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)

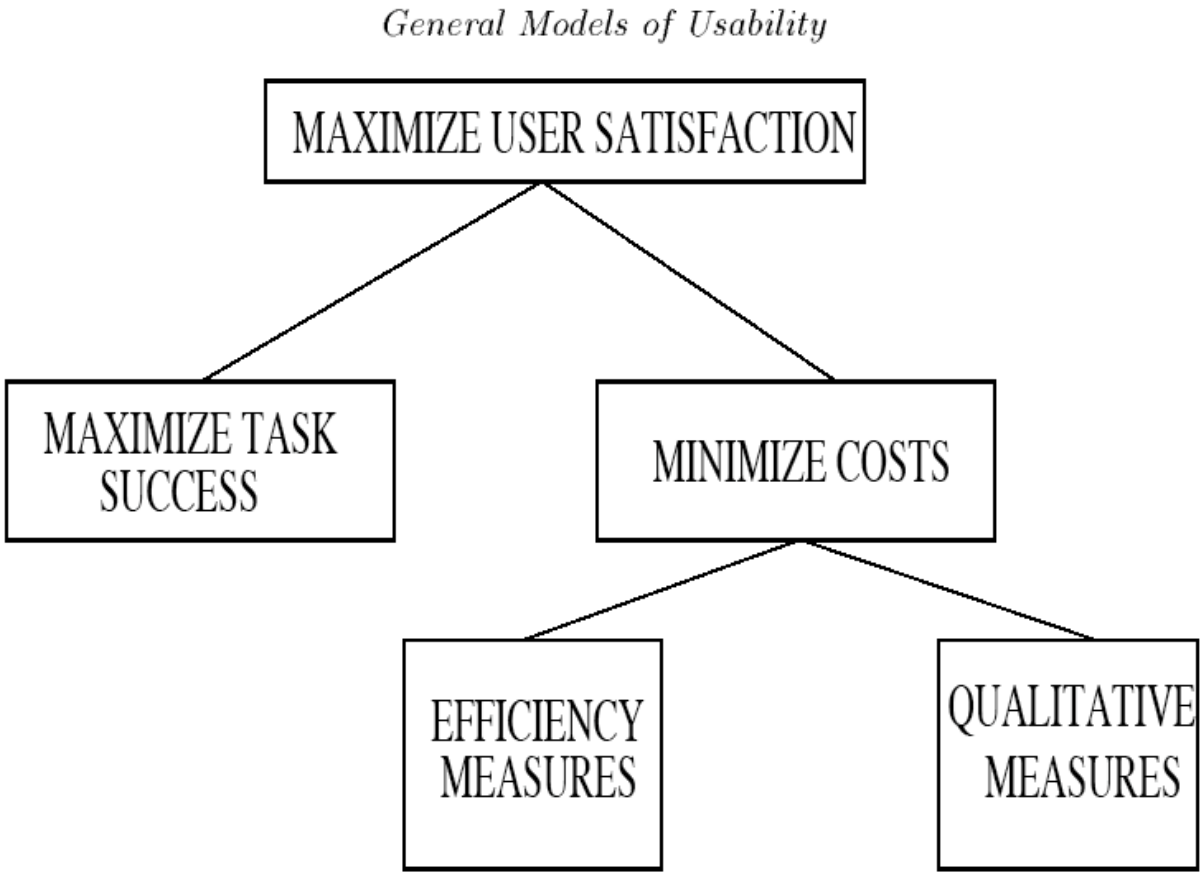
TRAINS-95 Results

- 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



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

Walker, Passonneau & Boland

- Examining communicator dialogues
- Using dialogue acts as part of “paradise” formula

DATE

- 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

DATE Dialogue Acts

Speech-Act	Example
REQUEST-INFO	*And, what city are you flying to?*
PRESENT-INFO	*The airfare for this trip is 390 dollars.*
OFFER	*Would you like me to hold this option?*
ACKNOWLEDGE	*I will book this leg.*
STATUS-REPORT	*Accessing the database; this might take a few seconds.*
EXPLICIT-CONFIRM	*You will depart on September 1st. Is that correct?*
IMPLICIT-CONFIRM	*Leaving from Dallas.*
INSTRUCTION	*Try saying a short sentence.*
APOLOGY	*Sorry, I didn't understand that.*
OPENING/CLOSING	*Hello. Welcome to the C M U Communicator.*

