Towards Cognitive Dialog Systems

Felix Putze[1]**, Tanja Schultz**[1]

[1]*Universität Karlsruhe (TH), Cognitive Systems Lab*

In this paper, we report on our initial setup and ongoing research on the development of cognitive dialog systems for dynamic environments. We describe the main components that we consider necessary to build dialog systems that estimate the user's mental processes (hence cognitive) and adapt their behavior accordingly: We require a realistic testing and recording environment to produce real-life data, e.g. a realistic driving simulator. We further need to observe the user during these interactions in a multimodal way to estimate the current user state based on this data. This information is integrated with cognitive modeling components that enrich the observational data. We finally need to find dialog strategies that adapt the behavior of the interactive system to optimally suit the current needs of the user. We report our progress in building these components, give an overview over the challenges we identified during this work and the solutions we aim for.

1 Introduction

Spoken dialog systems have matured to a point where they find their way to many real-world applications like tutoring systems (Litman, 2004) or automated call-center agents (Gorin, 1997). However, their application in more dynamic scenarios remains an open and very interesting task. Spoken dialog systems as an interface for in-car services are very desirable and at the same time very challenging: On the one hand, they offer an eyes-free and hands-free control without visual or manual distraction from the primary driving task. On the other hand, this task uses the user's cognitive capacity and we can no longer assume to deal with a fully attentive and perfect interaction partner as we can in more static environments. Another important aspect is the adaptation to individual preferences: As dialog sessions in driving scenarios may last for several hours, we have to take into account both changing user states, i.e. cognitive workload or emotions, as well as lasting user traits, e.g. his gender or personality. Both types of individual differences influence the optimal interaction behavior the system should use for maximizing user satisfaction, as several user studies show (Nass, 2000), (Nass, 2005). We see potential for a large range of adaptation measures: One example is reacting to increased cognitive workload by taking the initiative from the user, delaying non-critical information or reducing its complexity. Another one is adjusting the system to the user's emotional state and personality by selecting appropriate wording, voice and turn-taking behavior, as suggested by (Nass, 2005). In this paper, we concentrate on the aspect of adaptation to different levels of cognitive workload. Most approaches and techniques described here are relevant for other adaptation tasks as well.

2 Driving Simulator & Experimental Setup

Development, testing and evaluation of different interaction strategies requires a realistic experimental environment which reproduces all important effects and distractions seen in real-life applications. While recording in a real car in real traffic situations creates the most authentic sessions, the downsides of this approach are safety concerns with early prototypes, the lack of reproducibility and the missing ability of reliably provoking scenarios which are relevant for the current investigation.

Therefore, we decided to build a driving simulator which is designed to create a realistic driving experience. The main focus was not to build a physically correct car test bed but to simulate the most important influences and distractions that occur during real driving tasks, especially in situations where the application of a dialog system plays an important role. We based our driving simulator on a real car and kept the interior fully intact and functional to provide a realistic in-car feeling. The car is surrounded by a projection wall, covering the view of the frontal and lateral windows. The simulator features acoustic feedback via engine sound and environmental surround sound and haptic feedback in the seat (via tactile transducers) and steering wheel (via Force-Feedback). We installed a display in the driver's cockpit to provide means for visual distraction as produced by graphical user interfaces and to display a 3D avatar as a visual representation for the interaction system.

Image 1: The CSL driving simulator in action

For studies with different main focus, we use different simulation software and scenarios. For investigation of interaction patterns and dialog strategies, the employed simulator software is based on a modified gaming engine[1]. It was extended with multi-screen display, steering wheel support and simple ambient traffic control. Its support for scripting scenarios allows us to configure individual driving stages: We can position the driver in a wide artificial environment with realistic urban and rural areas, where we define a route represented by navigation directions for the system.

As simulation software for experiments on cognitive workload classification, we employ the established Lane Chance Task (LCT) as primary driving task (Mattes, 2003). The LCT asks

[1] MTA:SA: http://www.mtasa.com

the driver to follow a highway with a fixed speed, respecting signs that mark only one of three lanes as valid, which forces the driver to change lanes every few seconds. During one experiment, the driver goes through multiple differently configured driving sessions, including some training sessions to familiarize himself with the simulator and the typical tasks during the experiment. Each session lasts for three minutes. The whole set of all sessions is designed to cover a variety of different types and levels of cognitive workload. The driving task itself is of low complexity compared to many real-life situations (especially in urban environments) as it does not feature other cars, intersections or other complex traffic elements. Instead, the driver's cognitive load is controlled by additional secondary tasks. We use a visual task (included in the LCT) that asks the driver to identify certain symbols on a display installed in the cockpit and an arithmetic task that asks the driver to classify numbers given by a prerecorded voice according to their divisibility by certain fixed values. Both secondary tasks are designed to support multiple difficulty levels and we record both answer quantity (number of given answers) and answer quality (percentage of correct answers). In addition, the LCT software contains a simple model to assess the driving quality. It calculates an ideal route determined by the position and type of the lane change requests. This model is compared to the actual route of the driver to calculate an error measure that describes the driving quality.

