Edina: Building an Open-Domain Socialbot using Self-Dialogues

ILCC, School of Informatics, University of Edinburgh

ben.krause@ed.ac.uk, f.fancellu@sms.ed.ac.uk, bonnie@inf.ed.ac.uk
Conversational AI is everywhere

http://static4.uk.businessinsider.com/image/581ca089dd08954b518b45b6-1190-625/
we-put-siri-alexa-google-assistant-and-cortana-through-a-marathon-of-tests-to-see-whos-winning-
the-virtual-assistant-race--heres-what-we-found.jpg
2016: The year of the chatbot

from ‘Tracxn Research, Chatbot Startup Landscape’, June 2016
Chatbot Applications

- Customer service
- IoT
- Other: help people with disabilities, etc.
Amazon vs. Google vs. Microsoft

https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E
https://www.bhphotovideo.com/images/images2500x2500/google_ga3a00417a14_home_1297281.jpg
Amazon Alexa Prize

The Alexa Prize
$2.5 Million to Advance Conversational Artificial Intelligence
September 2016 – November 2017

- **Goal:** to build on *open-domain* conversation AI for *commercial purposes*
  - Currently, Alexa mostly is mostly rule-based (*skills*)
- 18 teams involved (12 sponsored by Amazon)
- Users in the U.S. evaluate the conversation with bot on a scale from 1 to 5
Our team

**Team Edina**  
(Undiversity of Edinburgh)
The problem(s)
Where do we start?

- How do we build a chatbot?
  - No idea!
  - Let’s look at previous work!
Rule-based bots: Mitsuku (try it at mistuku.com!)

Human: How are you?
Mitsuku: Everything is going extremely well. How are you?
  Human: I am good, thanks for asking
Mitsuku: No problem. Glad to hear it. What's new?
  Human: Nothing really, just writing a presentation about you
Mitsuku: Fair enough. What do you want to talk about?
  Human: Do you like Taylor Swift?
Mitsuku: Yes she has a very nice voice.
  Human: How old is she?
  Human: Where is she from?
Mitsuku: I'm not sure. She didn't tell me her address.
  Human: Have you seen the new Wonder Woman movie?
Mitsuku: I didn't even realise they had made a new one.
  Human: Have you seen the new X-men movie?
Mitsuku: I didn't even realise they had made a new one.
  Human: Have you seen Hidden Figures?
Mitsuku: No I have never seen one. What is it like?
  Human: It is a movie
Mitsuku: Not if I can help it.
Rule-based vs. Machine-learning

- Rule-based
  - ✔ Fully deterministic
  - ✔ Output fully intelligible
  - ✗ Very constrained
  - ✗ Time-consuming, Difficult to maintain
  - ✗ Full of fallback strategies
A Network-based End-to-End Trainable Task-oriented Dialogue System

Generative Deep Neural Networks for Dialogue: A Short Review

A Deep Reinforcement Learning Chatbot
Rule-based vs. Machine-learning

- **Rule-based**
  - ✔️ Fully deterministic
  - ✔️ Output fully intelligible
  - ✗ Very constrained
  - ✗ Time-consuming, Difficult to maintain
  - ✗ Full of fallback strategies

- **Machine-learning**
  - ✔️ Easy to maintain
  - ✔️ Flexible, broader-coverage
  - ✗ Non-deterministic
  - ✗ Constrained to the domain of the training data
Where do we start?

- How do we build a chatbot?
  - No idea!
  - Let’s look at previous work!

- What does Amazon want?
  - Open-domain
  - The user needs to be happy!!!
Open-domain

For topics, you should expect your socialbot to handle a variety of topics from high level subject matter domains including Entertainment, Fashion, Politics, Sports, and Technology. Each domain may contain categories like baseball playoffs, celebrity gossip, political events, or scientific breakthroughs.

