A Multi-modal Dialog System for a Mobile Robot

Ioannis Toptsis, Shuyin Li, Britta Wrede,
Gernot A. Fink

Faculty of Technology, Bielefeld University
33594 Bielefeld, Germany
{itoptsis, shuyinli, bwrede, gernot}@techfak.uni-bielefeld.de

Abstract

A challenging domain for dialog systems is their use for the communication with robotic assistants. In contrast to the classical use of spoken language for information retrieval, on a mobile robot multimodal dialogs and the dynamic interaction of the robot system with its environment have to be considered. In this paper we will present the dialog system developed for BIRON – the Bielefeld Robot Companion. The system is able to handle multi-modal dialogs by augmenting semantic interpretation structures derived from speech with hypotheses for additional modalities as e.g. speech-accompanying gestures. The architecture of the system is modular with the dialog manager being the central component. In order to be aware of the dynamic behavior of the robot itself, the possible states of the robot control system are integrated into the dialog model. For flexible use and easy configuration the communication between the individual modules as well as the declarative specification of the dialog model are encoded in XML. We will present example interactions with BIRON from the "home-tour" scenario defined within the COGNIRON project.

1. Introduction

In Human-Computer Interaction (HCI) the ultimate goal of research is to make the interaction with intelligent devices more "natural", i.e. intuitive and easy to use for humans. In human-human communication spoken language can be considered the most natural and effective means of communication, though it frequently is complemented by other modalities, e.g. mimic or gesture. Therefore, spoken language dialog systems are applied in many areas of HCI to achieve a natural communication.

The classical domain of dialogue systems are telephony-based services. Such systems mainly enable human users to access information stored in some database by using spoken language only. During the interaction the dialog system is in complete control of the information appliance.

A radically different and extremely challenging new domain for dialog systems is their use in so-called *robot companions* – mobile robots serving humans as assistants in private homes and eventually even as companions during everyday life. The communication with such complex devices can not be limited to spoken language only but has to take into account all modalities used in human-human dialogs, such as gesture or the expression of emotions. Furthermore, the robot's behavior is not only dependent on the communication with the user but also on the rather complex interaction of the mobile platform and its environment. Therefore, the dialog system can not be the central control unit of the robot companion. It will, however, be the central interfacing component between human users and the robot control system.

In this paper we will present the design of the dialog management system of BIRON – the Bielefeld Robot Companion [1]. It uses speech as the main modality for communication but is also able to augment information presented by spoken language with hypotheses derived from additional modalities, as e.g. in the case of speech accompanied by deictic gestures. As the dialog manager is not the central control unit of BIRON the internal state of the robot control system is periodically communicated with the dialog manager. Commands to the robot are derived from multi-modal semantic interpretation structures for dialog acts. Depending on the current state of the robot control unit the dialog manager can decide early about the possibility to perform actions required by the user or inform him about the internal state of the robot in case of communication problems.

The development of BIRON is currently focused on the scenarios defined within the COGNIRON project. One of the key experiments there is the so-called *home-tour*, where a robot companion is shown around a user's private home in order to familiarize it with this new environment.

In the following sections we will first review some related work on dialog systems with emphasis on systems used for the interaction with mobile robots. Then we will in detail describe the design of the dialog manager developed for BIRON covering the general architecture, the dialog model used, and the integration with the robot control system. In section 4 we will outline the capabilities of the current dialog model and present an example dialog with BIRON.

2. Related Work

The first generation of dialog systems, and also the majority of dialog systems today, only handle speech input since spoken language is the most important modality in human-human interaction. The dialog-system presented in [2] is applied to information retrieval tasks and employs a *slot-filling* strategy. A slot is an information item for which a value is required. The dialog system collects information from the user by filling slots to reach the dialog goal. This way, the system is able to support implicit verification of application responses, which reduces the duration of the dialog. The dialog model developed at AT&T [3] defines states and actions which is similar to our approach. However it employs a stochastic dialog strategy which can automatically be adapted by reinforcement learning. Also, the slot-filling technique is used to collect information for database inquiries as in [2]. The PHILIPS dialog system [4] is designed, among others, for portability. Therefore, it is application independent and based on a modular architecture like our

*The work described in this paper was partially conducted within the EU Integrated Project COGNIRON ('The Cognitive Companion") and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020. It was also partly funded by the German Research Foundation (DFG) within the Graduate Program 'Task Oriented Communication'.

