Speech, speech recognition, and dialog state tracking in spoken dialog systems
(and maybe situated language systems too)

Jason D. Williams
Microsoft Research
Quick intro

• After undergrad & masters...

• Worked for 3 years building spoken dialog systems in industry
  • Ran 100 usability tests
  • Listened to 1000s of calls

• PhD in statistical approaches to spoken dialog systems

• AT&T Research for 6 years
  • Also deployed 2 large-scale systems

• Now at MSR (1.5 years)
Aims for today

• Introduce spoken dialog systems
  • Emphasis on issues introduced by automatic speech recognition (ASR)
  • Directed at practitioners, not ASR researchers

• Present some techniques for state tracking in dialog systems
  • Improve robustness to ASR errors
  • And may also open doors to collaboration with the dialog community
Spoken dialog systems

```
“28x to the airport”
```

```
Update dialog state
```

```
Choose action
```

```
 route: 29X
destination: airport
certainty: 0.7
```

```
F(s,o) → s'
```

```
G(s') → a
```

```
“Dialog state tracking”
“Belief tracking”
“Dialog modeling”
```

```
“Planning”
“Policy”
“Control”
```

```
“Route 29X?”
```

```
NLG + TTS
```

```
act: confirm
route: 29X
```

```
```

```
```

```
```

```
```

```
```
Dialog state tracking

*Dialog state tracking is difficult largely because errors in speech recognition and language understanding are error-prone, and render the true state of the dialog partially observable.*
Responses to "How may I help you?"

- Silences and hesitations while users think
  - Leads to end-pointing problems
  - Leads to users confusing themselves
- "Robot" language (hence examples, "speak naturally")
  - Example 1
  - Example 2
- Recognition errors confused with competences
  > "i need to sign up for a get off benefit" [no parse]
  > "i would like to enroll in a get one" [no parse]
  > "i would like to get help with my dental insurance" <HELP>
  > "dental insurance" <INSURANCE>

Source: Live calls, human resources dialog system
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
<tr>
<td>% in-grammar/in-domain</td>
<td>92.3%</td>
<td>91.0%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>92.1%</td>
<td>77.6%</td>
<td>77.7%</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR errors are hard to detect

Source: Commercial deployed spoken dialog system
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
<tr>
<td>% in-grammar/in-domain</td>
<td>92.3%</td>
<td>91.0%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>92.1%</td>
<td>77.6%</td>
<td>77.7%</td>
</tr>
<tr>
<td>Accepted utts (False accepts)</td>
<td>89.6%</td>
<td>60.3%</td>
<td>73.3%</td>
</tr>
<tr>
<td></td>
<td>(1.8%)</td>
<td>(4.9%)</td>
<td>(8.3%)</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
There is useful information on the N-Best list

Jason D. Williams, Challenges and Opportunities for State Tracking in Statistical Spoken Dialog Systems: Results From Two Public Deployments, in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 8, pp. 959-970, IEEE SPS, December 2012
Users really care about recognition accuracy

• 3 spoken dialog systems: voice dialing and messaging, email, train schedules

• User satisfaction scores at end of each dialog

• Build regression models that predict user satisfaction as a function of dialog properties – main 3 are:
  • Speed, task completion, recognition accuracy

• Across all 3 systems, the results were the same:
  • Most important: Reco accuracy
  • Second more important: task completion
  • Distant third: speed

Example: Rule-based state tracker

- In commercial systems, dialog state tracking is most commonly done with hand-coded rules

- Example:

```python
if ($slu[0].score > 0.8):
    // keep SLU result
    $state.bus_route = $slu[0].bus_route
else:
    // discard SLU result
```
Can we do better?

• Toeholds:
  • Rules usually make decisions are made locally, but turns are correlated
  • N-Best lists contain information that most rules ignore
  • There are useful priors that rules usually don’t use

• Intuition
  • Cast state as a hidden variable and evidence as noisy emission from that state
  • View dialog state estimation as a tracking or filtering problem
  • Similar to particle filters or HMMs

• Methods
  • Generative models: partitions, bayes nets, more
  • Discriminative: maxent, SVMs, ANNs, more
  • More: heuristics
Tracking a distribution over interaction states

User goal
Interaction history
User's action
System's action
ASR/NLU result
Generative approach

\[ b'(g', h') = \eta \cdot \sum_{u} P(o' | u', a) \sum_{h} P(u' | g', h, a) P(h' | u', g', h, a) \sum_{g} P(g' | g, a) b(g, h) \]

- **Synthesizes:**
  - Current belief over user's goal \( b(g) \)
  - Model of how user goal changes
  - Model of user producing utterances
  - ASR/SLU confidence

AT&T Let's Go Call Visualizer

to airport. Is that right?

