Acquiring and Maintaining Knowledge by Natural Multimodal Dialog

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)

genehmigte

Dissertation

von

Hartwig Holzapfel

aus Karlsruhe

Tag der mündlichen Prüfung:	29. Juni 2009
Erster Gutachter:	Prof. Dr. Alex Waibel
Zweiter Gutachter:	Professor Rolf Carlson

ZUSAMMENFASSUNG

Dialogsysteme kommen heutzutage in einer Vielzahl von Anwendungsgebieten zum Einsatz. Hierzu gehören Navigationssysteme, Call Center, diverse interaktive Systeme und in den letzten Jahren, verstärkt in der Forschung, auch Dialogsysteme für die Mensch-Roboter Interaktion. Eine Beschränkung, die diese Systeme trotz signifikanten Fortschritts in den letzten Jahren noch innehaben, ist fehlende Adaptionsfähigkeit, da diese Systeme durch statische Vorgaben, statische Systemkomponenten und statische Wissensmodelle beschränkt sind. Im Gegensatz dazu sollte die nächste Generation von Dialogsystemen in der Lage sein, ihre Strategien und Wissensmodelle zu adaptieren und dadurch in die Lage versetzt werden, sich an ihre Umgebung anzupassen. Eine derartige Adaptionsfähigkeit ist z.B. für einen humanoiden Roboter von hohem Nutzen, insbesondere dann, wenn die Adaption durch das System autonom und ohne den manuellen Eingriff eines menschlichen Operators durchgeführt werden kann.

Diese Arbeit bildet einen Beitrag für das Fernziel, komplett autonome Systeme zu entwerfen, die sich an ihre Umgebung anpassen können. Hierfür wird in dieser Arbeit ein Dialog-basierter Lernansatz vorgestellt, um eine Wissensbasis über einen längeren Zeitraum hinweg zu pflegen, mit neuem Wissen anzureichern und Fehleinträge zu korrigieren. Der vorgestellte Ansatz wurde für Personenidentifikation, Soziale Netzwerkmodellierung und Lernen von Objekten im Bereich der Mensch-Roboter Interaktion untersucht und evaluiert. Neben robusten und fehlertolerante Lernmethoden werden in dieser Arbeit auch Fehlerkorrekturmechanismen eingesetzt, da fehlerhafte Eintragungen in einem autonom lernenden System insbesondere beim Einsatz über einen längeren Zeitraum in einer realistischen Umgebung unvermeidlich sind. Diese Arbeit liefert daher einen Beitrag für den Entwurf robuster Lernmethoden durch fehlertolerante Dialogstrategien für Lernaufgaben, Optimierung durch Reinforcement Learning, enge Kopplung der Systemkomponenten auf verschiedenen Verarbeitungsebenen des Dialogsystems, Fehlerkorrekturdialoge und Untersuchung des Lernverhaltens über einen längeren Zeitraum hinweg, was die Verbesserung der Qualität der Wissensbasis zur Folge hat.

Im Rahmen dieser Arbeit wurde der interACT Rezeptionsroboter entwickelt, der über einen Zeitraum von elf Monaten im Dauerbetrieb getestet wurde und anhand einer zu lernenden Population von Mitarbeitern zur

Evaluation des Lernansatzes verwendet wurde. Die Evaluationen belegen, dass der vorgeschlagene Dialog-basierte Lernansatz Lernergebnisse mit guter Qualität liefert, dass Personen automatisch modelliert, Objekte und deren Semantik gelernt werden können, dass durch die Kopplung der Erkennungskomponenten und durch multimodale Fusion die Erkennungsergebnisse und die Dialogführung verbessert werden konnten, und dass Korrekturdialoge signifikante Verbesserungen der Qualität der Wissensbasis in einem derartigen autonomen System erzielen.

Hartwig Holzapfel

Abstract

While there have been various improvements and signification advancements in spoken dialog systems in recent years, today, dialog systems still have restricted adaptation abilities as dialog systems are generally being built as static systems with predefined and static knowledge models. In contrast, the next generation of dialog systems will be able to adapt their strategies and knowledge models and thus gaining the ability to adapt to their environment. Such adaptation abilities can be of great service for humanoid robots, if adaptation is autonomously conducted by the system and without manual intervention by a human operator.

