CS388: Natural Language Processing

Lecture 23: Dialogue

Greg Durrett

Credit: Megha Desai on Medium

Administrivia

› Project 2 back next week

› Two talks coming up:
  › Friday, November 15 11am GDC 6.302: Yoav Artzi (Cornell Tech)
  › Friday, November 22 11am GDC 6.302: Jacob Andreas (MIT)

This Lecture

› Classical chatbots

› Neural chatbots

› Task-oriented dialogue

› QA as dialogue

Chatbots
Turing Test (1950)

- Imitation game: A and B are locked in rooms and answer C’s questions via typewriter. Both are trying to act like B

  Original Interpretation:

  ![Diagram of Turing Test]

  Standard Interpretation:

  ![Diagram of Turing Test]

- The test is not “does this computer seem human-like to random people with a web browser?”

ELIZA

- Created 1964-1966 at MIT, heavily scripted

  Original Interpretation:

  ![Original ELIZA Interaction]

  Standard Interpretation:

  ![Standard ELIZA Interaction]

- The test is not “does this computer seem human-like to random people with a web browser?”

Cleverbot

- Carpenter (1986), online system built in 2006

  Original Interpretation:

  ![Original Cleverbot Interaction]

  Standard Interpretation:

  ![Standard Cleverbot Interaction]

- The test is not “does this computer seem human-like to random people with a web browser?”

Weizenbaum (1966)
Data-Driven Approaches

- Can treat as a machine translation problem: “translate” from current utterance to next one
  
  what  
  time  
  you  
  get  
  out  
  ?  

- Filter the data, use statistical measures to prune extracted phrases to get better performance

Ritter et al. (2011)

Data-Driven Approaches

<table>
<thead>
<tr>
<th>Status</th>
<th>MT-CHAT</th>
<th>MT-BASELINE</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>riding to the</td>
<td>good luck on your</td>
<td>can't wait for it</td>
<td>make sure to ride</td>
</tr>
<tr>
<td>arc on my bike.</td>
<td>bike ride are.</td>
<td>I'm on are a year.</td>
<td>in the middle of the street</td>
</tr>
<tr>
<td>hopefully I won't</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>melt into the pavement.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I really hate watching movies in the living room.</td>
<td>me too, what movies are you doing in room?</td>
<td>me too! I love those? in my mouth.</td>
<td>Why? no comfy spots or just too open to public</td>
</tr>
<tr>
<td>Hello? Did I miss David? Getting his teen choice award? Or did they not show it?</td>
<td>hello? miss? * he was getting his will probably win.</td>
<td>it's just like what you just did the exact date hello?</td>
<td>nothing yet...he presented though he is so darn cute</td>
</tr>
</tbody>
</table>

Ritter et al. (2011)

Seq2seq models

- Just like conventional MT, can train seq2seq models for this task

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.33</td>
</tr>
<tr>
<td>MT</td>
<td>3.21</td>
</tr>
<tr>
<td>HUMAN</td>
<td>6.08</td>
</tr>
</tbody>
</table>

Hard to evaluate:
**Subtitles Data**

do you want to meet your sponsor for the last 10 years?
of course! but he doesn’t want to see me!

and where had you been just before?
i’d been to the palace of the legion of honor, the art gallery.

yeah, we were just going to hit up taco bell.
well, it’s my pleasure.

Why might this model perform poorly? What might it be bad at?

---

**Lack of Diversity**

Training to maximize likelihood gives a system that prefers common responses:

<table>
<thead>
<tr>
<th>Input: What are you doing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.86 I don’t know.</td>
</tr>
<tr>
<td>-1.03 I don’t know!</td>
</tr>
<tr>
<td>-1.06 Nothing.</td>
</tr>
<tr>
<td>-1.09 Get out of the way.</td>
</tr>
</tbody>
</table>

Why?

<table>
<thead>
<tr>
<th>Input: what is your name?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.91 I don’t know.</td>
</tr>
<tr>
<td>-0.92 I don’t know!</td>
</tr>
<tr>
<td>-0.92 I don’t know, sir.</td>
</tr>
<tr>
<td>-0.97 Oh, my god!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: How old are you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.79 I don’t know.</td>
</tr>
<tr>
<td>-1.06 I’m fine.</td>
</tr>
<tr>
<td>-1.17 I’m all right.</td>
</tr>
<tr>
<td>-1.17 I’m not sure.</td>
</tr>
</tbody>
</table>

---

**Solution:** mutual information criterion; response $R$ should be predictive of user utterance $U$ as well

Standard conditional likelihood: $\log P(R|U)$

Mutual information: $\log \frac{P(R, U)}{P(R)P(U)} = \log P(R|U) - \log P(R)$

$log P(R) = \text{probabilities under a language model}$

---

**OpenSubtitles data**

---

Li et al. (2016)
Specificity

- Train a specificity classifier on labeled data
  - I don’t know => SPECIFICITY=1
  - Going to the store => SPECIFICITY=3

- When training the decoder, condition on the automatically predicted specificity of the response
  - I don’t know [STOP] => SPECIFICITY=1 (nonspecific)
  - Going to the store [STOP] => SPECIFICITY=4 (specific)

- At test time, set the specificity level higher to get less generic responses

Can use other models to try to fix these issues. But the facts are still all made up, even if they make sense.