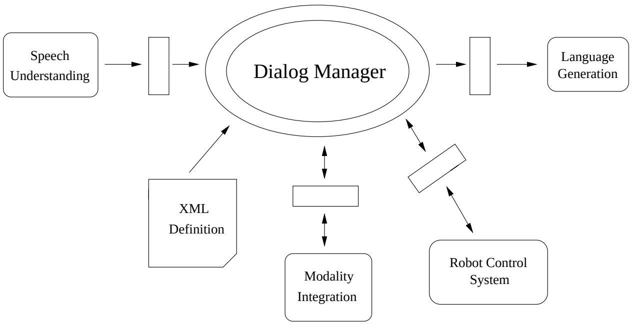

Figure 1: The dialog system of BIRON with the interface to the robot control system

system. For dialog control and speech understanding a definition language called *high-level dialogue description language* (HDDL) is used. With HDDL it is possible to divide the whole dialog into sub-dialogs, so called HDDL modules. This system is mainly used in automatic inquiry applications, where only spoken language is supported.

In the recent years the research of intelligent interfaces has focused on multi-modal dialog systems. The support of additional modalities enhances the robustness and the naturalness of the HCI. A representative of such interfaces is the MASK-kiosk [5]. It can handle multi-modal travel inquiries in form of spoken language and pointing on a touch screen. However, this kind of gesture can only approximate natural gesture used in human-human communication to a certain degree. In this system both modalities are fused on a semantic level inside the dialog-manager while in our system the fusion is achieved by a separate component. A multi-modal user registration system is presented in [6]. The dialog-manager contains states and actions and is similar to ours. But in this system the action to be taken does not depend on the state as in our system, but on the transition. Furthermore, the integration of the speech understanding and the modality fusion into the dialog-manager differs from our system. Information collected by different modalities is fused via a Bayesian network.

At present, only a small number of dialog systems supports intelligent human-machine interaction for mobile robots because of the higher complexity and dynamics of the task and the underlying system. The dialog system developed for an autonomous robot helicopter within the WITAS project [7] applies a combination of spoken language and pointing on a map. Its goal directed dialog strategy is not based on the slot-filing method and dialogs are open ended. The Hygeiorobot [8] is a mobile robotic assistant for hospital use. It can fulfill tasks like delivery of medicine or message and replying of inquiries of information about patients. Its uni-modal dialog system is state-based and designed to perform relatively short dialogs only. CARL is a mobile service robot [9], that is able to process input in form of spoken language and pointing gestures on a touch screen. Its system differs from ours in two points: First, their state-based and event-driven dialog-manager interprets user input via high-level reasoning. Second, the human-robot communication is modeled as an exchange of messages.

3. Dialog Manager

In the following, we first present the architecture of the dialog system developed for BIRON and then describe the dialog model in detail. We will close this section by emphasizing our system's capability of handling the internal robot states directly.

3.1. System Architecture

In many dialog systems the dialog manager is merged with other components, e.g. with speech understanding. This can lead to heavy dependencies of the dialog system on the application. We developed a modular architecture that separates the dialog management from speech processing as shown in Figure 1. The dialog manager is the main component of the dialog system and is also the focus of this paper. It communicates with other components over well defined interfaces, which use XML-structures for data exchange. This modular architecture of the dialog system enhances its portability.

The dialog manager receives the result of the semantic analysis of the speech input from the speech understanding component. In case that the semantic structure indicates the involvement of other modalities, the dialog manager will consult the modality integration component for further information. Consider the following example: The user says ”This green cup” while pointing to it. The semantic structure delivered by the speech understanding contains anaphora ”this” which indicates a possible involvement of gesture. The dialog manager then sends a request to the modality integration component to ask for integration of the semantic structure and the possible gestural information that can specify which object, in this case, which green cup, the user meant. Feedback to the user can be presented by the language generation module.

The dialog manager interprets the user's commands and sends them to the robot control system for execution. The robot control system is an independent component and can only process commands if the current status of the overall system allows it. Therefore, we implemented the control flow in a bidirectional way: The dialog manager sends user commands to the robot control and periodically receives messages from the robot control reporting its current status. Thus, the robot control system is not under control of the dialog system, but an equal ”partner” of it.