This thesis presents a contribution to the far goal of building completely autonomous systems that can adapt to their environment, by introducing a dialog-based learning approach for maintaining a knowledge base in a long-term run including knowledge acquisition and knowledge correction. The presented approach has been applied to and evaluated on person identification, social network models and object learning in the domain of human-robot interaction. Besides robust and error tolerant learning methods, this thesis applies error correction mechanisms, as knowledge base errors caused by an autonomously learning system are inevitable especially when the system is exposed to a real-life environment over a longer period of time. Thus, further contributions of this thesis are design of robust error-tolerant learning methods with error-tolerant dialog strategies for learning tasks, optimization by Reinforcement Learning, tight coupling of multimodal dialog system components, knowledge mending dialogs and analysis of the learning behavior over a longer period of time, which results in better quality of the knowledge base.

Within this work we have developed the interACT robot receptionist, which autonomously acquires a model of a research laboratory population with almost “24/7-availability” and which has been evaluated over a long-term period of eleven months. Results of this on other evaluations show that the dialog-based learning approach produces a knowledge base of good quality, that the learning approach can successfully be applied to automatically model persons and to learn objects and their semantics, that recognition and dialog quality can be improved by tight coupling and multimodal fusion and that knowledge mending can lead to significant improvements of knowledge base quality in such an autonomous system.

Universität Karlsruhe (TH)

Acknowledgements

This thesis would not have been possible without the support and without the contributions of a lot of people to whom I'm really grateful. During six years of constructive research at the Institute for Anthropomatics at the Universität Karlsruhe (TH) which has now become the KIT, and six months at Carnegie Mellon University in Pittsburgh, a lot of people have influenced or inspired my work.

First I would like to thank my advisor Prof. Alex Waibel who has given me the opportunity to do research in an excellent and creative environment, for supporting and believing in my work, and for visionary discussions and feedback. And I would like to thank professor Rolf Carlson, for being part of the thesis committee and for valuable feedback and discussions.

I would like to thank all my colleagues at the Institute for Anthropomatics and the International Center for Advanced Communication Techniques at the Karlsruhe Institute of Technology for many discussions and joint work that we have conducted together in different research projects. I would also like to thank my colleagues at Carnegie Mellon and the Dialogs-on-Dialogs reading group for great and fruitful discussions. I cannot think of a better place to conduct such kind of research. I greatly thank my colleagues not only for cutting edge research but especially for letting me use their high quality software which could be integrated in a real-time processing system and for continuing discussions on integration issues. This includes a wide variety of multimodal technologies with speech recognition, dialog and vision. During my time at the lab I had the pleasure to work with and in some cases share offices with Keni Bernardin, Maria Danninger, Matthias Denecke, Hazim Kemal Ekenel, Mika Fischer, Christian Fügen, Hua Gao, Petra Gieselmann, Teresa Herrmann, Shajith Ikbal, Klaus Joas, Florian Kraft, Ian Lane, Kornel Laskowski, Robert Malkin, John McDonough, Florian Metze, Kai Nickel, Jan Niehues, Matthias Paulik, Ivica Rogina, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Rainer Stiefelhagen, Sebastian Stüker, Dan Valsan, Stephan Vogel, and Matthias Wölfel. The speech recognizer Janus plays a major part in our dialog systems and I have also been lucky to have colleagues who have been building great visual processing components, which I could use in my thesis project. It was in Pittsburgh where I developed the initial idea to build dialogs for a learning robot. Since then, more and more facets have been added to what has finally become the interACT Receptionist Robot. One of the first supporters of this idea was Thomas Schaaf, whom I would like to thank for his support and discussions about detecting and learning new words. I greatly thank Christian Fügen, my first mentor, and Florian Kraft for integration of the speech recognizer on the robot. I greatly thank Kai Nickel for long years of collaboration in the context of humnoid robots and for the visual processing library Arthur, which was a pleasure to integrate. I greatly thank Hazim Ekenel who did great work on visual person identification, which ended up being a core part of person identification and learning in dialog, and Mika Fischer for his contributions to the software. I also thank Rainer Stiefelhagen and Keni Bernardin for valuable discussions on multimodal integration. I greatly thank Pedram Azad for providing visual object recognition software, which could be used for interactive learning of objects, and for long hours jointly working on the demonstrator. I would like to thank Qin Jin for letting me use her voice recognition software. I also want to thank Petra Gieselmann for the years that we spent together building dialogs for the robot and running experiments.