PersonaChat

<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 love watching Game of Thrones</td>
</tr>
<tr>
<td>I like to eat cheetos</td>
<td></td>
</tr>
</tbody>
</table>

Zhang et al. (2018)
State of Chatbots

‣ Can force chatbots to give consistent answers with a persona, but still probably not very interesting

‣ “Wizard of Wikipedia:” chatbot that can discuss topics by retrieving from Wikipedia [Dinan et al., 2019]

‣ XiaoIce: Microsoft chatbot in Chinese, 20M users, average user interacts 60 times/month…people do seem to like talking to them…?

Task-Oriented Dialogue

Task-Oriented Dialogue

‣ Personal assistants / API front-ends:

  - Siri, find me a good sushi restaurant in Chelsea
  - Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google
  - How expensive is it?
  - Entrees are around $30 each
  - Find me something cheaper

 Task-Oriented Dialogue

‣ Customer service:

  - Hey Alexa, why isn’t my Amazon order here?
  - Let me retrieve your order. Your order was scheduled to arrive at 4pm today.
  - It never came
  - Okay, I can put you through to customer service.
Air Travel Information Service (ATIS)

- Given an utterance, predict a domain-specific semantic interpretation

<table>
<thead>
<tr>
<th>Utterance</th>
<th>How much is the cheapest flight from Boston to New York tomorrow morning?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal:</td>
<td>Airfare</td>
</tr>
<tr>
<td>Cost:</td>
<td>cheapest</td>
</tr>
<tr>
<td>Depart:</td>
<td>Boston</td>
</tr>
<tr>
<td>Arrive:</td>
<td>New York</td>
</tr>
<tr>
<td>Depart:</td>
<td>tomorrow</td>
</tr>
</tbody>
</table>

- Can formulate as semantic parsing, but simple slot-filling solutions (classifiers) work well too

DARPA (early 1990s), Figure from Tur et al. (2010)

Intents

- 29 different intents
- which flights go from cleveland to indianapolis on april fifth
  - Intent: flight
- does tacoma airport offer transportation from the airport to the downtown area
  - Intent: ground_service
- what days of the week do flights from san jose to nashville fly on
  - Intent: day_name
- what meals are served on american flight 811 from tampa to milwaukee
  - Intent: meal

Joint Intent Classification and Tagging

- RNN jointly predicts intent and slot tags

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
<th>Intent Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RecNN [8]</td>
<td>93.22</td>
<td>4.60</td>
</tr>
<tr>
<td>RecNN+Viterbi [8]</td>
<td>93.96</td>
<td>4.60</td>
</tr>
<tr>
<td>Attention Encoder-Decoder NN (with aligned inputs)</td>
<td>95.87</td>
<td>1.57</td>
</tr>
<tr>
<td>Attention BiRNN</td>
<td>95.98</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Liu and Lane (2016)

Air Travel Information Service (ATIS)

- Need to use dialogue context to do the right thing. Here we’re appending American Airlines as a constraint to the previous query
- seq2seq model mapping to query with copy mechanism

Suhr et al. (2018)
**Full Dialogue Task**

- Parsing / language understanding is just one piece of a system
- Dialogue state: reflects any information about the conversation (e.g., search history)
- User utterance -> update dialogue state -> take action (e.g., query the restaurant database) -> say something
- Much more complex than chatbots!

Young et al. (2013)

---

**POMDP-based Dialogue Systems**

- POMDP: user is the “environment,” an utterance is a noisy signal of state
- Dialogue model: can look like a parser or any kind of encoder model
- Generator: use templates or seq2seq model
-Where do rewards come from?

Young et al. (2013)

---

**Full Dialogue Task**

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

*Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google*

How expensive is it?

```
get_value(cost, curr_result)
```

Entrees are around $30 each

---

**Reward for completing task?**

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

*Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google*

How expensive is it?

```
+1 make_reservation(curr_result)
```

Okay make me a reservation!
User gives reward?

Find me a good sushi restaurant in Chelsea

restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How does the user know the right search happened?

+1

How expensive is it?

get_value(cost, curr_result)

+1 Entrees are around $30 each

Wizard-of-Oz

› Learning from demonstrations: “wizard” pulls the levers and makes the dialogue system update its state and take actions

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

wizard enters these

restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

› Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do

Learning from Static Traces

› Using either wizard-of-Oz or other annotations, can collect static traces and train from these

Kelley (early 1980s), Ford and Smith (1982)
Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```r
restaurant_type <- sushi
location <- Chelsea
stars <- 4+
curr_result <- execute_search()
```

- User asked for a “good” restaurant — does that mean we should change our model to filter by star rating? What does “good” mean?
- Hard to change system behavior if training from static traces, especially if system capabilities or desired behavior change

Goal-oriented Dialogue

- Tons of industry interest!
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Facebook, Samsung Bixby, Tencent WeChat, ASAPP
- Lots of cool work that’s not public yet

Other Dialogue Applications

Search/QA as Dialogue

- “Has Chris Pratt won an Oscar?” / “Has he won an Oscar”
Dialogue is a very natural way to find information from a search engine or a QA system.

Challenges: hard to annotate good dialogue datasets in a purely static way.

UW QuAC dataset: Question Answering in Context

Conversational machine reading: answer repeated questions based on a passage.

Interesting and potentially useful idea, but annotating data is very hard!

Error analysis
System
Better model
Data
Harder Data
Fixed distribution (e.g., natural language sentences), error rate $\to 0$

Error rate $\to ???$; “mission creep” from HCI element
Dialogue Mission Creep

- High visibility — your product has to work really well!

Takeaways

- Some decent chatbots, but unclear how to make these more sophisticated than they are right now
- Task-oriented dialogue systems are growing in scope and complexity — really exciting systems on the way
- More and more problems are being formulated as dialogue — interesting applications but challenging to get working well