The difficulty of the secondary task yields a natural label for the recorded sessions for evaluation purposes on cognitive workload. However, experiments show that not all users follow the expected error pattern of performing worse on more difficult tasks. This is due to training effects beyond the initial training phase, fatigue or individual influences like emotions. As we want to assign session labels that reflect not the expected but the actual performance, we prefer error scores over task difficulty levels as objective label. To define an error based workload measure, we combine the scores for answer quantity, answer quality and driving quality into a vector, z-normalize its components to ensure equal distributions along all three dimensions and use an euclidean distance or city block distance to compare two sessions. The vectors are then sorted according to this metric and clustered to form a small number of classes (e.g. low, medium, high). Scaling of the individual dimensions allows to weigh the three performance aspects differently, e.g. by giving the driving quality a larger impact on the overall score.

As we are also interested in the subjective workload impression as an indication for user satisfaction, we also collect questionnaire data on different scales of workload. These questionnaires are handed to the driver immediately after each driving session so we can extract a label for it. Differences between subjective and objective workload scores are for example visible for extremely difficult tasks, where the driver gives up on (parts of) his assignment, reducing subjective workload by basically removing a task. In contrast, the objective workload deduced from error measures or task difficulty does not fall (or even rises), leading to a large gap between both values. To test subjective workload, one has to select form different subjective workload models, ranging from simple single-dimensional models to sophisticated multi-dimensional scales. Two of the most established ones are the Workload Profile (WP) (Tsang, 1996) and the NASA Task Load Index (NASA-TLX) (Hart, 1988). While a comparative work (Rubio, 2004) recommends WP as slightly better, it is not applicable without extensive training and explanation. In small preliminary user tests, many test persons were unclear about several aspects of the WP questionnaire. In contrast, the NASA-TLX is intuitive, quick to answer and also offers reasonable discriminative power. It is therefore suited for repeated application. One drawback in comparison to the WP is the lack of discrimination between different modalities (e.g. vocal vs. manual output) and cognitive processing schemes (e.g. spatial vs. symbolic).

3 Recording Setup & User State Classification

During the experiment, we employ a variety of signals to observe the user in the car. This is done for multiple reasons: First, an adaptive dialog system needs data streams from which it can extract meaningful features describing the user's state. Second, to train automatic recognizers that perform this user state classification, we need to provide large amounts of labeled training data. To that end, we installed multiple biosignal sensors in the car to get a reliable, continuous data stream without obstructing or distracting the user too much. We employ the following equipment to observe the user:

- Small cameras to record videos of the face and the upper body of the driver to catch facial expressions and body pose

- A close-talking microphone to record the user’s utterances

- A comfortable headband to record electroencephalography (EEG) data of the prefrontal cortex (positions Fp1, Fp2, F7 and F8 in the international 10-20 positioning system)

- A light sensor glove which measures skin conductance and pulse (via plethysmography)

- A respiration belt on top of the clothes to measure respiration frequency

Image 2: (Part of) the recording setup with EEG headband and headset in the driving simulator

The last three items all use the same recording interface and are either attached to a universal signal recorder[2] or directly connected via Bluetooth, which reduces obstruction to a minimum. In addition, we employ indirect motion monitoring by continuously recording the angle of the steering wheel and the acceleration and brake pedals in the car.

[2] VarioPort, Becker MediTec

To record all biosignal streams in a synchronized fashion with support for arbitrary input and output formats, time-stamping, distributed recording and convenient logging and storing of session packages, we employ our modular Biosignal-Studio software (Putze, 2009).