The integration of all these components into the dialog system could not have been done by a single person. Therefor I especially want to thank all students who contributed to parts of this thesis with their student projects (Studienarbeiten), diploma theses (Diplomarbeiten), and HiWi-jobs either by integrating software or by conducting research and running experiments with vast amounts of multimodal data in different settings. In the order of appearance I would like to thank Ulf Krum, Boris Schulz, Patrycja Holzapfel, Carolin Manolikakis, Stefan Ziesemer, Christoph Schaa, Marek Wester, Thomas Prommer, Stephan Könn, Daniel Neubig, Ronny Händel, Philipp Hüthwohl, Felix Putze, Philipp Grosse, Stefan Ultes, Matthias Ernst and Paul Märgner. I want to thank Kornel Laskowski, Philipp Grosse and Patrycja Holzapfel for reading parts of this thesis and for helpful comments.

Parts of this work have been conducted within the Collaborative Research Center (Sonderforschungsbereich) SFB588 Humanoid Robots - Learning and Cooperating Multimodal Robots. I would like to thank all project partners who contributed to this enormous project. Especially I would like to thank Tamim Asfour and his team for sheer endless work on creating a great humanoid robot, for providing valuable time to run experiments with the demonstrator, and for his ability to integrate experimental research software on a robot that actually does practical things. Furthermore I especially want to thank Catherina Burghart, Roger Häussling and Ralf Mikut for joint work, experiments and evaluation of social human-robot interaction.

Last and most importantly I want to say thanks to my wife Patrycja who had to endure all these years of research interests, late-night work and

Hartwig Holzapfel

laptop-holidays, but always actively supported me to successfully complete this project.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Dialog systems today are used for a great variety of applications, such as interactive systems, call centers, navigation systems, and in recent years human-robot interaction has also become the focus of dialog systems research. The term 'dialog system' in this sense refers to a system for human-machine interaction using speech and optionally other modalities for input and ouput. Various architectural diagrams can be found in the literature. The common ground of state-of-the art systems is speech input processing with speech recognition and language understanding, dialog management, natural language generation and text-to-speech output. Figure 1.1 shows a diagram with these five components, following the categorization by Pietquin (2004) including the annotation of processing levels.

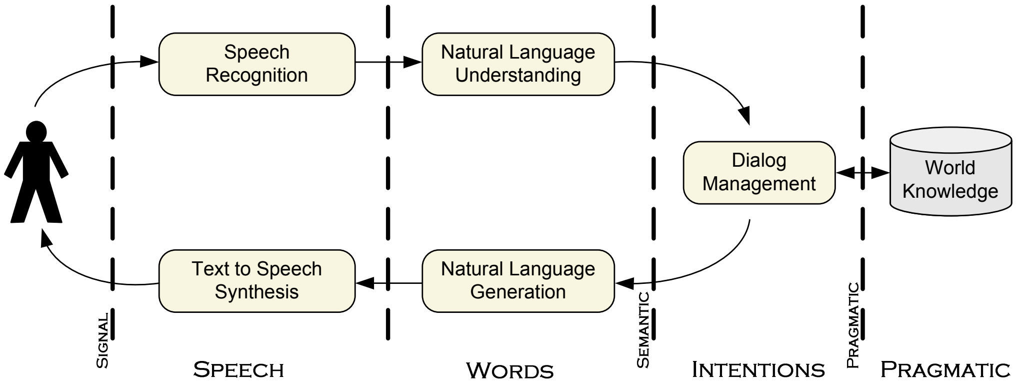


Figure 1.1: Classical speech-based dialog architecture

Each of these components requires specific knowledges sources for operation. The relevant knowledge sources and their knowledge models are shown in figure 1.2. The figure depicts the perceptual side of the multimodal architecture that will be used throughout this thesis and already indicates, how the knowledge sources are shared between different components. The components for multimodal fusion represent general approaches, which also exist

