A User Model for Dialog System Evaluation Based on Activation of Subgoals

Klaus-Peter Engelbrecht

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin

Abstract

User models have become increasingly popular to conduct simulation-based testing of spoken dialog systems. These models usually describe users' overt behavior, as opposed to the underlying reasons for the observed actions. While such models are useful to generate test data, a causal model might be more generally applicable to different systems, and in addition allow to derive useful information for data analysis and prediction of user judgments. Thus, a modeling approach trying to explain user behavior is proposed in this paper, which is based on Dörner's PSI theory. The evaluation shows that the utterances generated by this model are similar to those of real users.

Introduction

- User simulation is used for **dialog system testing** or training
- Currently, statistical models of **overt user behavior** are used
  - Finding **interaction problems** works well
  - **Performance** prediction somewhat possible
- **Assumption**: causal user models...
  - are **more generally** applicable to different systems
  - will be useful to automatically **detect interaction problems** in simulation data
  - will allow to derive useful **predictors of user judgments**
- **Modeling approach**
  - **PSI Theory** (Dörner, 2002)
  - Explains human behavior as a consequence of **needs** (e.g. hunger)
  - Actions are taken to fulfill needs
  - **Application to HCI:** the user's subgoals are understood as needs

Task, Belief and Needs

- **Task** = set of constraints the user wants to communicate to the system
- **System state** = understood constraints
- **Belief** = constraints believed by the user to be understood
- **Need strength** = urgency of a need, leading to a motif the user might pursue (edit distance between task and belief)

Task: You want to invite your friend to a typical German dinner on Saturday. Choose a good restaurant in the city centre!

Slot	State	Belief	Task	Need strength
price	„expensive"	„expensive"	„expensive"	0
location	„south"	„south"	„centre"	2
date			„saturday"	1
time	„evening"		„evening"	1
foodtype	„german"	„italian"	„german"	2

Start with all slots empty.

In case of a no-match prompt, no change is required.

If the system provides no feedback on values specified by the user, add these values, but only if the system continues consistently (e.g. not asking for one of the provided slots)

System: Welcome to the Bochum restaurant information system BoRIS. I'll help you find a restaurant based on the following criteria: Food type, price range, time, date, and location. How may I help you?

User: I'm looking for a restaurant in the university district serving traditional German food.

System: OK, when do you want to go there?

User: Saturday night.

System: I couldn't understand you. When do you want to go there?

User: Saturday night.

System: Please confirm Saturday night.

Add AVPs explicitly confirmed by the system.

User: Yes

(SU: {logical=„no"})

System: When do you want to go there?

User: Saturday.

(SU: {date=„Thursday" ; location=„centre"})

In case affirmation of the confirmation by the user is required, and the user does not affirm or the system asks for any of the confirmed values in the next exchange, remove all confirmed values.

System: Please repeat the location.

User: City centre.

System: How much do you want to spend?

User: No more than fifteen Euro.

System: How much do you want to spend?

...

If the system asks for repetition of the slot value, fill the slot with the placeholder "XXX" (unknown inserted value)

Probability of Success

If need was the only criterion for action selection, the users would always take all actions leading to a need reduction immediately. However, users seem to evaluate the chances that an utterance will be understood for the given dialog context.

The estimated probability of success depends on...

- **Dialog context** (in particular the system prompt)
- **Previous experience** of the user with the system (i.e. the dialog history), particularly recognition errors
- The **user's general confidence** in interacting with the system. May be formed in interactions with other systems.

Simple model at this stage:

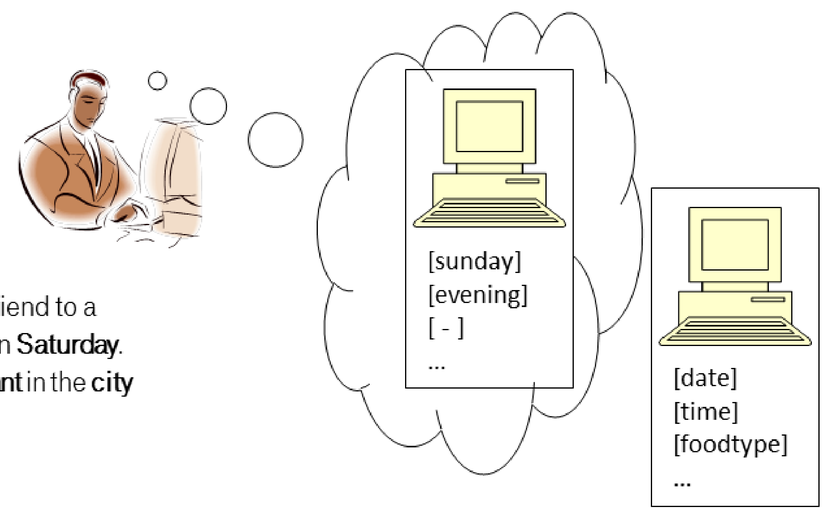
- The **constraint the system asked for** receives a fixed high probability of success
- All other constraints receive a lower probability (value varies between users in order to model different degrees of user initiative).