The collected biosignal streams are passed on to a set of statistical classifiers that estimate the current states and traits of the user which are relevant for system adaptation. Currently, we implement a classifier for cognitive workload and a recognizer of personality traits like extroversion or emotional stability. Both classifiers use the same framework of preprocessing (artifact detection and removal, windowing), feature extraction, feature selection and reduction via Forward Feature Selection and classification with Support Vector Machines. We perform feature fusion with a subsequent Linear Discriminant Analysis on the joint feature space. For feature extraction, we mostly rely on well established routines: We use the software Praat[3] to extract prosodic features like pitch, jitter or shimmer from the user's voice. The single-channel biosignals respiration, skin conductance and pulse are all treated using a similar processing chain to smooth the signal, extract peaks, calculate sliding means and variances as well as derivations as dynamic features.

4 Cognitive Model

For user models in interactive systems, there are two different applications: We need user state models to represent relevant information on the user during the interaction. This does not only comprise the ability to store the recently observed user states, but the model may also be capable to predict future values or estimate states that are not directly observable, like the state of the user's memory. As this model must be applicable to all users that interact with the system, we start with a general stochastic model, representing the "average" user. Observations made by the system during the interaction introduce bias for a certain state or type of user in this model and modify the prediction probabilities of future observations, behavior, internal state etc. The other major application for user (behavior) models is user simulation where one predicts user behavior given the context of an ongoing interaction. User simulation is employed for evaluation and training of dialog systems and interaction strategies where the user simulation replaces expensive trials with real users. For user simulation, the user model is not an averaged representation of all users but a personalized model of a single individual. This is necessary so that the system is confronted not only with one standard behavior but instead with very different types of users.

In traditional systems, user state models and user behavior models are separated and based on quite different approaches: While user state models typically are little more than a collection of user state variables (e.g. as in (Gnjatović, 2008)), user behavior models are mostly behavioristic models represented as statistics of user actions dependent on (a short-time window of) the discourse (Eckert, 1997). We propose to bring both types of user models together as the simulation model profits from a model of user states to coherently adjust the simulated user's behavior to the simulated user state. This is a typical problem of purely statistical user behavior models that do not maintain a representation of the mental state of the simulated user, often leading to arbitrary and unpredictable behavior. How this fusion can be done is explained with an example in the following section.

[3] http://www.praat.org

For representing the user in a realistic way, we need *cognitive* user models, i.e. user models that take into account the cognitive processes in the human mind during the interaction. Ideally, a cognitive model would reflect the whole complexity of human cognition in a general, comprehensive way. This is the goal of cognitive architectures like ACT-R (Anderson, 2004) or many others which already can predict many well-known phenomena of the mind. Most current applications for ACT-R models however only work in controlled environments, using strong assumptions and mostly ignoring individual differences like emotions. Until these models progress to a more mature stage, we take a different stance, using individual cognitive components to represent certain aspects of our user model. The following section will present two examples for this approach.

4.1 Urges as Cognitive Motivators

While interacting with a dialog system in a dynamic scenario, the user pursues different, often adversarial goals. For example, he is trying to extend his domain knowledge but this will often conflict with the desire to maintain a low cognitive workload, especially if most resources are already occupied by the driving task. He also has the desire to understand and be able to predict the system he is interacting with, which limits the options of the system to react to the cognitive overload problem by constantly adjusting its behavior. We model these different goals using the concept of urges as proposed by the PSI cognitive architecture (Bach, 2003). These urges describe different desires on a scale from zero (completely satisfied) to one (extremely high demand). Once a desire exceeds a certain threshold, it creates motivations which trigger behavior that aims at changing the situation which causes the increased urge. This way, urges regulate the decision making as they function as weights for several possible actions. They also help the dialog system to capture the most relevant problems and an overall representation of the user's "well-being". The user tries to maintain a homeostasis of all urges being low or within certain tolerable limits. This mechanism drives and influences the user's behavior and at the same time gives an indication of which urges are most critical or urgent to cope with.

The currently implemented urges for a dynamic interaction scenario are (in analogy to the original urges from the PSI architecture): *Information competence*, i.e. the desire to gather domain information, is calculated by inspecting the memory model of the user (see below) to derive his interest in additional information. *Inertia*, i.e. the desire to maintain a status of low cognitive workload, is directly derived from the cognitive workload variable of the user model. And finally, there is *interface competence*, i.e. the desire to interact with a predictable interface that follows the general implicit rules of spoken interaction, depends on two factors: First, we count the number of violations agains a predefined set of interaction rules (avoid barge-in, complete utterances). In addition, an adaptive expectation model stores statistics on the expected system reactions to user actions.