As you can imagine, the topics that Alexa customers will want to chat about may change over time. The only guidance we will provide to customers wanting to engage with Alexa Prize socialbots will be to stay within the five high level topical domains mentioned above. Within those domains, customers may wish to engage in conversations on longer-term topics like self-driving cars or immigration reform, as well as fluid or short-lived topics that are representative of trending news items or their particular interests.
Rule-based vs. Machine-learning

- **Rule-based**
  - ✔️ Fully deterministic
  - ✔️ Output fully intelligible
  - ✗ Very constrained
  - ✗ Time-consuming, Difficult to maintain
  - ✗ Full of fallback strategies

- **Machine-learning**
  - ✔️ Easy to maintain
  - ✔️ Flexible, broader-coverage
  - ✗ Non-deterministic
  - ✗ Constrained to the domain of the training data
Rule-based vs. Machine-learning

- **Rule-based**
  - ✔ Fully deterministic
  - ✔ Output fully intelligible
  - ✗ Very constrained
  - ✗ Time-consuming, Difficult to maintain
  - ✗ Full of fallback strategies

- **Machine-learning**
  - ✔ Easy to maintain
  - ✔ Flexible, broader-coverage
  - ✗ Non-deterministic
  - ✗ Constrained to the domain of the training data
Corpora

- **OpenSubtitles**: Crowdsourced movie subtitles
- **Movie Scripts from IMDB**
- **Fisher**: Phone conversations
- **Ubuntu Dialogue Corpus**: Technical support for Ubuntu-related problems

No corpora that fit our domain!!!
The user needs to be happy

- Avoid offensive language
- Avoid sensitive topics (politics, religion, sex)
- Be empathetic
Be empathetic

[User is looking for info on a stage 4 cancer]

BOT: Who is your favorite actor?
USER: <says they have cancer>
BOT: I like a few of his other films.
USER: <mentions cancer again>
BOT: Me too! I like him in the notebook.
Rule-based vs. Machine-learning

- Rule-based
  - ✔ Fully deterministic
  - ✔ Output fully intelligible
  - ❌ Very constrained
  - ❌ Time-consuming, Difficult to maintain
  - ❌ Full of fallback strategies

- Machine-learning
  - ✔ Easy to maintain
  - ✔ Flexible, broader-coverage
  - ❌ Non-deterministic
  - ❌ Constrained to the domain of the training data
What is ideal?

- A model that...
  - mostly machine-learning based
  - feeds on clean data that is relevant to the task (what and how the user wants it!)
  - maintainable from an engineering and financial perspective
  - outputs intelligible responses
What is ideal?

- A model that...
  - mostly machine-learning based
  - feeds on clean data that is relevant to the task (what and how the user wants it!)
  - maintainable from an engineering and financial perspective
  - outputs intelligible responses
Ask people!

- If you want to know *what* do people talk about and *how* they do it, ask people.
- Two people conversing with each other on a topic
Ask people the Turkers!

- Crowdsourcing platform
- Create and upload a task (e.g. ‘have a conversation with another user on a topic’)
- Have people around the world solve the task
- Collect data

https://pbs.twimg.com/profile_images/661394940816035840/1R9_KPHN.png
Visual Dialogue (Abhishek et al., 2016)
However…

- Having two turkers to chat with each other requires good timing and a <strong>common ground</strong> (the image in VisDial)

E.g.
A: Hey, have you seen Guardians of the Galaxy?
B: No
A: Not your type I guess.
B: Have you?
A: I have
B: Sounds nice

- Costs double (when people two people at a time)
Self-dialogues

The Turker makes up a fictitious conversation

---

**Task and instructions**

Assume there are two users talking to each other about **NFL football**. Your job is to pretend to be both people and hold a conversation:

**Note** - The topic must be "NFL football". Keep each turn to 1-2 sentences. Avoid bad, political, sexist and inappropriate language and avoid IM/chat abbreviations.

**Example:**

**User1**: Hey are you excited for the upcoming NFL season?

**User2**: Yeah for sure! What team do you think will win it all?

**User1**: I think the patriots are going to repeat.

**User2**: Tom Brady is getting old, I think this year will belong to the Seahawks.

**User1**: I don't know about that, the Seahawks haven't been the same without Marshawn Lynch.

[Including some facts in the conversations are a great touch!]

etc...