Planning

Some actions are not immediately related to constraints or needs. Then, actions inherit value from the subgoals they are related to.

- **Explicit confirmation**: value of affirmation and negation equal the collective value of the confirmed constraints.
- **Constraint relaxation**: The constraint is first changed in the task model, leading to a need strength of 2. The accept action allows to reduce this need in the next step, and thus receives a value corresponding to this need's strength.
- **Naming slots to fill next**: In the analyzed system, this action allows to control the dialog flow. It receives a value which is a fraction of the value of the respective constraint. The size of this fraction is a user parameter.

Action Selection



Task:
You want to invite your friend to a typical **German dinner** on **Saturday**. Choose a **good restaurant** in the **city centre**!

Actions:	[date]	[time]	[foodtype]	...[slot=date]	[slot=time]	...[yes]	[no]
Value:	[2]	[0]	[1]	...[1.6]	[0]	...[0]	[0]
pSucc:	[0.8]	[0.8]	[1]	...[0.8]	[0.8]	...[0.8]	[0.8]
Value*pSucc:	[1.6]	[0]	[1]	...[1.3]	[0]	...[0]	[0]

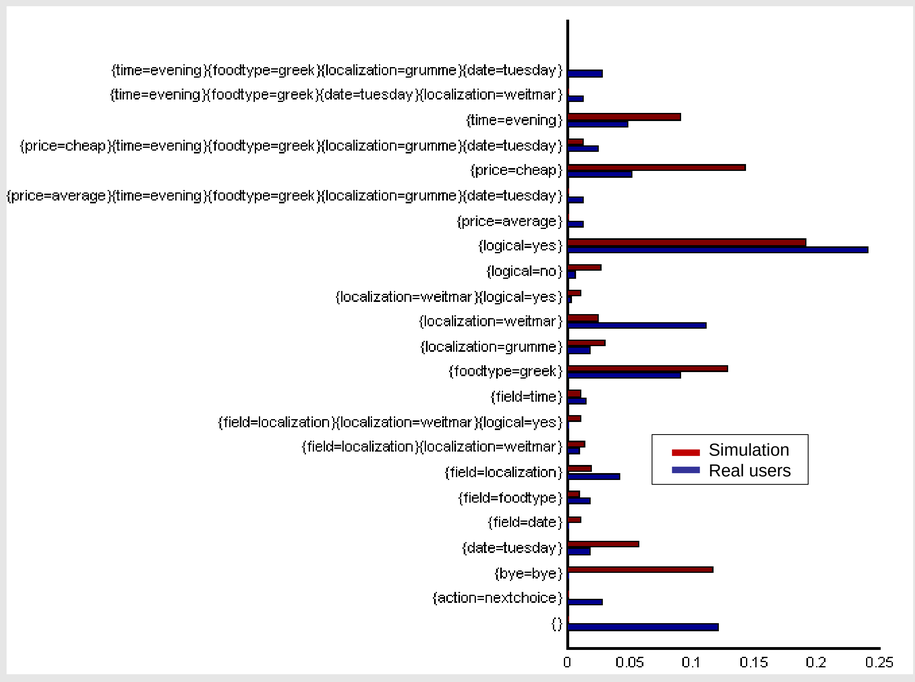
Prompt: What type of food are you looking for?

Further parameters needed to select actions:

- *Threshold* above which actions are selected
- *pSuccNotAsked* models user initiative
- *noiseLevel* models difference in importance of each constraint
- *dampEqualAttributes* so the user does not say opposing things
- *needScent* – value of a semantically related concept is

Comparison with Real User Data

- 2000 simulated dialogs
- MAD = mean absolute difference of user turn prior probabilities in each corpus
- Figure: only turns with a prior probability above 0.01 are displayed (table presents results for all



Database

- **BoRIS: Bochum restaurant information system** (Möller 2005)
  - Mixed-initiative, slot filling (date, time, price range, food type, location)
  - Constraint relaxation or refinement, depending on number of solutions
- **Wizard-of-Oz test** with speech recognition simulation
  - Different target **word accuracies** (60,...,100%)
  - Different **confirmation** strategies (explicit/ none)
  - Different **voices** (male/ female/ TTS)
- 40 Users (29m, 11f; *M* = 29.0y, *SD*=9.7)
- 5 tasks per user, only one used here

Agreement between utterances in both corpora	
N(emp)	40
N(sim)	161
N(common)	21
Recall	0.53
Precision	0.13
MAD	0.005

Conclusions & Future Work

- A new, causal model of user behavior was presented
- Model parameters are directly related to characteristics of the users and the task
- Proof-of-concept evaluation successful
- Test generalizability to other systems, compared to statistical models
- Derive useful predictors of user judgments
- Long-term goal: emotional user behavior