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...
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
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?
...
Okay make me a reservation!

```
+1 make_reservation(curr_result)
```
User gives reward?

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
Wizard-of-Oz

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

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

Find me a good sushi restaurant in Chelsea

wizard enters these
\[
\begin{align*}
\text{Intent} &= \text{restaurant} \\
\text{restaurant\_type} &\leftarrow \text{sushi} \\
\text{location} &\leftarrow \text{Chelsea} \\
\text{curr\_result} &\leftarrow \text{execute\_search()}
\end{align*}
\]

wizard types this out or invokes templates
\[
\text{Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google}
\]

- Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do
Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```plaintext
Intent=restaurant
restaurant_type <- sushi
location <- Chelsea
stars <- 4+
curr_result <- 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>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_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)
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</td>
<td>95.87</td>
<td>1.57</td>
</tr>
<tr>
<td>(with aligned inputs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention BiRNN</td>
<td>95.98</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Liu and Lane (2016)
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

Common trick in QA models (discussed next week)
Air Travel Information Service (ATIS)

<table>
<thead>
<tr>
<th>Model</th>
<th>Query</th>
<th>Denotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Relaxed</td>
</tr>
<tr>
<td><strong>Development Results</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEQ2SEQ-0</td>
<td>28.7±1.7</td>
<td>48.8±1.4</td>
</tr>
<tr>
<td>SEQ2SEQ-H</td>
<td>35.1±2.2</td>
<td>59.4±2.4</td>
</tr>
<tr>
<td>S2S+ANON</td>
<td>37.6±0.7</td>
<td>61.6±0.7</td>
</tr>
<tr>
<td>FULL-0</td>
<td>36.3±0.5</td>
<td>61.5±0.8</td>
</tr>
<tr>
<td>FULL</td>
<td>37.5±0.9</td>
<td>63.0±0.7</td>
</tr>
</tbody>
</table>

- H has access to previous utterances
- Denotation: right answer but not exact match on query

Suhr et al. (2018)
Goal-oriented Dialogue

- Tons of industry interest!
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri (VocalIQ), Google Allo, Amazon Alexa, Microsoft Cortana, Facebook M, Samsung Bixby, Tencent WeChat
- 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
Alexa Skills

- Instantiate skill, intents, and slots
- Custom slot types let you define values
- To do really complicated things, need phrase slots — this actually gives you the text
- More information: EE596 from UW

Slide credit: Hao Fang / Hao Cheng (UW)
Other Dialogue Applications
“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>Challenge</th>
<th>Legion of Super Heroes Post-Infinite Crisis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Character: Harmonia 2011 Home World: Earth Powers: Elemental</td>
</tr>
</tbody>
</table>

Iyyer et al. (2017)
QA as Dialogue

- UW QuAC dataset: Question Answering in Context

Choi et al. (2018)
Google can deal with misspellings, so more misspellings happen — Google has to do more!
Most NLP tasks

- Error analysis
- Better model
- Data

- System

Dialogue/Search/QA

- Error analysis
- Better model
- Data

- System

- Harder Data

- Error rate -> ???; “mission creep” from HCI element

- Fixed distribution (e.g., natural language sentences), error rate -> 0
High visibility — your product has to work really well!