ASR Result: yes that's correct
Origin: carnegie mellon in oakland; carnegie mellon university
Destination: to airport; airport; to the airport; the airport

Bus Route: twenty eight x; the twenty eight x
Date
Time

Do you want times for the next few buses? Say yes or no.
Why aren’t more people working on this?

• Building dialog systems is expensive -- difficult to obtain dialog data
• Can this be cast as a challenge task?

http://research.microsoft.com/en-us/events/dstc/

• 15K dialogs with real users
• Transcriptions, labels
• Common log format, with user and system semantics
• 3 totally different dialog systems
• Evaluation tools; baselines; template code

## Dialog state tracking challenge data sets

<table>
<thead>
<tr>
<th></th>
<th>Train 1A</th>
<th>Train 1B</th>
<th>Train 1C</th>
<th>Train 2</th>
<th>Train 3</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year(s)</td>
<td>2009</td>
<td>2009</td>
<td>2009</td>
<td>2010</td>
<td>2010</td>
<td>2011</td>
<td>2012</td>
<td>2011-2</td>
<td>2010</td>
</tr>
<tr>
<td>Dialogs</td>
<td>1013</td>
<td>1117</td>
<td>9502</td>
<td>643</td>
<td>688</td>
<td>715</td>
<td>750</td>
<td>1020</td>
<td>438</td>
</tr>
<tr>
<td>Turns/Dialog</td>
<td>14.7</td>
<td>13.3</td>
<td>14.5</td>
<td>14.5</td>
<td>12.6</td>
<td>14.1</td>
<td>14.5</td>
<td>13.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Sys acts/turn</td>
<td>4.0</td>
<td>3.8</td>
<td>3.8</td>
<td>4.0</td>
<td>8.4</td>
<td>2.8</td>
<td>3.2</td>
<td>8.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Av N-best len</td>
<td>21.7</td>
<td>22.3</td>
<td>21.9</td>
<td>22.4</td>
<td>2.9</td>
<td>21.2</td>
<td>20.5</td>
<td>5.0</td>
<td>3.2</td>
</tr>
<tr>
<td>Acts/N-best hyp</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>1.0</td>
<td>2.1</td>
<td>2.0</td>
<td>1.0</td>
<td>1.6</td>
</tr>
<tr>
<td>Slots/turn</td>
<td>44.0</td>
<td>46.5</td>
<td>45.6</td>
<td>49.0</td>
<td>2.1</td>
<td>41.4</td>
<td>36.9</td>
<td>4.3</td>
<td>3.5</td>
</tr>
<tr>
<td>Transcribed?</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Labelled?</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>1-best WER</td>
<td>42.9%</td>
<td>41.1%</td>
<td>42.1%</td>
<td>58.2%</td>
<td>40.5%</td>
<td>57.9%</td>
<td>62.1%</td>
<td>48.1%</td>
<td>55.6%</td>
</tr>
<tr>
<td>1-best SLU Prec.</td>
<td>0.356</td>
<td>-</td>
<td>-</td>
<td>0.303</td>
<td>0.560</td>
<td>0.252</td>
<td>0.275</td>
<td>0.470</td>
<td>0.334</td>
</tr>
<tr>
<td>1-best SLU Recall</td>
<td>0.522</td>
<td>-</td>
<td>-</td>
<td>0.388</td>
<td>0.650</td>
<td>0.362</td>
<td>0.393</td>
<td>0.515</td>
<td>0.376</td>
</tr>
<tr>
<td>N-best SLU Recall</td>
<td>0.577</td>
<td>-</td>
<td>-</td>
<td>0.485</td>
<td>0.738</td>
<td>0.456</td>
<td>0.492</td>
<td>0.634</td>
<td>0.413</td>
</tr>
</tbody>
</table>
Results

9 teams participated (27 entries)

Graph shows accuracy of top-ranked dialog state

Results at SigDial (August, Metz, France – roughly co-located with Interspeech)
Integration point for situated language

• Observations can account for more than just speech sensor
  • Vision
  • Object attributes
  • Distance