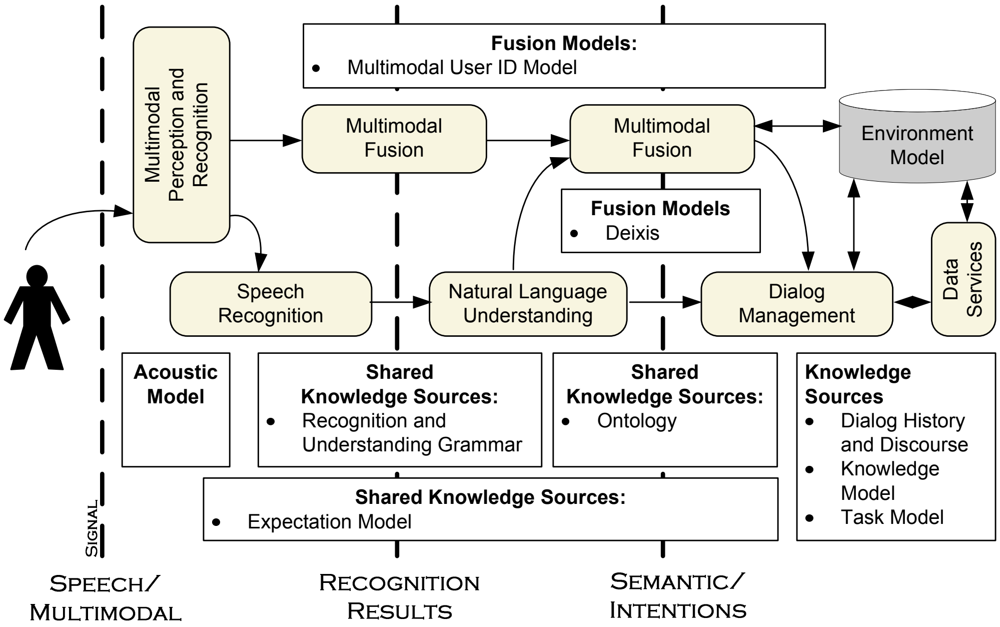


Figure 1.2: Multimodal dialog architecture with shared knowledge sources

in other state-of-the art multimodal systems, where fusion is conducted at a symbolic and/or subsymbolic level (e.g. multimodal person identification) and at the semantic level (e.g. deictic references).

While there have been various improvements and signification advancements in spoken dialog systems in recent years, today, dialog systems are generally being built as static systems with predefined static knowledge models. With static knowledge sources, systems have limited adaptation abilities. And with preprogrammed dialog strategies, also the behavior of the systems is static. However, the next generation of dialog systems will be able to adapt to the environment by adapting their knowledge sources and interaction strategies. Such adaptation abilities can be of great service, if the adaptation can happen autonomously by the system itself and without a human operator who needs to program the system's database manually.

The focus of this work is a spoken dialog-based interactive learning approach for extending and creating multimodal knowledge bases. This approach enables a system to extend its knowledge (i.e. *Environment Model*, *Grammar*, and *Ontology*) through explicit spoken dialog-based interaction, and presents tight coupling (i.e. *Shared Knowledge Sources* and *Multimodal Fusion*) and machine learning techniques for dialog strategies (i.e. the *Task*

Model). In contrast to classical learning methods, this approach does not require a human operator to annotate data manually such as in supervised learning. Rather, knowledge is extracted from information acquired from natural interactive spoken dialog (i.e. *Dialog Management*) in human-robot interaction scenarios and from background processing of data retrieved from the World Wide Web (i.e. *Data Services*).

1.1 Problem Definition and Scenarios

Recent work addresses the development of humanoid robots that in the future will support everyday life of humans. There, humanoid robots enter an open environment, where they encounter persons and objects that have not been known during design time of the system. Therefore, the robot domain offers a good test environment for this thesis.