The main influence of the urges on the behavior of the user and the system is on action selection: They offer reward functions that are used to evaluate past actions which allows the system and the user simulation to learn from observed interactions. Both have access to the same set of urges, although the two systems might see different values as its observation of the simulated user is noisy. The learned feedback is stored and reviewed later in similar situations. This process is formalized as reinforcement learning, which is already established for cognitive behavior modeling (Wai-Tat, 2006), however not in the context of man-machine-interaction.

4.2 Memory and Cognitive Workload

In the current cognitive model, cognitive workload is represented as a single value on a continuous scale. Two sources influence this value: During interaction, it basically represents the output of the multimodal workload classifier. During simulation (where no actual sensor is present), the value is determined by triggering scripted or randomly generated workload events that stand for difficult traffic situations or ongoing speech understanding. When no event is triggered, the workload gradually decreases. Future effort will go into the combination of both sources, e.g. using Bayesian networks or tracking schemes which can combine model prediction and observations via workload sensors. To include the workload variable in the state space of the system (see chapter 5), we discretize and normalize the cognitive workload variable based on previous observations of the raw values so that occurring workload values are equally distributed among all classes.

The cognitive workload estimate from the biosignal sensors and the cognitive model influences the user model and the dialog strategy in different ways. The user model (which according to chapter 4 also contains a user behavior model) selects different actions and speech act representations based on its own cognitive workload level. The higher the workload level, the simpler the employed actions. The average cognitive workload level during a system utterance also determines the quality of the user's speech understanding: A high average workload (in combination with a high complexity of the system utterance) leads to a high probability of non-understanding. Future work will also introduce partial understanding (in the case of short-time changes of workload) and the possibility of misunderstandings. The workload is (indirectly, see below) also propagated to the system and is used by the interaction strategy to determine the currently appropriate complexity of system moves.

For simulation purposes, we also model the output of virtual biosignal sensors. This value is based on the "true" value of the workload attribute as represented by the cognitive simulation model but skewed by an error model that reflects noise and systematic errors typically made by statistical user state classifiers. The output of the virtual sensor is visible to the system instead of the true value which is only available to the user model during simulation, e.g. to determine its influence on action selection. Currently, the recognition error is modeled as additive white Gaussian noise.

5 Dialog Strategy Adaptation

Adaptation of dialog behavior cannot be seen as a short-term task but must instead be considered from a strategic, long-term perspective. This is the case even if all decisions for themselves are local in nature (e.g. changing the system voice for a single utterance), because every adaptation comes with costs: decisions based on noisy signals may be wrong, too frequent adaptation (even if locally appropriate) can confuse the user, obfuscate the interface or produce the impression of an inconsistent system persona. In addition, there are decisions which are inherently strategic in nature since they cannot be undone, for example the decision to cancel a complete subdialog. We believe that the manual design of adaptation strategies based on a complex cognitive user model is unfeasible. Therefore, we employ reinforcement learning (Sutton, 1998), an automatic learning technique, to discover optimal strategies for a given dialog and behavior model. This has been successfully done for strategies of dialog systems on a speech act level (Singh, 2002) (Scheffler, 2002) and can be extended to adaptive systems. We built a learning framework which uses modularization and user simulation based on cognitive models and the concept of urges. This process requires several steps: First, one has to define a suitable state space that represents the relevant components of the cognitive user model in a compressed and discretized form. One has to define a suitable action space that contains enough parameters to allow the system to modify the surface form of its speech acts. As not all adaptation measures take place on the level of a single dialog move, also the general conditions of move execution (e.g. when and how often new moves can be executed) have to be flexible enough.

As described in Chapter 4, cognitive models can be used to simulate coherent user behavior. This is used for evaluating dialog sessions, but can also be employed to simulate dialog sessions for training purposes. As reinforcement learning requires a large number of training epochs which cannot all be recorded with real users, user simulation is employed to generate interaction sessions from which rewards are deduced and attributed to the system actions leading to them.

To enable adaptation to user states, one has to provide different means of influencing the outputs of the system. While classical systems usually only allow the selection of several predefined utterances, dialog strategies for dynamic environments need to be more flexible. Our current system supports the following adaptation techniques, enabling it to adjust its behavior to the user's current state.

