Speech Processing 11-492/18-492

Spoken Dialog Systems Evaluation
How Good is Your System

What are the metrics?
- Task success
- Minimum call escalation
- Minimum call time
- Error recovery vs WER

What should I do to make it better
Spoken Dialog Challenge Goals

- **Improve Spoken Dialog Systems**
- **Community wide effort**
  - Accessible to new groups
- **Find common task**
  - Hard enough to be interesting
  - Easy enough that groups can build systems
- **Test systems on common user base**
  - Real users (not paid users)
- **Provide framework for better evaluation**
  - Not just task completion
Find bus schedules in Pittsburgh
  - CMU’s Let’s Go SDS
  - (120K live calls since 2005)

Answers calls via Port Authority
  - Weekdays 7pm-6am
  - Weekends 4:30-7am

Resources
  - Full source code for system
  - Full dialog logs
SDC Options

- **Build a complete system**
  - From provided system
  - Or from your own system
- **Test an evaluation technique**
- **Build a user simulator**
SDC Timeline

- Sep 2009 Announced
- Nov 2009 Web seminars
- Jan 2010 Data, system distributed
- Jan-Apr 2010 Development
- May 2010 Control Tests
- Jul-Aug 2010 Live Tests
Test Scenarios

Control Tests
- 4 systems (sys1-4)
- Around 100 calls
- Provided scenarios
- SDS Experts as callers
- 4 systems (sys1, sys2, sys3, sys4)

Live Tests
- 14 days each (evenings/weekend)
- Real callers who want bus information
- 3 systems (sys1, sys3, sys4)
Control Tests

Time you want to leave: 8:50 pm
You want: Any bus you can take before 9:45
Control Tests

- **Web based instructions**
  - To align calls and scenarios
- **8 calls per user**
- **Required international calls**
  - Remember time zones!
- **User pool is SDS researchers**
  - Unfamiliar with Pittsburgh buses
  - More non-native speech
  - Don’t care about next bus
  - Better/more forgiving to SDS
- **Finding callers is hard**
Measuring Success

- **Non-empty calls**
  - User might just hang up

- **No information**
  - Call ends before any information given

- **“I don’t have that”**
  - No bus for those arrival/departure point
  - No bus at that time
  - Result can be **correct** or **incorrect**

- **Positive Output**
  - There is a bus from X to Y at time Z
  - Result can be **correct** or **incorrect**
## Control Test

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS2</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Calls</td>
<td>91</td>
<td>61</td>
<td>75</td>
<td>83</td>
</tr>
<tr>
<td>no_info</td>
<td>3.3%</td>
<td>37.7%</td>
<td>1.3%</td>
<td>9.6%</td>
</tr>
<tr>
<td>donthave</td>
<td>17.6%</td>
<td>24.6%</td>
<td>14.7%</td>
<td>9.6%</td>
</tr>
<tr>
<td>donthave_corr</td>
<td>68.8%</td>
<td>33.3%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>donthave_incorr</td>
<td>31.3%</td>
<td>66.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>pos_out</td>
<td>79.1%</td>
<td>37.7%</td>
<td>84.0%</td>
<td>80.7%</td>
</tr>
<tr>
<td>pos_out_corr</td>
<td>66.7%</td>
<td>78.3%</td>
<td>88.9%</td>
<td>80.6%</td>
</tr>
<tr>
<td>pos_out_incorr</td>
<td>33.3%</td>
<td>21.7%</td>
<td>11.1%</td>
<td>19.4%</td>
</tr>
<tr>
<td></td>
<td>SYS1</td>
<td>SYS3</td>
<td>SYS4</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td><strong>Total Calls</strong></td>
<td>678</td>
<td>451</td>
<td>742</td>
<td></td>
</tr>
<tr>
<td><strong>Non-empty calls</strong></td>
<td>633</td>
<td>430</td>
<td>670</td>
<td></td>
</tr>
<tr>
<td><strong>no_info</strong></td>
<td>18.5%</td>
<td>14.0%</td>
<td>11.0%</td>
<td></td>
</tr>
<tr>
<td><strong>donthave</strong></td>
<td>26.4%</td>
<td>30.0%</td>
<td>17.6%</td>
<td></td>
</tr>
<tr>
<td><strong>donthave_corr</strong></td>
<td>47.3%</td>
<td>40.3%</td>
<td>37.3%</td>
<td></td>
</tr>
<tr>
<td><strong>donthave_incorr</strong></td>
<td>52.7%</td>
<td>59.7%</td>
<td>62.7%</td>
<td></td>
</tr>
<tr>
<td><strong>pos_out</strong></td>
<td>55.1%</td>
<td>56.0%</td>
<td>71.3%</td>
<td></td>
</tr>
<tr>
<td><strong>pos_out_corr</strong></td>
<td>86.8%</td>
<td>93.8%</td>
<td>91.6%</td>
<td></td>
</tr>
<tr>
<td><strong>pos_out_incorr</strong></td>
<td>13.2%</td>
<td>6.2%</td>
<td>8.4%</td>
<td></td>
</tr>
</tbody>
</table>
# Control vs Live Tests