A goal posed by this thesis is to study an integrated learning approach and to examine a complete dialog system for learning. To pursue this goal, a dialog system for human-robot interaction has been developed and extended by learning capabilities, which are tested in different scenarios. These learning capabilities address interactive, autonomous updates of the system's knowledge sources. Each of these knowledge sources as depicted in figure 1.2 models important information necessary for the functionality of the dialog system. Learning processes are necessary to extend, update or delete information from the knowledge base. The knowledge base must be updated to adapt to new observations, or it must be extended, e.g. when the robot acquires information about a previously unkown object. The learning capabilities studied in this thesis are person identification, social user models, object representation, and reinforcement learning for acquisition of error tolerant dialog strategies. In terms of the above knowledge sources, this might include adding new words to the *Recognition and Understanding Grammar*, to the speech recognizer's vocabulary (graphemic representation) and to the speech recognizer's pronunciation model (phonetic representation), adding new concepts and relations to the *Ontology*, adding new entities to the *Environment Model*, and adding new samples to the visual classification models. Other learning tasks require similar updates of knowledge sources.

For testing these capabilities, the following three scenarios are described in this thesis. All scenarios are depicted by use case diagrams with pictures from the settings. Use cases, as shown in the picture, are well applicable to dialog interaction, and can be interpreted as dialog tasks. The links between the use cases are shown as <*includes*> and as <*extends*> relations, arrows which are unannotated for better readability represent <*includes*> relations.

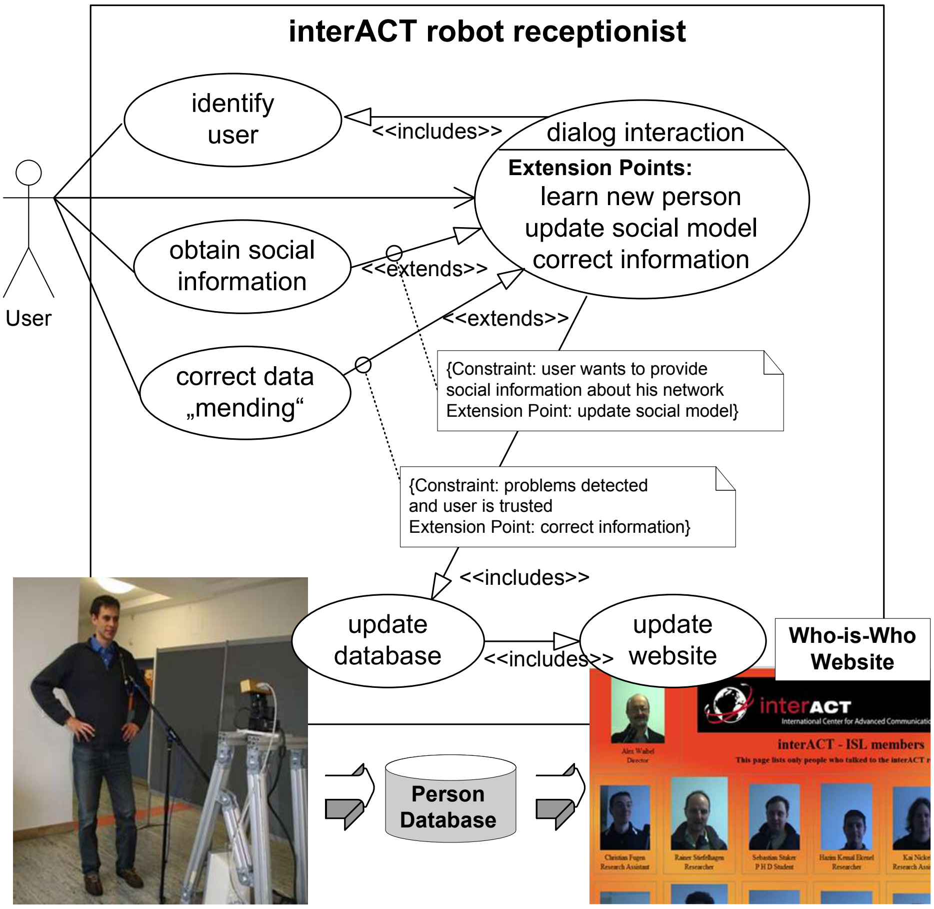


Figure 1.3: Use case and pictures from interaction with the interACT receptionist, showing person identification and social network modeling

- the interACT robot receptionist, which models persons and their social relations in the interACT laboratory. Figure 1.3 shows the main use cases of the interACT robot receptionist with user interaction. The use cases include person identification, new person learning, social network modeling and knowledge mending.