Task-subtask

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

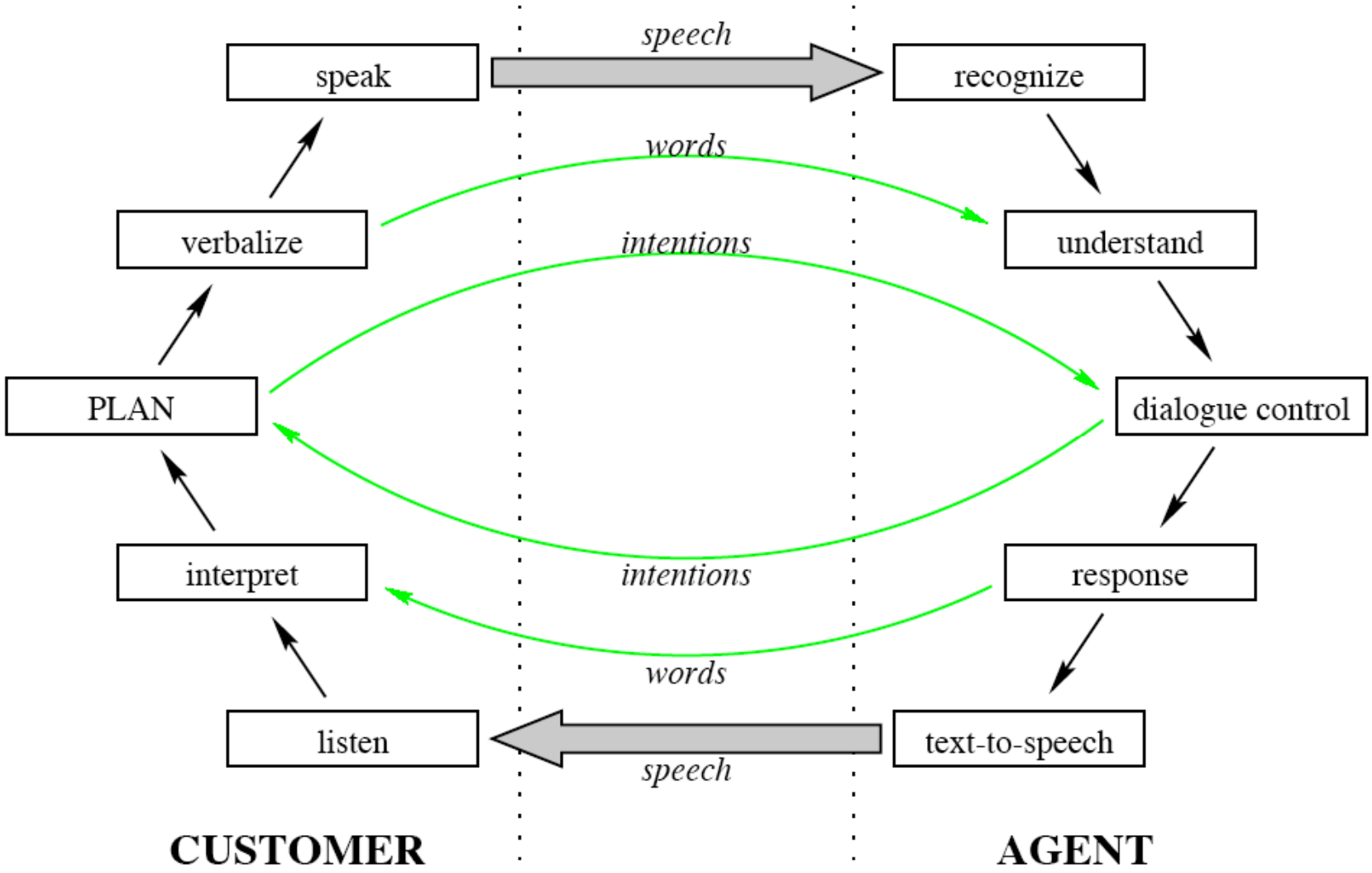
WPB: DATE usage

- Automatic tagging of system utterances
  - Easy because of template generation

