Evaluating Spoken Dialogue Systems

Julia Hirschberg

CS 4706

Dialogue System Evaluation

- Key point about SLP.
- Whenever we design a new algorithm or build a new application, need to evaluate it
- Two kinds of evaluation
  - **Extrinsic:** embedded in some external task
  - **Intrinsic:** some sort of more local evaluation.

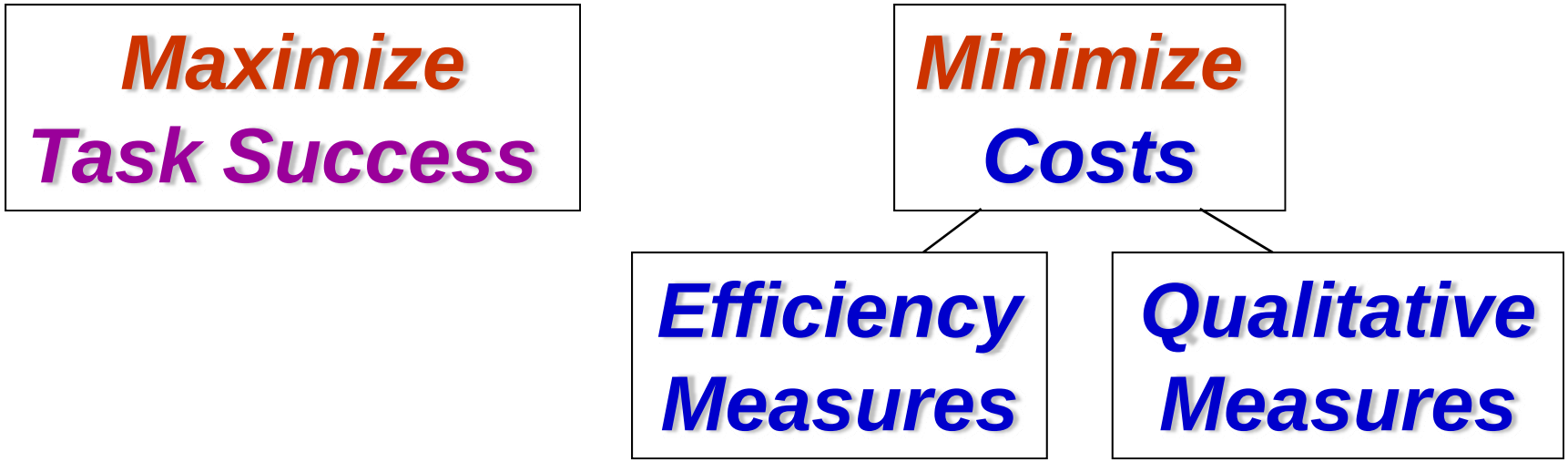
- How to evaluate a dialogue system?
- What constitutes success or failure for a dialogue system?

Dialogue System Evaluation

- Need evaluation metric because
  - 1) Need metric to help compare different implementations
    - Can't improve it if we don't know where it fails
    - Can't decide between two algorithms without a goodness metric
  - 2) Need metric for "how well a dialogue went" as an input to reinforcement learning:
    - Automatically improve our conversational agent performance via learning

Evaluating Dialogue Systems

- PARADISE framework (Walker et al ’00)
- “Performance” of a dialogue system is affected both by *what* gets accomplished by the user and the dialogue agent and *how* it gets accomplished



Task Success

- % of subtasks completed
- Correctness of each questions/answer/error msg
- Correctness of total solution
  - Attribute-Value matrix (AVM)
  - Kappa coefficient
- Users' perception of whether task was completed

Task Success

- **Task goals seen as Attribute-Value Matrix**

  ELVIS e-mail retrieval task ***(Walker et al '97)***

  "Find the time and place of your meeting with Kim."

Attribute	***Value***
Selection Criterion	***Kim or Meeting***
Time	***10:30 a.m.***
Place	***2D516***

- **Task success can be defined by match between AVM values at end of task with "true" values for AVM**

Slide from Julia Hirschberg

Efficiency Cost

- Polifroni et al. (1992), Danieli and Gerbino (1995) Hirschman and Pao (1993)
- Total elapsed time in seconds or turns
- Number of queries
- Turn correction ratio:
  - Number of system or user turns used solely to correct errors, divided by total number of turns