- a robot parcel receptionist, which acts as a receptionist in a parcel delivery task. Figure 1.4 shows the main stages of the dialog interaction with the parcel robot receptionist as use cases including person identification.

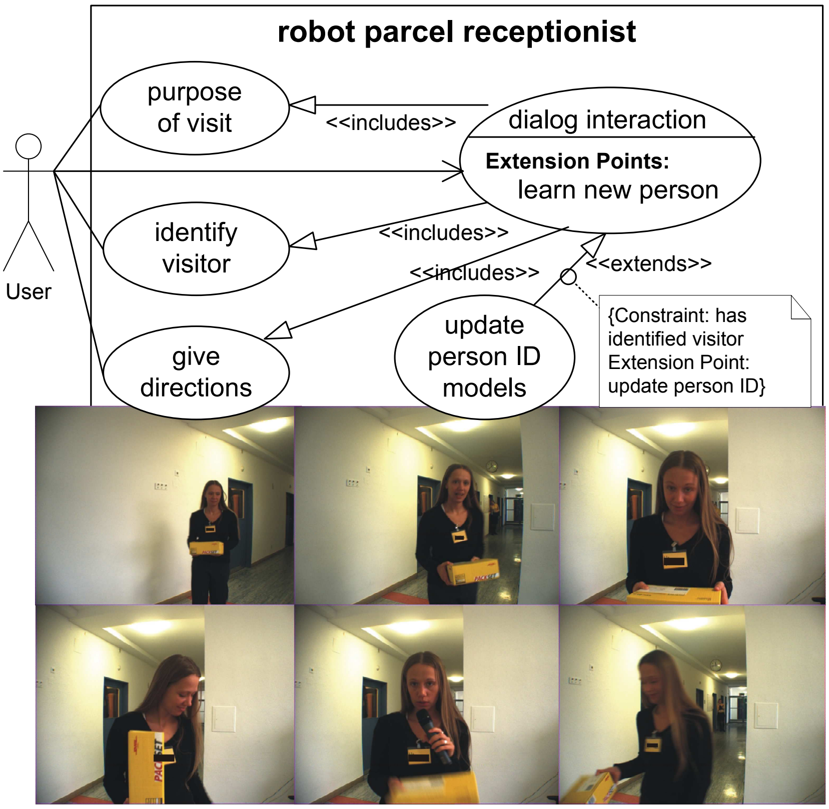


Figure 1.4: Use case and pictures from interaction with the parcel receptionist

- interaction with the humanoid robot Armar III[1] with fetch and bring services. Dialogs for object learning are tested in this scenario. Figure 1.5 shows the two main use cases that involve user interaction plus a set of subtasks that need to be conducted in dialog to learn an object.

All three scenarios have in common that they are tested in an open environment, where new persons, object, words, etc. occur, which the system needs to learn to fulfill its task. The robot parcel receptionist and the interACT robot receptionist are similar as both are robot receptionists, and in their functionality as a receptionist, both get to know new persons. They are distinguished here, as the robot parcel receptionist is an earlier version of the interACT receptionist and was mainly used for analysis of interaction and social studies. The interACT robot receptionist finally is the largest system among these in terms of the dialog-based learning model and covers

[1] Armar III is being developed within SFB588:
http://www.sfb588.uni-karlsruhe.de

