Lecture 5

Dialogue System Evaluation
Why Evaluate?

- Is system good (enough)?
- Is (system/module/strategy) A better than B?
- What are the problems with the system?
- How do we make it better?
Types of Evaluation

- Glass Box vs Black Box
- System-wide vs component level metrics
- Subjective vs objective metrics
- Task-performance vs satisfaction
- Satisfy who?
  - User
  - Owner
  - teacher
Offline vs Online Evaluation

- Online: evaluated as to actual dialogue run
- Offline: use pre-collected dialogue corpora as test set
- Online: Who are the subjects?
  - Agents/simulations?
  - Humans
    - Novices?
    - Experts?
    - Real target population?
Task performance

• Performance quality
  – Task completed?
  – Parts of task completed?
  – Quality of solution?

• Performance efficiency
  – Time metrics
    • Elapsed time
    • Number of turns
    • Number of words
  – Other resource metrics
Subjective measures

• User satisfaction
• User perceived completion/correctness
• Hand-coded features
  – Transcription
  – Concept ID/correct understanding
  – Speech acts
  – Correct responses
  – initiative
• How reliable is the coding?
  – Kappa
Component-level analysis

• ASR: WER
• NLU: “concept accuracy”
• Dialogue: ??
• Generation: concept accuracy, fluency
• Synthesis: understandability
TRAINS-95 Evaluation

- Trains-95 system
  - Simpler, robust version of trains
- Main evaluation: task performance
  - Quality of solution
  - Time to completion
- Studying:
  - Is system usable?
  - Is speech feasible (compared to text input)?
  - How does a speech post-processor correcting off-the-shelf recognizer effect dialogue quality?
TRAINS-95 procedure

- 16 subjects, 2x2 grid
- Tutorial video & practice session for training
- 5 tasks (last one choice of mode)
TRAINS-95 Results

- Speech just as good and faster than text (but occasionally fail)!
- Subjects preferred to use speech (but perhaps from novelty rather than efficiency)
- Limited correlation between WER (actually WRA) and dialogue time, perhaps because:
  - Robust parsing
  - Nonunderstanding vs misunderstanding
  - Differences in system strategy
Paradise

- Paradigm for Dialogue System Evaluation
- User satisfaction is primary
- What accounts for User Satisfaction?
- Method:
  - Collect sample dialogues
  - User satisfaction by compound interview
  - Collect system parameters
  - Find best correlation between system parameters and user satisfaction (what features ‘explain’ differences in satisfaction)
    - Linear regression
Paradise Models

General Models of Usability

MAXIMIZE USER SATISFACTION

MAXIMIZE TASK SUCCESS
MINIMIZE COSTS

EFFICIENCY MEASURES
QUALITATIVE MEASURES
Walker, Kamm, & Litman

- Comparison of three systems (Elvis, Annie, Toot)
- Two different domains
- How do paradise models generalize across data?
Communicator Evaluation Metrics

- **Dialogue Efficiency**: Task Duration, System turns, User turns, Total Turns

- **Dialogue Quality**: Word Accuracy, Response latency, Response latency variance

- **Task Success**: Exact Scenario Completion

- **User Satisfaction**: Sum of TTS performance, Task ease, User expertise, Expected behavior, Future use.
Communicator Evaluation

• Many systems (9), different styles, architectures
• Same tasks
• How to evaluate across systems?
  – Standard log files
  – Users use multiple systems
  – Paradise style evaluation
• Examining communicator dialogues
• Using dialogue acts as part of “paradise” formula
• Dialogue Act Tagging for Evaluation
• 3 dimensions of acts
  – Speech act
  – Task-subtask
    • “effort” on subtask - sum of lengths of utterances in subtask
  – Conversational domain
    • About task
    • About communication (managing channel, grounding)
    • Situation frame (how to talk to system)
• Tagging only system utterances
<table>
<thead>
<tr>
<th>Speech-Act</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>REQUEST-INFO</td>
<td>And, what city are you flying to?</td>
</tr>
<tr>
<td>PRESENT-INFO</td>
<td>The airfare for this trip is 390 dollars.</td>
</tr>
<tr>
<td>OFFER</td>
<td>Would you like me to hold this option?</td>
</tr>
<tr>
<td>ACKNOWLEDGE</td>
<td>I will book this leg.</td>
</tr>
<tr>
<td>STATUS-REPORT</td>
<td>Accessing the database; this might take a few seconds.</td>
</tr>
<tr>
<td>EXPLICIT-CONFIRM</td>
<td>You will depart on September 1st. Is that correct?</td>
</tr>
<tr>
<td>IMPLICIT-CONFIRM</td>
<td>Leaving from Dallas.</td>
</tr>
<tr>
<td>INSTRUCTION</td>
<td>Try saying a short sentence.</td>
</tr>
<tr>
<td>APOLOGY</td>
<td>Sorry, I didn’t understand that.</td>
</tr>
<tr>
<td>OPENING/CLOSING</td>
<td>Hello. Welcome to the C M U Communicator.</td>
</tr>
</tbody>
</table>
## Task-subtask

<table>
<thead>
<tr>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP-LEVEL-TRIP</td>
<td>What are your travel plans?</td>
</tr>
<tr>
<td>ORIGIN</td>
<td>And, what city are you leaving from?</td>
</tr>
<tr>
<td>DESTINATION</td>
<td>And, where are you flying to?</td>
</tr>
<tr>
<td>DATE</td>
<td>What day would you like to leave?</td>
</tr>
<tr>
<td>TIME</td>
<td>Departing at what time?.</td>
</tr>
<tr>
<td>AIRLINE</td>
<td>Did you have an airline preference?</td>
</tr>
<tr>
<td>TRIP-TYPE</td>
<td>Will you return to Boston from San Jose?</td>
</tr>
<tr>
<td>RETRIEVAL</td>
<td>Accessing the database; this might take a few seconds.</td>
</tr>
<tr>
<td>ITINERARY</td>
<td>I found 3 flights from Miami to Minneapolis.</td>
</tr>
<tr>
<td>PRICE</td>
<td>The airfare for this trip is 390 dollars.</td>
</tr>
<tr>
<td>GROUND</td>
<td>Did you need to make any ground arrangements?</td>
</tr>
<tr>
<td>HOTEL</td>
<td>Would you like a hotel near downtown or near the airport?</td>
</tr>
<tr>
<td>CAR</td>
<td>Do you need a car in San Jose?</td>
</tr>
</tbody>
</table>
WPB: DATE usage

• Automatic tagging of system utterances
  – Easy because of template generation
Eckert et al: Automatic Evaluation

- Goal: be able to compare systems
- Method: automated users, generate “random” dialogues according to a user model
- Assign a quality metric for a dialogue as sum of weighted cost functions
- Evaluation of dialogue system on user model as sum over all possible dialogues of quality of dialogue times probability of dialogue
Eckert et al: Evaluation Environment
Eckert et al

• Advantages:
  – More testing than available data
  – Cheaper (not human-intensive)
  – “reliable” - same model for all systems/variations

• Disadvantages
  – How can you tell when you have a good sample?
  – Building a user model can be as complex or more than building a good system/system model