Dialogue System Evaluation

Evaluation

- things we can measure about how well a dialogue went
  - user satisfaction
  - learning
  - task completion
  - how long they stayed with it
- outcomes
  - tell us how well a dialogue went
  - can be represented numerically in some way and then predicted based on what happened within the dialogues themselves
  - you need to keep records of what happened in the dialogues themselves

PARADISE Framework

- used to evaluate dialogue systems
- *performance* of a dialogue system is affected by both:
  - *what* gets accomplished by the user and the dialogue agent and
  - *how* it gets accomplished
- maximize user satisfaction
  - maximize task success
  - minimize costs
    - efficiency measures
    - qualitative measures
- regress against user satisfaction
  - questionnaire to assign each dialogue a user satisfaction rating - *dependent* measure
  - cost and success factors - *independent* measures
  - use regression to train weights for each factor

Experimental Procedures

- subjects given specific tasks
- spoken dialogues recorded
- cost factors, states, dialogue acts automatically logged
- ASR accuracy, barge-in hand-labeled
- users specify task solution via web page
- users complete user satisfaction survey of some kind
- use **multiple linear regression** to model user satisfaction as a function of task success and costs

    - test for significant predictive factors

Success Metric

- could we use the success metric to drive automatic learning?
- methods for automatically evaluating system performance
- way of obtaining training data for further system development
- can we find intrinsic evaluation metrics that correlate with extrinsic results?