3.2. Dialog Model

The model of the dialog manager is based on a *Finite State Machine* (FSM) that is extended with the ability of recursive activation of other FSMs and the execution of an action in each state. Actions that can be taken in certain states are specified in the *policy* of the dialog manager.

The implementation of the dialog manager is based on the so-called *slot-filling* strategy [2]. A slot is an information item for which a value is required. The task of the dialog manager is to fill enough slots to meet the dialog goal, which is defined as a goal state in the FSM. This can be viewed as a quantization of the semantic content of user's utterance into the required information items. Every state of the model is determined by the status of its slots. The slots can be empty, be filled with an attribute, or have logical values *true* or *false*. The incoming information from the user and the robot control system fills the slots, which are categorized into three sections and collected in a so-called dialog frames as shown in Figure 2. The USER section contains information provided by the user, the SYSTEM section represents the internal status of the robot control (see subsection 3.3 for details) and the CONTROL section contains items for internal use of the dialog manager.

The slot-filling technique alone is not powerful enough to support the complex interaction scenarios in robot domain [10]. To overcome this limitation we modeled the dialog in a modular way by dividing the dialog into sub-dialogs. Each sub-dialog is associated with a task and is modeled as a separate FSM. Each FSM has a goal state which indicates the completion of the current task. The processing of each sub-dialog can be interrupted by another sub-dialog, which enables alternated instruction processing. The interrupted sub-dialog can be resumed later.

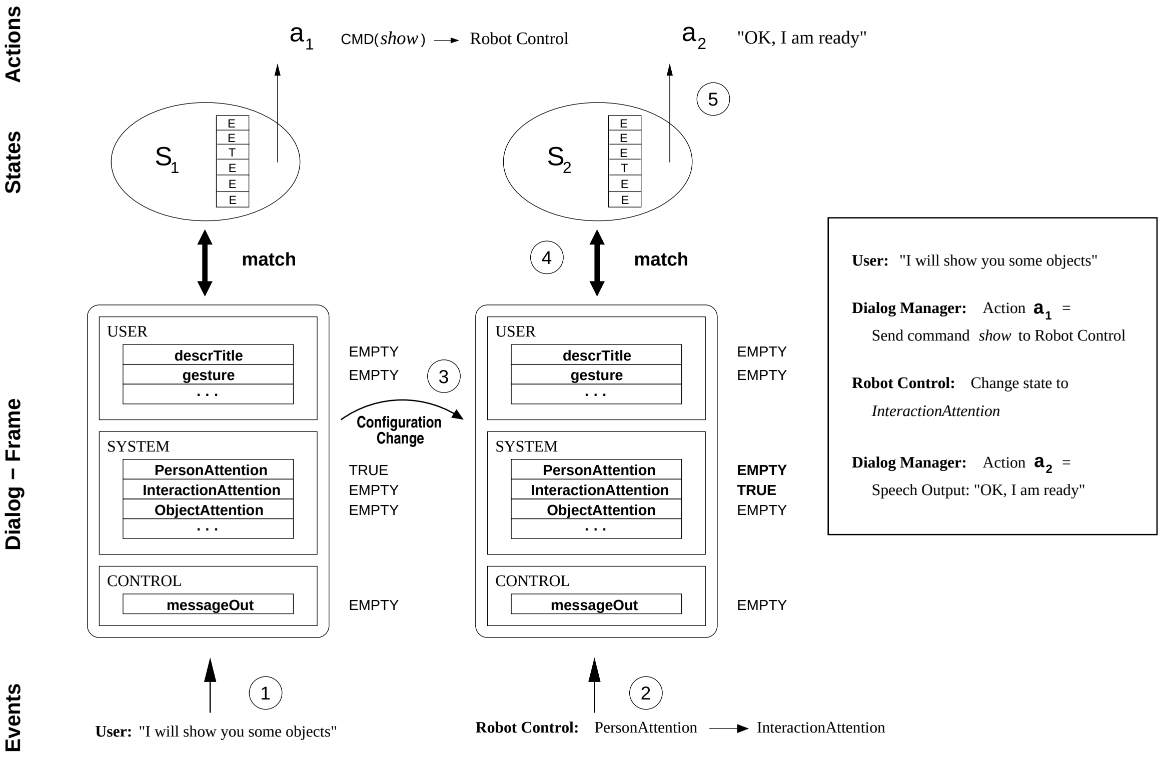
Actions
States
Dialog – Frame
Events
a_1 CMD(*show*) → Robot Control
a_2 "OK, I am ready"
S_1
E E T E E E
S_2
E E E T E E
match
match
1
2
3
4
5
USER
descrTitle
gesture
...
SYSTEM
PersonAttention
InteractionAttention
ObjectAttention
CONTROL
messageOut
EMPTY
EMPTY
TRUE
EMPTY
EMPTY
EMPTY
EMPTY
EMPTY
EMPTY
TRUE
EMPTY
EMPTY
Configuration Change
User: "I will show you some objects"
Robot Control: PersonAttention → InteractionAttention
User: "I will show you some objects"
Dialog Manager: Action $\mathbf{a_1}$ = Send command *show* to Robot Control
Robot Control: Change state to *InteractionAttention*
Dialog Manager: Action $\mathbf{a_2}$ = Speech Output: "OK, I am ready"

