

Machine Learning and Dialogue Evaluation Techniques

James Henderson

School of Informatics
University of Edinburgh

ESLLI 2006

<http://homepages.inf.ed.ac.uk/jhender6/esslli2006/>

Outline

- 1 Outline of Remaining Material
- 2 Evaluation Techniques: Real Users, Simulated Users
 - Real versus Simulated Users
 - Real Users
 - Simulated Users

Outline

- 1 Outline of Remaining Material
- 2 Evaluation Techniques: Real Users, Simulated Users
 - Real versus Simulated Users
 - Real Users
 - Simulated Users

Outline of Remaining Material

- Evaluation Techniques
 - Real Users
 - Simulated Users
- Reinforcement Learning with Complex States
 - State Features
 - Reinforcement Learning with Fixed Datasets
 - Exploratory Reinforcement Learning
- Partially Observable Markov Decision Processes
 - POMDPs
 - Approximating POMDPs
- Future Research Topics

Outline

- 1 Outline of Remaining Material
- 2 Evaluation Techniques: Real Users, Simulated Users
 - Real versus Simulated Users
 - Real Users
 - Simulated Users

Evaluation Methods

- Evaluation is a major bottleneck in work on dialogue
- Experiments with real users take hundreds of person-hours to run
- Simulated users help address this bottleneck

Real Users versus Simulated Users

- Performance with real users is the real objective. Simulated users are only useful to the extent that they correlate with real users
- Experiments with real users are very expensive to run. Simulated users can generate much more data at much less cost
- Real users can fill out usability questionnaires. Simulated users require objective measures of dialogue quality

Evaluation with Real Users

- Dialogues are generated by first setting users a task

You are on a business trip on your own. You need to find a hotel room in the middle of town. Price is no problem.

- During the session, everything that happens is recorded, and various objective evaluation measure are calculated
- After the end of the session, the user fills out a usability questionnaire

Objective Quality Measures

Some quality measures can be collected just from looking at the dialogues:

- actual task completion
- dialogue length
- etc.

Usability Questionnaire

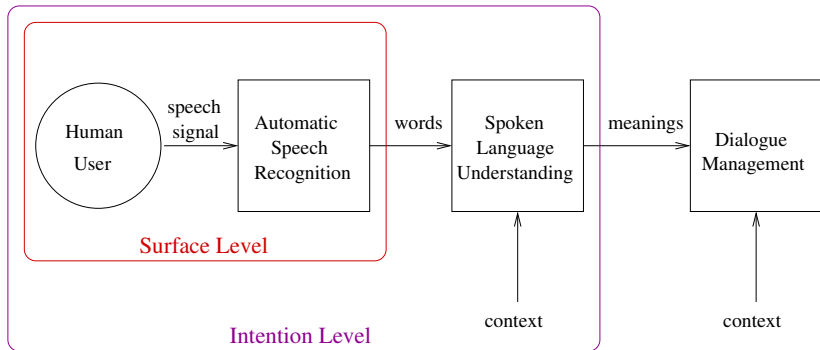
Some quality measures can only be determined by asking the user's opinion:

- perceived task completion
- ease of use
- expected behavior
- like to use again

Experiments with Simulated Users

- Dialogues can be generated by substituting a real user with a simulated user in the processing loop
- Large numbers of dialogues can be generated in this way in a relatively short time
- The dialogues can be evaluated using objectively measurable criteria which approximate the measures collected from real users

Levels of User Simulation



- Surface level: produces sequences of user utterances.
- Intention level: produces sequences of user dialogue acts, tasks, etc.

User Simulations

- User simulation is a **supervised learning** problem
- Given a representation of the dialogue state and previous system action, the user model needs to output a probability distribution over user actions
- The user simulation choose an action **stochastically** from the distribution output by the user model

State Features in User Simulation

- The main question in user simulations is what to condition on:
 - only the previous system action [Schatzmann et al., *SIGDial 2005*]
 - n-grams of previous system and user actions [Georgila et al., *Eurospeech 2005*]
 - n-grams of previous system and user actions plus filled/confirmed features for associated slots [Georgila et al., *Interspeech 2005*]
 - all available features of the dialogue state [Georgila et al., *Eurospeech 2005*]
- Even very simple user simulations can result in realistic dialogues

Supervised Learning Methods

- If there are few enough states that you have enough data for each state, then it is enough to calculate a table of relative frequencies (e.g. only using the previous system action)

$$P(a|s) = \frac{\# \langle s, a \rangle}{\# \langle s \rangle}$$

- With larger numbers of states, it is good to use a smoothing method (such as backoff or linear interpolation), which allows similar states to share counts (e.g. with 5-grams of actions)
- With very large, or infinite, state spaces, log-linear feature combination (a.k.a. “maximum entropy” models) is a good approach. This can be used in combination with kernel methods

Summary

- Experiments with real users are the real objective, but are expensive to run
- Simulated users can be used do intermediate evaluations cheaply