- Selection of type and timing (to the second) of the next system speech act. Using small time slices as a dialog timing model allows to precisely select when the next utterance is issued. With this freedom, the system can delay non-critical information in situations of high cognitive workload.
- The system is able to abort own speech acts and interrupt utterances of the user. This is again supported by the temporal slicing that enables the system to review the state of user and interaction at a very fine level. If a change in the user state indicates the necessity of immediate reaction, the ongoing speech act can be interrupted. This feature comes with support for partial understanding of interrupted utterances.
- The system can change the system voice by sending parameters to the speech synthesis component that determine the basic voice and other parameters like pitch or speed. Instead of directly modifying those technical parameters, the system uses a mapping (provided by the synthesis component) to map a desired emotional voice scheme, expressed on a discretized version of the dimensional model of affect (Russel, 1980) to those technical parameters. This model allows to select from calm and excited voices by changing the activation parameter.
- The system can select utterances based on a (currently manually assigned) complexity score. This score represents an estimate on how difficult it is to understand the given utterance and may reflect utterance length, number of items mentioned in this utterance or

linguistic cues of complexity. From the cognitive model and the observed workload, we derive an optimal complexity score (as high as possible without risking cognitive overload) which is compared to the complexity of the utterance. This distance is used as a criterion to weigh all available speech acts.

6 Conclusion

Although we only recently started our studies, we already made several steps towards our goal of flexible, generic and natural adaptation mechanisms: We implemented and tested a realistic driving simulator which will allow a large number of experiments under controlled but nevertheless authentic conditions. We are working towards a framework of statistical classifiers that are able to determine the user's current state. We investigate cognitive modeling architectures to structure the user's adversarial desires and to model the user's cognitive load.

For future developments, we need to extend our work on these components such that larger quantitative user evaluation in our scenario becomes possible. These user studies are necessary to study the influences of different adaptation schemes on the users' satisfaction and to train and evaluate the user state classifiers.

In addition, cognitive models need to be further investigated to achieve a tighter integration of all components and a coherent model of all relevant cognitive processes during the interaction. This will improve the predictive power of the models and improve their contribution to cognitive dialog systems.

7 Literature

Litman, D. & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. *In Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Boston, USA*

Gorin, A.; Riccardi, G. & Wright, J. (1997). How may I help you? In: *Speech Communication 23*

Nass, C.; Jonsson, I.-M.; Harris, H.; Reaves, B.; Endo, J.; Brave, S. & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In: *CHI '05 extended abstracts on Human factors in computing systems, Portland, USA*

Nass, C. & Lee, K. M. (2000). Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In: *Proceedings of the SIGCHI conference on Human factors in computing systems, The Hague, The Netherlands*

Mattes, S. (2003). The lane-change-task as a tool for driver distraction evaluation. In *Proceedings of IgfA*

Tsang, P. S. & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. In: *Ergonomics*, Volume 39, Issue 3, pp. 358 - 381

Hart, S. & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P. & Meshkati, N. *(ed.). Human mental workload*, Amsterdam: North Holland B.V., pp. 139-183

Rubio, S.; Diaz, E.; Martin, J. & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. In: *Applied Psychology*, 53(1), pp. 61-86

Putze, F. & Schultz, T. (2009). Cognitive Dialog Systems for Dynamic Environments: Progress and Challenges. In: *Proceedings of the 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety, Dallas, USA*

Sutton, R. S. & Barto, A. G. (1998). Reinforcement Learning: An Introduction *MIT Press, Cambridge*

Singh, S.; Litman, D.; Kearns, M. & Walker, M. (2002). Optimizing Dialogue Managment with Reinforcement Learning: Experiments with the NJFun System. In: *Journal of Artificial Intelligence Research* 16

Scheffler, K. & Young, S. (2002). Automatic learning of dialogue strategy using diaogue simulation and reinforcement learning. In: *Proceedings of the second international conference on Human Language Technology Research*

Bach, J. (2003). The micropsi agent architecture. In: *Proceedings of the 5th International Conference on Cognitive Modeling*

Gnjatović, M. & Rösner, D. (2008). Emotion Adaptive Dialogue Management in Human-Machine Interaction: Adaptive Dialogue Management in the NIMITEK Prototype System. In: *19th European Meetings on Cybernetics and Systems Research*

Eckert W., Levin, E. and Pieraccini, R. (1997). User modeling for spoken dialogue system evaluation. In *Proceedings of the IEEE ASR Workshop*

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C. & Qin, Y. (2004). An integrated theory of the mind. In: *Psychological Review 111*, pp. 1036-1060

Wai-Tat, F. and Anderson, J.R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. In: *Journal of experimental psychology*, vol. 135, no. 2

Russell, J. A. (1980). A circumplex model of affect. In: *Journal of Personality and Social Psychology*, Vol 39(6), 1161-1178