Figure 2: Dialog part with internal actions of the dialog manager and the structure of the dialog frame

The dialog management is event-based. Switching between the dialog states depends on the status composition of all slots in the dialog frame. In an ongoing dialog, the dialog manager compares slots in the newly updated dialog frame with those in the FSM to find out in which state the current dialog is. According to the specification of the state-action-association in the *policy* the appropriate action is executed, e.g., generation of speech output or sending events to the robot control system. We will illustrate this process with an example in subsection 3.3.

The dialog model is defined in a declarative definition language which is encoded in XML. This increases the portability of the dialog manager and allows an easier configuration and extension of the defined dialogs.

3.3. Integration of Internal Robot States

In robot applications it is important for the dialog manager to be informed about the current status of the sensory-motor system of the robot. This is often realized via message exchange in a multi-agent system in other applications [9]. Our approach is to integrate internal states of the robot control into the dialog model by representing the states of the robot control as an FSM. In an ongoing dialog, the current status of the robot control is represented as slots in the SYSTEM section of the corresponding dialog frame. Therefore, the dialog manager is permanently ”aware” of the status of the robot control.

In Figure 2 we demonstrate our approach with an example. Suppose the robot control system is in *PersonAttention* status, this means, that the robot is ready to start communication with the user. This status is represented as the slot PersonAttention in the SYSTEM section in the left dialog frame, its value is set to TRUE. The user speech input ”I will show you some objects” activates the sub-dialog ”show” and the result of its comparison with the current dialog frame is the state S$_1$. The associated action a$_1$ ”send command ’show’ to robot control” is then triggered as specified in the *policy*. After receiving this message the robot control system changes its status from *PersonAttention* to *InteractionAttention* which results in a change in the corresponding slots in the dialog frame’s SYSTEM section. After the match between this updated dialog frame with the sub-dialog ”show” the action a$_2$ is triggered. The robot generates the utterance ”OK, I’m ready!”.

The integration of the robot control states into the dialog model has several advantages. The dialog manager has dynamic knowledge about the abilities of the robot control system and can immediately make the decision if a certain user request can be processed or not without a transmission to the robot control. This reduces the reaction time of the robot. Another advantage is that the information about the task currently processed by the robot control system are available for the dialog manager. In case that the user tries to interrupt the current task the robot can give detailed information about the robot’s current status. This information can also be used to maintain the communication during long-term actions, e.g. by informing the user periodically about the current status of the task.

4. Scenario and Dialogs

Within the COGNIRON project we are currently implementing the home tour experiment. The central idea of this scenario is that a robot is delivered at home where the user familiarizes it with the environment by showing it different rooms and objects. During the home tour the robot should build internal representations of the environment and objects.