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS2</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Calls</strong></td>
<td>91</td>
<td>61</td>
<td>75</td>
<td>83</td>
</tr>
<tr>
<td>no_info</td>
<td>3.3%</td>
<td>37.7%</td>
<td>1.3%</td>
<td>9.6%</td>
</tr>
<tr>
<td>donthave</td>
<td>17.6%</td>
<td>24.6%</td>
<td>14.7%</td>
<td>9.6%</td>
</tr>
<tr>
<td>donthave_corr</td>
<td>68.8%</td>
<td>33.3%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>donthave_incorr</td>
<td>31.3%</td>
<td>66.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>pos_out</td>
<td>79.1%</td>
<td>37.7%</td>
<td>84.0%</td>
<td>80.7%</td>
</tr>
<tr>
<td>pos_out_corr</td>
<td>66.7%</td>
<td>78.3%</td>
<td>88.9%</td>
<td>80.6%</td>
</tr>
<tr>
<td>pos_out_incorr</td>
<td>33.3%</td>
<td>21.7%</td>
<td>11.1%</td>
<td>19.4%</td>
</tr>
</tbody>
</table>

## Control

## Live

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Calls</strong></td>
<td>678</td>
<td>451</td>
<td>742</td>
</tr>
<tr>
<td>Non-empty calls</td>
<td>633</td>
<td>430</td>
<td>670</td>
</tr>
<tr>
<td>no_info</td>
<td>18.5%</td>
<td>14.0%</td>
<td>11.0%</td>
</tr>
<tr>
<td>donthave</td>
<td>26.4%</td>
<td>30.0%</td>
<td>17.6%</td>
</tr>
<tr>
<td>donthave_corr</td>
<td>47.3%</td>
<td>40.3%</td>
<td>37.3%</td>
</tr>
<tr>
<td>donthave_incorr</td>
<td>52.7%</td>
<td>59.7%</td>
<td>62.7%</td>
</tr>
<tr>
<td>pos_out</td>
<td>55.1%</td>
<td>56.0%</td>
<td>71.3%</td>
</tr>
<tr>
<td>pos_out_corr</td>
<td>86.8%</td>
<td>93.8%</td>
<td>91.6%</td>
</tr>
<tr>
<td>pos_out_incorr</td>
<td>13.2%</td>
<td>6.2%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>
## Call Lengths

<table>
<thead>
<tr>
<th></th>
<th>Length (s)</th>
<th>Turns/call</th>
<th>Words/turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYS1 control</td>
<td>155</td>
<td>18.29</td>
<td>2.87 (2.84)</td>
</tr>
<tr>
<td>SYS1 live</td>
<td>111</td>
<td>16.24</td>
<td>2.15 (1.03)</td>
</tr>
<tr>
<td>SYS2 control</td>
<td>147</td>
<td>17.57</td>
<td>1.63 (1.62)</td>
</tr>
<tr>
<td>SYS3 control</td>
<td>96</td>
<td>10.28</td>
<td>2.73 (1.94)</td>
</tr>
<tr>
<td>SYS3 live</td>
<td>80</td>
<td>9.56</td>
<td>2.22 (1.14)</td>
</tr>
<tr>
<td>SYS4 control</td>
<td>154</td>
<td>14.70</td>
<td>2.25 (1.78)</td>
</tr>
<tr>
<td>SYS4 live</td>
<td>126</td>
<td>11.00</td>
<td>1.63 (0.77)</td>
</tr>
</tbody>
</table>
# Completion Results

- Live rates lower than Control rates
- Different system orders between Control and Live

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>64.9% (5.0%)</td>
<td>89.4% (3.6%)</td>
<td>74.6% (4.8%)</td>
</tr>
<tr>
<td>Live</td>
<td>60.3% (1.9%)</td>
<td>64.6% (2.3%)</td>
<td>71.9% (1.7%)</td>
</tr>
</tbody>
</table>
### Completion vs WER

#### Completion

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>64.9% (5.0%)</td>
<td>89.4% (3.6%)</td>
<td>74.6% (4.8%)</td>
</tr>
<tr>
<td>Live</td>
<td>60.3% (1.9%)</td>
<td>64.6% (2.3%)</td>
<td>71.9% (1.7%)</td>
</tr>
</tbody>
</table>

#### Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>SYS1</th>
<th>SYS3</th>
<th>SYS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>38.4</td>
<td>27.9</td>
<td>27.5</td>
</tr>
<tr>
<td>Live</td>
<td>43.8</td>
<td>42.5</td>
<td>35.7</td>
</tr>
</tbody>
</table>
Observations

- **Control vs Live is different**
  - Addressing real users is different
  - Control callers speak longer, don’t give up

- **Word error rate matters**
  - But for real users not so much for control

- **Confirmation vs Call Length**
  - Most correct calls vs most correct calls per hour

- **WER Important**
  - But SYS3/SYS4 equal in Control different in Live
  - SYS4 had more controlled dialogs (longer)
  - SYS3 had much shorter dialogs
Further Analysis

- **Automatic Measures**
  - Can we get automatic measures
- **Too much to hand label in real systems**
  - Cannot label every day
  - Need daily automatic measure
  - WER, completion rate etc not accurate enough
  - Can we find something better
- **Full database release**
  - [http://dialrc.org/data/sdc2010](http://dialrc.org/data/sdc2010)
  - Control and Live Test Logs
  - Transcription of all turns
  - Hand labeled success rate
System Deployment
From the lab to real world

- **Developing a real usable system**
  - Strategy to deployment
  - Engineering issues
  - Stability issues

- **Call analysis**
  - Finding issues in calls
From the lab to the real world

- **Build system that works for you**
- **Have locals (developers) use it**
  - Analyze their usage and build a new system
- **Have subset of target users use it**
  - Analyze their usage and build a new system
- **Have target users use it**
  - Analyze their usage and build a new system
- **Over time after successful deployment**
  - Analyze usage and build a new system
Improving components

- Acoustic models (channel-appropriate data)
- Language models (what people actually say)
- Grammar models (how to interpret it)
- Dialog strategies (how people get what they want)
- New functionality (asking for things you could give)
Improving Acoustic models

◆ Collect date of target audience using system
  ● It’s from the right channel
  ● It’s the right distribution of vocabulary
  ● It’s the right distribution of dialects
  ● It’s the right distribution of style

◆ Collect over time
  ● You may have different users and different times of the week.
Collecting real data:

- What words and phrases are used
  - (Are people polite, rude, include greetings …)
- How do people actually ask for things you offer

Grammar additions

- Modify your grammar to deal with what real users actually say
Dialog Strategies and New Functions

🔹 What are people doing
  - Analyze dialog states and look for novel events
    - (e.g. adding “next” and “previous” bus)

🔹 What are they saying when you tell them you can’t help.
  - What functionality are they asking for
How well is my system working

- Number of callers
- Measuring success
  - (Real measuring of success)
- Number of Turns
- Number of User Complaints
- User satisfaction
- Proportion of first time callers
- Call through to operators
User satisfaction

- Caring beyond “success”
  - Good success, painful success

- How to detect it
  - Word level information
    - Politeness (or lack off)
  - Prosodic information
    - Detect frustration, anger etc
  - Human labeling
    - Subsample the data and estimate it
Call analysis

- What is happening in your 100K calls?
- Estimate success
- Count number of turns
- Count number of known errors
- Trace different dialog state sequences
  - Are some more likely to fail.
- What should you do next to improve task success.
Call Analysis

- Cannot listen to them all
  - Need optimal way to sub-sample
- Find the “interesting” calls
- Who are your users
  - First time callers, repeat callers
  - What classes succeed
  - What classes fail
Dialog System Customer

- **USAir, Amtrak, AT&T ...**

- They wish to minimize operator calls
  - % of calls dealt with automatically
  - *(successfully or not ?)*
Dialog Users

- They want the service to work
- They want it to work in an obvious way
- It must be better than waiting for a human
- It must be able to deal with their task
- It must get better for them
System Designer

- **Must be easy to maintain**
  - Not require re-design every day

- **Must be fast**
  - The more calls you serve the money you make
  - The faster calls are the less equipment you need

- **Must adapt to the needs**
  - Customer and users
Must be reliable hardware/software
Can it deal with (near) simultaneous calls
Can it deal with very long calls
POTS is not a very stable system
  • Hard to detect hang up
It has to run 24/7
You must detect hangs automatically
  • Not easy on some operating systems
Cost per call
- What is the average length of a call
- How many simultaneous calls per machine

Can it scale to 10, 100, 1,000 …

Can you deal with call volume
- What are the peaks
- What are the down times

Can your (Amazon) Cloud deal with that?
Spoken Dialog Systems

- **Types of systems**
  - Task oriented, question/answering systems
  - Mixed initiative systems
  - HMIHY: classification tasks

- **Dialog Components**
  - ASR, Parsing
  - Dialog Manager
  - Generation, TTS
Development Systems

- VoiceXML
- Olympus

Deploying systems

- Iterative development
- Call analysis
- Adapting to improve usage