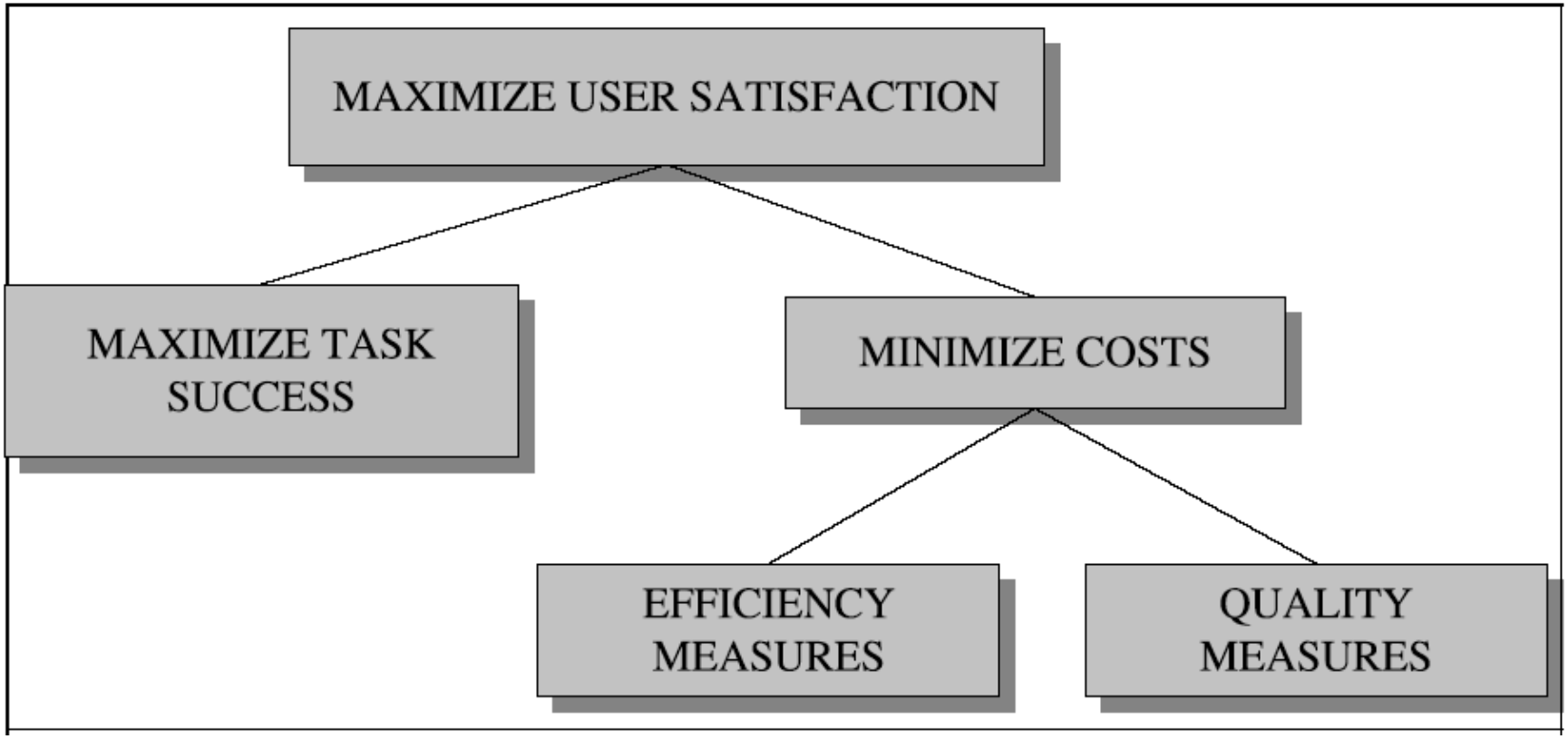
Quality Cost

- # of times ASR system failed to return any sentence
- # of ASR rejection prompts
- # of times user had to barge-in
- # of time-out prompts
- Inappropriateness (verbose, ambiguous) of system’s questions, answers, error messages

Another Key Quality Cost

- “Concept accuracy” or “Concept error rate”
- % of semantic concepts that the NLU component returns correctly
- I want to arrive in Austin at 5:00
  - DESTCITY: Boston
  - Time: 5:00
- Concept accuracy = 50%
- Average this across entire dialogue
- “How many of the sentences did the system understand correctly”

PARADISE: Regress against user satisfaction



Regressing 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 specified tasks
- Spoken dialogues recorded
- Cost factors, states, dialog acts automatically logged; ASR accuracy,barge-in hand-labeled
- Users specify task solution via web page
- Users complete User Satisfaction surveys
- Use multiple linear regression to model User Satisfaction as a function of Task Success and Costs; test for significant predictive factors

