bot
Thank You!