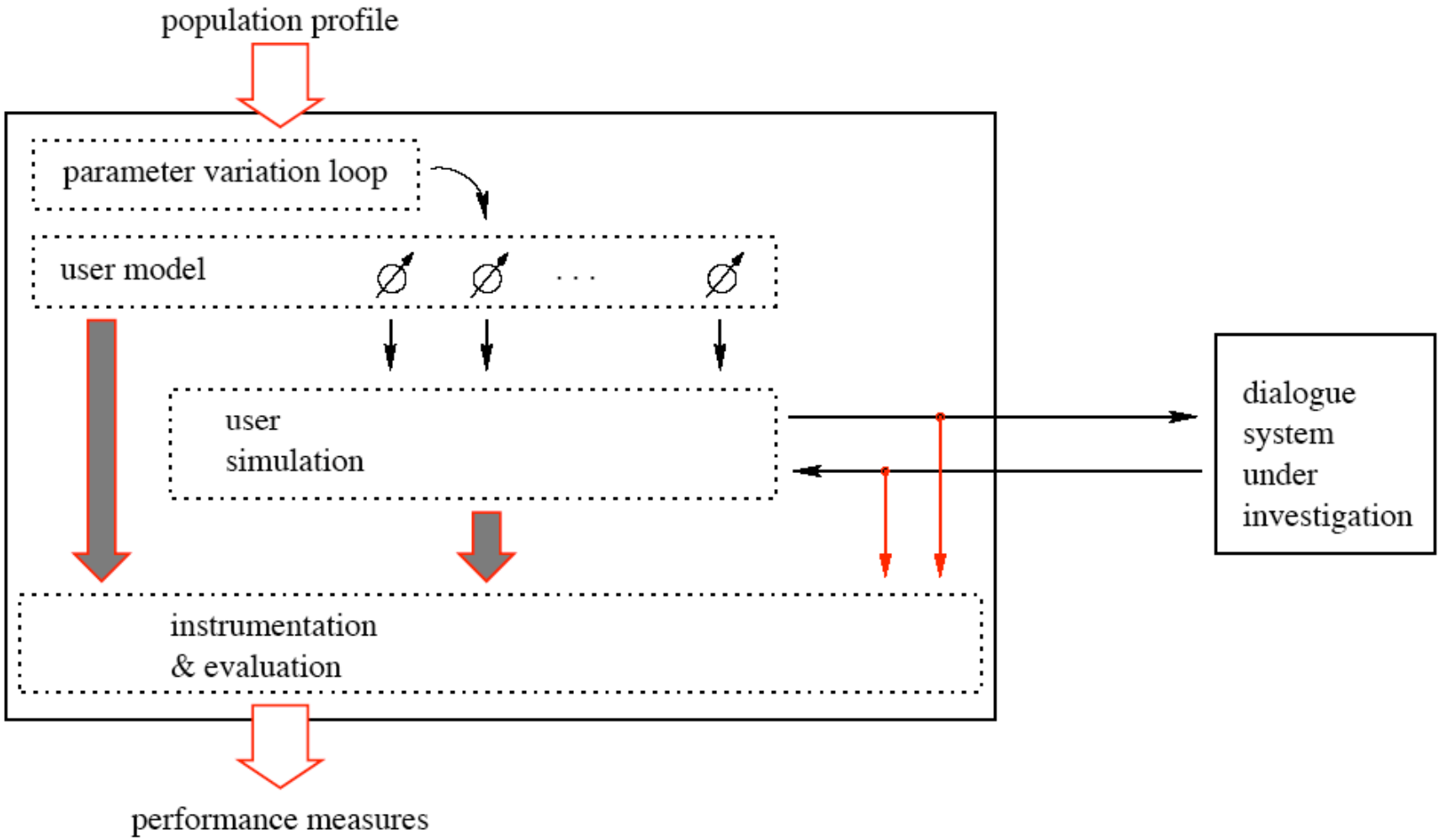
Eckert et al: Automatic Evaluation

- 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 Feedback model



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