Data for Task-Oriented Dialogue

Data

- How do you get training data when you don’t have a working dialogue system?
- Somehow need to annotate what should happen in response to user utterances. But then you need to know how those users would respond...

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
Intent=restaurant
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

```
Intent=restaurant
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google
```

Very indirect signal of what should happen up here

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

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

\[
\text{Intent}=\text{restaurant} \\
\text{restaurant\_type} \leftarrow \text{sushi} \\
\text{location} \leftarrow \text{Chelsea} \\
\text{curr\_result} \leftarrow \text{execute\_search()} \\
\text{Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google}
\]

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

\[
\text{Intent}=\text{restaurant} \\
\text{restaurant\_type} \leftarrow \text{sushi} \\
\text{location} \leftarrow \text{Chelsea} \\
\text{curr\_result} \leftarrow \text{execute\_search()} \\
\text{stars} \leftarrow 4+ \\
\text{curr\_result} \leftarrow \text{execute\_search()}
\]

- User asked for a “good” restaurant — does that mean we should 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
ATIS

Air Travel Information Service (ATIS)

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

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Goal:</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much is the cheapest flight from Boston to New York tomorrow morning?</td>
<td>Airfare</td>
</tr>
<tr>
<td>Cost.Relative</td>
<td>cheapest</td>
</tr>
<tr>
<td>Depart.City</td>
<td>Boston</td>
</tr>
<tr>
<td>Arrival.City</td>
<td>New York</td>
</tr>
<tr>
<td>Depart.Date.Relative</td>
<td>tomorrow</td>
</tr>
<tr>
<td>Depart.Time.Period</td>
<td>morning</td>
</tr>
</tbody>
</table>

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)
Suhr et al. (2018) 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) Detect and anonymize entities for better performance

Goal-oriented Dialogue

Suhr et al. (2018) Tons of industry interest!

Suhr et al. (2018) Dozens of startups + medium-sized companies in this space

Suhr et al. (2018) Big Companies: Apple Siri (VocalIQ), Google Allo, Amazon Alexa, Microsoft Cortana, Facebook M, Samsung Bixby, Tencent WeChat

Suhr et al. (2018) Lots of cool work that’s not public yet
Alexa Skills

- Let you add functionality to Amazon Alexa
- Can deploy to Alexa devices, develop and debug through AWS (no device necessary)
- Plugs into Amazon’s ASR and TTS, so no need to wrestle with these services yourself

More information: EE596 from UW

Slide credit: Hao Fang / Hao Cheng (UW)

Search/QA as Dialogue

- “Has Chris Pratt won an Oscar?” / “Has he won an Oscar”
**QA as Dialogue**

- Dialogue is a very natural way to find information from a search engine or a QA system
- Challenges:
  - QA is hard enough on its own
  - Users move the goalposts

```
<table>
<thead>
<tr>
<th>Original Intent: What super hero from Earth appeared most recently?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
</tr>
<tr>
<td>--------------------</td>
</tr>
<tr>
<td>Night Girl</td>
</tr>
<tr>
<td>Dragonwing</td>
</tr>
<tr>
<td>Gates</td>
</tr>
<tr>
<td>X3</td>
</tr>
<tr>
<td>Harmonia</td>
</tr>
</tbody>
</table>
```

Iyyer et al. (2017)

**Search as Dialogue**

- Google can deal with misspellings, so more misspellings happen — Google has to do more!

**Dialogue Mission Creep**

- Most NLP tasks:
  - Error analysis
  - System
  - Data
  - Better model

- Data:
  - Harder Data

- Error rate -> ??
  - “mission creep” from HCI element
High visibility — your product has to work really well!