User Satisfaction: Sum of Many Measures

Was the system easy to understand? (TTS Performance)

Did the system understand what you said? (ASR Performance)

Was it easy to find the message/plane/train you wanted? (Task Ease)

Was the pace of interaction with the system appropriate? (Interaction Pace)

Did you know what you could say at each point of the dialog? (**User Expertise**)

How often was the system sluggish and slow to reply to you? (System Response)

Did the system work the way you expected it to in this conversation? (Expected Behavior)

Do you think you'd use the system regularly in the future? (Future Use)

Performance Functions from Three Systems

- ELVIS User Sat.= .21* COMP + .47 * MRS - .15 * ET
- TOOT User Sat.= .35* COMP + .45* MRS - .14*ET
- ANNIE User Sat.= .33*COMP + .25* MRS +.33* Help
  - COMP: User perception of task completion (task success)
  - MRS: Mean (concept) recognition accuracy (cost)
  - ET: Elapsed time (cost)
  - Help: Help requests (cost)

Performance Model

- Perceived task completion and mean recognition score (concept accuracy) are consistently significant predictors of User Satisfaction
- Performance model useful for system development
  - Making predictions about system modifications
  - Distinguishing ‘good’ dialogues from ‘bad’ dialogues
  - Part of a learning model

Now that we have a Success Metric

- Could we use it to help drive automatic learning?
  - Methods for automatically evaluating system performance
  - Way of obtaining training data for further system development

Recognizing `Problematic' Dialogues

- Hastie et al, "What's the Trouble?" ACL 2002
- Motivation: Identify a Problematic Dialogue Identifier (PDI) to classify dialogues
- What is a Problematic Dialogue
  - Task is not completed
  - User satisfaction is low
- Results:
  - Identify dialogues in which task not completed with 85% accuracy
  - Identify dialogues with low user satisfaction with 89% accuracy

Corpus

- 1242 recorded dialogues from DARPA Communicator Corpus
  - Logfiles with events for each user turn
  - ASR and hand transcriptions
  - User information: dialect
  - User Satisfaction survey
  - Task Completion labels
- Goal is to predict
    - User Satisfaction (5-25 pts)
    - Task Completion (0,1,2): none, airline task, airline+ground task

DATE Dialogue Act Extraction

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?*
ACKNOWLEDGMENT	*I will book this leg.*
BACKCHANNEL	*Okay.*
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.*

Features Used in Prediction

- **Efficiency Measures**
  - *Hand-labelled*: WERR, SERR
  - *Automatic*: TimeOnTask, TurnsOnTask, NumOverlaps, MeanUsrTurnDur, MeanWrdsPerUsrTurn, MeanSysTurnDur, MeanWrdsPerSysTurn, DeadAlive, Phone-type, SessionNumber
- **Qualitative Measures**
  - *Automatic*: DATE Unigrams, e.g. present-info:flight, acknowledgement:flight_booking etc.
  - *Automatic*: DATE Bigrams, e.g. present-info:flight+acknowledgement:flight_booking etc.
- **Task Success Features**
  - *Hand-labelled*: HL Task Completion
  - *Automatic*: Auto Task Completion

Results

	Baseline	Auto Logfile	ALF + GC	ALF + GC+ DATE
TC	59%	59%	79%	85%
BTC	86%	86%	86%	92%

Table 1: Task Completion (TC) and Binary Task Completion (BTC) prediction results, using automatic logfile features (ALF), GroundCheck (GC) and DATE unigram frequencies

Feature used	Log features	LF + unigram	LF + bigram
HL TC	0.587	0.584	0.592
Auto TC	0.438	0.434	0.472
HL BTC	0.608	0.607	0.614
Auto BTC	0.477	0.47	0.484

Table 2: Correlation results using logfile features (LF), adding unigram proportions and bigram counts, for trees tested on either hand-labelled (HL) or automatically derived Task Completion (TC) and Binary Task Completion (BTC)

Task Completion	Dialogue	Recall	Prec.
Hand-labelled	Good	90%	84.5%
Hand-labelled	Problematic	54.5%	66.7%
Automatic	Good	88.5%	81.3%
Automatic	Problematic	66.7%	58.0%

Table 3: Precision and Recall for good and problematic dialogues (where a good dialogue has User Satisfaction>12) for the PDI using hand-labelled Task Completion and Auto Task Completion

Summary

- Intrinsic vs. extrinsic evaluation methods
  - Key: can we find intrinsic evaluation metrics that correlate with extrinsic results?
- What other sorts of measures can you think of?