USER, COME BACK!!

Maxine Eskenazi
Language Technologies Institute, CMU
June 16, 1999
What is the goal?

Assessment differs when the application becomes “real”

*Then you can based it on the volume of calls*

As long as we do not make “real” reservations, is user satisfaction valid?

Users can point to places ("hotspots") where they had a hard time

*What do they remember that would make them call or not call again?*

We would want the system with the least “hotspots”
System “hotspots”

What do you remember about the last system you called?

“It kept repeating the information about one hotel.”

Why was it irritating?

“I had to repeat myself”
“The system asked the same thing over and over”
“I told it Pittsburgh and then it asked where I wanted to go”
“It wasn’t talking about the same thing I was”
Measures

label turns by subgoal (departure time, hotel location)

sr = no. syst repeats; ur = no. user repeats;
T = no. turns; us = user then syst said same thing;
sn = syst subgoal /= user subgoal

repeats (R) = sr + ur / 2T - 2 (per goal)

“I had to repeat myself”, “The system asked the same thing over and over”:

“shadows” (S) = us / (T - 1) (per goal)

“I told it Pittsburgh and then it asked where I wanted to go”

system/=user (N) = sn / T

“It wasn’t talking about the same thing I was”
System Assessment

Compare semi-automatic measures to what users say

Panel of 9 callers

9 different scenarios, 3x3 different difficulty levels

easy = 1 leg; med. = 2 legs+; hard = 3 legs+
each caller did: 1 easy, 1 medium, 1 hard
Repetitions - questions

Tell me where in the dialogue:

Something didn’t go as you thought it should

You had to change what you wanted

You wanted to give up

You had no idea what to do next
Comparative results

For 20 dialogues, user/label agreement:

highest number of turns: 65%

highest percent of repeats (R): 85%

highest percent of system “shadows (S): 60%

highest percent of system/=user (N): 55%

(S) seems to refine the information that is in (R)
divides problem repeats from simple navigation
## Comparative Results - Task Difficulty

<table>
<thead>
<tr>
<th></th>
<th>Turns</th>
<th>Repetitions</th>
<th>Shadows</th>
<th>S/=U</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>65%</td>
<td>85%</td>
<td>60%</td>
<td>55%</td>
</tr>
<tr>
<td>“Easy”</td>
<td>50%</td>
<td>100%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>“Medium”</td>
<td>62.5%</td>
<td>75%</td>
<td>62.5%</td>
<td>75%</td>
</tr>
<tr>
<td>“Hard”</td>
<td>80%</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
</tr>
</tbody>
</table>
Assessment - Sanders questionnaire

- Task success
- Usefulness
- Efficiency
- ASR understanding
- Comprehensible TTS
- Response time
- User initiative
- Know what to say
- S: understood all info
- Task success
- Efficiency
- Usefulness
- ASR understanding
- Comprehensible TTS
- Response time
- User initiative
- Know what to say
- S: understood all info
Sanders questionnaire - cont.

- Easy to correct
- Can be successfully corrected
- Realise there is a pb
- Got info efficiently
- How to get info
- U: got all info

0 1 2 3 4 5
Length of dialogues

- Longest
- Shortest
- Hard
- Medium
- Easy
- MEAN

0 200 400 600 800
## Acquired Data

<table>
<thead>
<tr>
<th>Type</th>
<th>No. Dialogs</th>
<th>No. Utts.</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Human</td>
<td>58</td>
<td>1800</td>
<td>~1 hr</td>
</tr>
<tr>
<td>WOZ1</td>
<td>107</td>
<td>1992</td>
<td>~1.3 hrs</td>
</tr>
<tr>
<td>WOZ2</td>
<td>16</td>
<td>487</td>
<td>~20 min</td>
</tr>
<tr>
<td>Movieline</td>
<td>112</td>
<td>7527</td>
<td>~3 hrs</td>
</tr>
<tr>
<td>System</td>
<td>2861</td>
<td>44783</td>
<td>~11.4 hrs</td>
</tr>
</tbody>
</table>