**Avoid bad, political, sexist and inappropriate language.**

Now it’s your turn. Fill in for both users for 10 turns. Incomplete submissions will be rejected. You can do up to 20 of these; though obvious copies and exploitations will be rejected.
What is your absolute favorite movie?
I think Beauty and the Beast is my favorite.
The new one?
No, the cartoon. Something about it just feels magical.
It is my favorite Disney movie.
What’s your favorite movie in general?
I think my favorite is The Sound of Music.
Really? Other than cartoons and stuff I can never get into musicals.
I love musicals. I really liked Phantom of the Opera.
Self-dialogues, cont’d

- **Speed and set-up**: takes less effort and waiting time to gather data from a single user
- **Cost effectiveness**: halves the cost; after an initial bulk, only sporadic updates to keep on track with trendy topics
- **Quality**: the user is always an expert in what is talking about; knows about the entities introduced in the dialogues
- **Naturalness**: the flow conversation is natural
- **Not 2-person conversations**: further analysis (dialogue acts etc.) are hindered
Data collected

- A peak of 2,307 conversations a day
- **Total cost:** US $17,947.54

You need a lot of $$$ for these tasks!
Data collected, cont’d

<table>
<thead>
<tr>
<th>Topic/subtopic</th>
<th># Conversations</th>
<th># Words</th>
<th># Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movies</td>
<td>4,126</td>
<td>814,842</td>
<td>82,018</td>
</tr>
<tr>
<td>Action</td>
<td>414</td>
<td>37,037</td>
<td>4,140</td>
</tr>
<tr>
<td>Comedy</td>
<td>414</td>
<td>36,401</td>
<td>4,140</td>
</tr>
<tr>
<td>Fast &amp; Furious</td>
<td>343</td>
<td>33,964</td>
<td>3,430</td>
</tr>
<tr>
<td>Harry Potter</td>
<td>414</td>
<td>44,220</td>
<td>4,140</td>
</tr>
<tr>
<td>Disney</td>
<td>2,331</td>
<td>232,573</td>
<td>23,287</td>
</tr>
<tr>
<td>Horror</td>
<td>414</td>
<td>428,331</td>
<td>4,138</td>
</tr>
<tr>
<td>Thriller</td>
<td>828</td>
<td>77,975</td>
<td>8,277</td>
</tr>
<tr>
<td>Star Wars</td>
<td>1,726</td>
<td>178,351</td>
<td>17,260</td>
</tr>
<tr>
<td>Superhero</td>
<td>414</td>
<td>40,967</td>
<td>4,140</td>
</tr>
<tr>
<td>Music</td>
<td>4,911</td>
<td>924,993</td>
<td>98,123</td>
</tr>
<tr>
<td>Pop</td>
<td>684</td>
<td>62,383</td>
<td>6,840</td>
</tr>
<tr>
<td>Rap / Hip-Hop</td>
<td>684</td>
<td>66,376</td>
<td>6,840</td>
</tr>
<tr>
<td>Rock</td>
<td>684</td>
<td>63,349</td>
<td>6,837</td>
</tr>
<tr>
<td>The Beatles</td>
<td>679</td>
<td>68,396</td>
<td>6,781</td>
</tr>
<tr>
<td>Lady Gaga</td>
<td>558</td>
<td>49,313</td>
<td>5,566</td>
</tr>
<tr>
<td>Music and Movies</td>
<td>216</td>
<td>37,303</td>
<td>4,320</td>
</tr>
<tr>
<td>NFL Football</td>
<td>2,801</td>
<td>562,801</td>
<td>55,939</td>
</tr>
</tbody>
</table>
The system
System overview
A deterministic queue

- Queue of components: when a component fails, the next one is called
  1. **EVI**: a factoid Q&A component provided by Amazon
  2. **Rule-based**: deals with general chit-chat
  3. **Edina’s likes and dislikes**: a bit of personality (based on Wiki views)
  4. **Matching score**: our main component. Retrieves the most-likely answer from the self-dialogue database.
  5. **Proactive**: change the topic on its own volition
  6. **Neural network**: A generative neural network kicks in if everything else fails.

<table>
<thead>
<tr>
<th></th>
<th>RB</th>
<th>MS</th>
<th>NN</th>
<th>EVI</th>
<th>LD</th>
<th>Proactive</th>
<th>Backup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.69%</td>
<td>46.29%</td>
<td>20.55%</td>
<td>0.47%</td>
<td>0.37%</td>
<td>16.56%</td>
<td>0.08%</td>
</tr>
</tbody>
</table>
A deterministic queue

- Queue of components: when a component fails, the next one is called

1. **EVI**: a factoid Q&A component provided by Amazon
2. **Rule-based**: deals with general chit-chat
3. **Edina’s likes and dislikes**: a bit of personality (based on Wiki views)
4. **Matching score**: our main component. Retrieves the most-likely answer from the self-dialogue database.
5. **Proactive**: change the topic on its own volition
6. **Neural network**: A generative neural network kicks in if everything else fails.
Rule-based