• Hidden variable can account for more state than just user’s goal
  • State of objects in a world
  • State of a robot in a world
  • State of currently inferred semantics about the world

• An instantiated example of this: troubleshooting dialog system
Troubleshooting spoken dialog system

In some applications such as help-lines for DSL modem faults, there are additional sources of uncertainty. These can be easily incorporated into the Bayesian network. Eg

Applying POMDPs to dialog systems in the troubleshooting domain, Williams (Proc W'Shop Bridging the Gap, ACL, 2007)
DSL troubleshooting SDS as a Bayesian network

Hidden state components

- Service outage
- Upstream network failure
- Unknown, unfixable problem
- Correct username in browser
- Correct username on modem
- Correct password in browser
- Correct password on modem
- Correct service type in browser
- State of modem network light
- Correct service type on modem
- Config screen visible in browser
- Modem configuration is correct
- DSL connection is working
- State of modem power light
- User opened a webpage
- State of DSL modem
DSL troubleshooting
spoken dialog system
demonstration
Ideation: interactive image ID from attributes

**Dialog**

S: Is this an animal?
U: Yes ~ 0.9

**Attribute detectors**
- has-antlers: 0.9
- has-paws: 0.3
- jumps: 0.8
- is-furry: 0.7

**Inferred object class**
- is-deer: 0.4
- is-rabbit: 0.3
- is-moose: 0.6
Ideation: interactive image ID from attributes

**Attribute detectors**

- has-antlers: 0.9
- has-paws: 0.3
- jumps: 0.8
- is-furry: 0.7

**Inferred object class**

- is-deer: 0.5
- is-rabbit: 0.4
- is-moose: 0.7

**Dialog**

S: Is this an animal
U: Yes ~ 0.9

S: Is this a moose
U: ha! no way ~ 0.7
Ideation: interactive image ID from attributes

Attribute detectors
- has-antlers: 0.9
- has-paws: 0.3
- jumps: 0.8
- is-furry: 0.7

Inferred object class
- is-deer: 0.6
- is-rabbit: 0.5
- is-moose: 0.1

Dialog

S: Is this an animal
U: Yes ~ 0.9

S: Is this a moose
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9
Ideation: interactive image ID from attributes

**Attribute detectors**
- has-antlers: 0.9
- has-paws: 0.3
- jumps: 0.8
- is-furry: 0.7

**Inferred object class**
- is-deer: 0.1
- is-rabbit: 0.7
- is-moose: 0.1

**Dialog**
- S: Is this an animal?
  - U: Yes ~ 0.9
- S: Is this a moose?
  - U: ha! no way ~ 0.7
- S: Does it have hooves?
  - U: No ~ 0.9
- S: Is this a rabbit?
  - U: No, much more vicious ~0.6
Ideation: interactive image ID from attributes

Attribute detectors
has-antlers  0.9
has-paws  0.3
jumps  0.8
is-furry  0.7

Inferred object class
is-deer  0.1
is-rabbit  0.2
is-moose  0.1

Dialog

S: Is this an animal?
U: Yes ~ 0.9

S: Is this a moose?
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

S: Is this a rabbit?
U: No, much more vicious ~0.6

S: What kind of animal is it
U: Jackalope! ~ 0.7
Ideation: interactive image ID from attributes

Attribute detectors

- has-antlers: 0.9
- has-paws: 0.3
- jumps: 0.8
- is-furry: 0.7

Inferred object class

- is-deer: 0.1
- is-rabbit: 0.2
- is-moose: 0.1
- is-jackalope: 0.8

Dialog

S: Is this an animal?
U: Yes ~ 0.9

S: Is this a moose?
U: ha! no way ~ 0.7

S: Does it have hooves?
U: No ~ 0.9

S: Is this a rabbit?
U: No, much more vicious ~0.6

S: What kind of animal is it
U: Jackalope! ~ 0.7
Graphical model for conversations about objects and their attributes

- Object ID
- Object ID
- Attributes
- User action
- Detect or output
- System action
- ASR/SLU output
Summary

• Speech recognition errors are common and difficult to detect reliably
• Dialog systems that do not handle errors well are doomed to failure
• Techniques exist to improve the accuracy of dialog state tracking
• These techniques suggest integration points for building conversational systems in new and exciting domains:
  • Machine vision conversational systems
  • Conversational robots
  • Learning within the dialog
Thanks!

Speech, speech recognition, and dialog state tracking in spoken dialog systems

Jason D. Williams
Microsoft Research