We have implemented five sub-dialogs for this scenario: (1) Greeting: the user logs into the system with common greeting phrases like “Hello”. The dialog manager sends the command “register” to the robot control system that changes its status from *Per-*

sonAlertness to *PersonAttention*. The robot then registers the user as an active communication partner and centers its focus on the user. (2) Parting: the user logs out of the system with common parting phrases like "Goodbye". The corresponding dialog manager command is "checkout" and the status of the robot control system is set back from *PersonAttention* to *PersonAlertness*. The robot returns to its standby mode. (3) Person following: the user can activate this function by saying "Please follow me". The dialog manager's command "follow" results in a status transition of the robot control system from *PersonAttention* to *PersonFollow* and the robot starts to follow the user. (4) Initiating gesture detection[1]: gesture detection can be triggered by user commands like "Look" or "I will show you some objects", which activate the dialog manager's command "show". This command changes the status of the robot control system from *PersonAttention* to *InteractionAttention* and the robot turns its camera to the direction of the user's hand. (5) Initiating object detection: The robot looks for the corresponding object in its current camera view if the user says, e.g., "This is a TV set". This process is initiated by the dialog manager's command "describe" and the following status transition of the robot control system from *InteractionAttention* to *ObjectAttention*.

In the following we illustrate the described procedures with a dialog example. (U: User; R: Robot, DM: dialog manager, RC: robot control)

U: Hello BIRON!
(DM: register, RC: PersonAlertness ⇒ PersonAttention)
R: Hello, what can I do for you?
U: Please follow me.
(DM: follow, RC: PersonAttention ⇒ PersonFollow)
R: OK, I'm following.
U: I will show you some objects.
(DM: show, RC: PersonAttention ⇒ InteractionAttention)
R: OK, I'm ready.
U: This is my TV set.
(DM: describe, RC: InteractionAttention ⇒ ObjectAttention)
R: OK, I can see it.
U: Thank you, BIRON, Good-bye.
(DM: checkout, RC: ObjectAttention ⇒ PersonAlertness)
R: Bye-bye.

As shown above, our system design can help to ensure smooth cooperation between the dialog manager and the robot control system and thus improve the robot's performance as a whole.

5. Conclusion

In this paper we presented the dialog system developed for the mobile robot BIRON. It assumes that speech is the main modality used for communication. However, the system is able to augment the semantic representations derived from user utterances by hypotheses for additional modalities as e.g. speech-accompanying gestures. The central component of the system is the dialog manager which communicates with its supporting modules via well defined interfaces using XML-encoded data structures. XML is also used for the declarative definition of the dialog model. As the dialog manager is not the central control unit of BIRON the internal states of the robot control system are periodically communicated and integrated into the current configuration of the dialog. In the current implementation a dialog model for the so-called "home-tour" scenario is defined[2].

6. References

[1] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proceedings International Conference on Multimodal Interfaces*. Vancouver, Canada: ACM, November 2003, pp. 28–35.

[2] B. Souvignier, A.Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant - strategies for spoken dialog systems," in *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, 2000, pp. 51–62.

[3] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human machine interaction for learning dialog strategies," in *IEEE Transactions on Speech and Audio Processing*, 2000.

[4] H. Aust and O. Schröer, "An overview of the PHILIPS dialog system," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[5] J. Gauvain, S. Bennacef, L. Devillers, L. Lamel, and S. Rosset, "Spoken language component of the MASK kiosk," in *Human Comfort & Security of Information Systems*, K. Varghese and S. Pfleger, Eds. Springer, 1997, pp. 93–103.

[6] F. Huang, J. Yang, and A. Waibel, "Dialogue management for multimodal user registration," in *Proceedings International Conference on Spoken Language Processing*, Beijing, China, October 2000.

[7] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "A multimodal dialogue system for human robot conversation," in *Proceedings North American Chapter of the Association for Computational Linguistics*, Pittsburgh, USA, June 2001.

[8] D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proceedings of the European Workshop on Service and Humanoid Robots (ServiceRob '2001)*, Santorini, Greece, 25-27 June 2001.

[9] L. S. Lopes, A. Teixeira, M. Rodrigues, D. Gomes, C. Teixeira, L. Ferreira, P. Soares, J. Giro, and N. Snica, "Towards a personal robot with language interface," in *Proceedings European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003.

[10] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "The WITAS multi-modal dialogue system I," in *Proceedings European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001, pp. 1559–1562.

[1] Currently, the gesture detection is not yet integrated in our system.

[2] A video of an example interaction with BIRON using German language can be found on our web site http://www.techfak.uni-bielefeld.de/ags/ai/projects/BIRON/.