- **Bot’s identity**: anonymized until the finals
- **Edina’s favorites**: favorite actor, artist, singer, etc.
- **Sensitive topics**: suicide, cancer, death as well as prompts containing offensive contents that needed to be ‘gracefully’ caught
- **Topic shifting**: deals with requests of topic shifting
- **Games and jokes**
- **+ a set of the most frequent prompts from Alexa users**, provided by Amazon
Our main component

Matches a user query $q$ with the conversation contexts $c$ of all potential responses from the pool of self-dialogues gathered through AMT, to return the most likely response $r$ (and a confidence score).

E.g.

$q$: Have you seen Hidden Figures?

$c_{-2}$: Any cool new movie?

$c_{-1}$: What about Hidden Figures?

$r$: I thought Hidden Figures was very thin on the actual mathematics of it all. S: 0.87
The matching score is an interpolation of bag-of-words, IDF-based scores (rare words are upweighted).

\[
S(q, r_i, c_i) = \frac{(S^c + S^{cr})(S^c)^n}{\eta} + \lambda S^{2cq} \quad (1)
\]

where \(S^c, S^{cr}, (S^c)^n\) and \(S^{2cq}\) are subscores and \(\lambda, \eta\) and \(n\) are constants.
Neural network

- Language model with multiplicative LSTM (Krause et al., 2017)
- Trained on OPENSUBTITLES and fine-tuned on our data
Evaluation
Evaluation

- Evaluating the usefulness of the matching score
- Qualitative evaluation
- Evaluations we haven’t done but we would like to do
Evaluating the matching score

- We sample conversation triplets from our self-dialogue pool
  A: It is my favorite Disney movie.
  B: I think Beauty and the Beast is my favorite.
  A: I think my favorite is ‘Sound of Music’

- We manually score the actual reply against what the matching score returns.
Evaluating the matching score - Results

<table>
<thead>
<tr>
<th>Response type</th>
<th>Mean human evaluation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human AMT worker</td>
<td>4.09 ± 0.15</td>
</tr>
<tr>
<td>Matching score top 4, confidence &gt; 0.7</td>
<td>3.60 ± 0.07</td>
</tr>
<tr>
<td>Matching score top-4, confidence &gt; 0.5</td>
<td>3.50 ± 0.08</td>
</tr>
<tr>
<td>Matching score top-4</td>
<td>3.08 ± 0.09</td>
</tr>
</tbody>
</table>

Corr.: $r = 0.259$ ($p < 0.0005$)
Qualitative evaluation of conversation

B: What bands have you seen live?
H: I’ve seen Sufjan Stevens, do you like them?
B: What’s he like?
H: Quite boring to be honest
B: I know what you are talking about.
H: Right?! What kind of bands have you seen live?
B: Yeah! I saw Cake a few months ago, and just last week I saw The Protomen.
H: Really! How was Cake?
B: Eating cake?
H: No i’m not eating a cake
B: No, the band. Have you ever heard of them?
H: No i haven’t
B: She does more traditional type music, but I absolutely love the way she sings.
H: I’ll have to check her out
B: You should! she’s the best singer for sure.
Things we haven’t done but we would have loved to do

- Assessing whether the entities mentioned in the self-dialogues reflect the entities Alexa users talk about
- **Using the scores from Alexa users to tune our system**

Evaluating open-domain chatbots is difficult!
Conclusion
Conclusion

▶ Open-domain conversational AI is hard and still a (very) open problem
▶ **Data collection/annotation** is a real challenge, but **self-dialogues** are efficient and surprisingly effective.
▶ **Evaluation** is still an open problem
▶ An **hybrid-system** is a reasonable solution for this challenge (rule-based x machine-learning x IR)
Final remarks

- We got 6th place (out of 15 teams)
- ...despite being the underdogs of the competition
- **Teamwork is hard but pays off!**
What’s next?

The Alexa Prize
$2.5 Million to Advance Conversational Artificial Intelligence
Since November 1, 2017

2018
YOU, YOU AND YOU

MSC PROJECTS FOR EVERYONE

https://imgflip.com/i/1yie8f