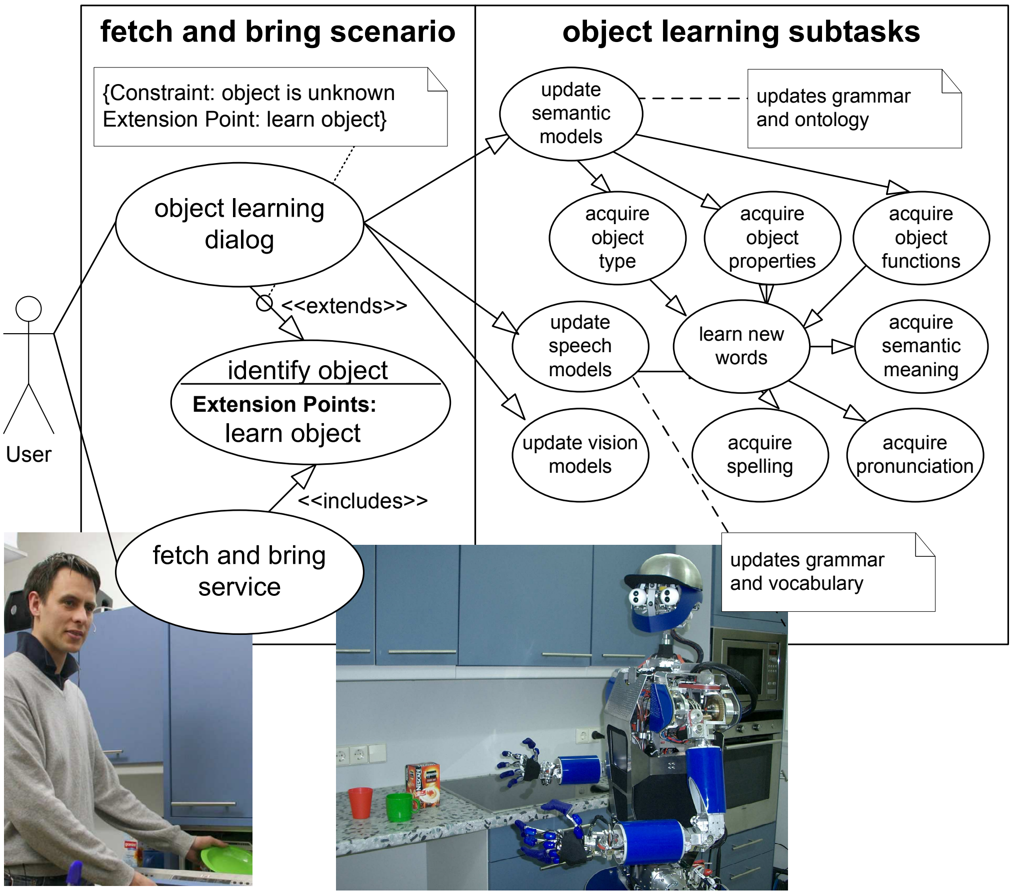


Figure 1.5: Use case and pictures from interaction with Armar III, showing fetch and bring services and object learning

several experiments including user identification, reinforcement learning of strategies, social network modeling and knowledge maintenance over time.

To measure if the proposed approaches are effective, we employ evaluation criteria which are applied to each capability separately (e.g. dialog success rates, dialog efficiency and learning success rate), and evaluation criteria which evaluate the success of the interACT receptionist in creating and maintaining a social database in a long-term study (e.g. knowledge base quality and subjective user feedback).

So far, the terms 'knowledge base' and 'learning' have already been used without further specification. As both terms can have quite different meanings, the following definitions specify how these terms are used in this thesis. In addition, the term 'dialog-based learning' is introduced and the scope of 'learning' is further specified.

We use the term ***knowledge base*** in this thesis for a model, which represents certain aspects of the environment. It stores descriptive information about acquaintances and objects, i.e. representations of real-world entities, ontological concepts as semantic categories, relations between the entities, and multimodal recognition models.

We use the term ***learning process*** for a process which updates the knowledge base by adding, modifying or removing information, concepts, relations and recognition models, with the purpose of applying the updated models for recognition of entities in the environment (e.g. persons), or to reason about the stored information.

We use the term ***dialog-based learning*** for a learning process which is realized by a dialog system.

We use the term ***learning*** throughout this thesis with different meanings of what and how something is learned, including procedural and declarative knowledge. The term 'learning' is used for machine learning techniques which include training of dialog strategies by reinforcement learning. The system can 'learn' a new word during the dialog, which requires at phonetic and graphemic representations and semantic categorization (without manual annotation by a human), or can 'learn' the semantic categorization of an object, which includes modifications of the ontology. The term 'learning' is also used for automatic creation/training of classification models, e.g. for visual person identification. Finally, we use the term 'learning' to describe the overall system behavior while maintaining a knowledge base including adding, updating and removing entities from the knowledge base, including their semantic representation and classification models.

1.2 Thesis Statement

Knowledge acquisition by a humanoid robot can be conducted robustly using multimodal spoken dialog. The robot can adapt its knowledge base to new observations in the real world and over a longer period of time. The system can correct errors in the knowledge base by automatic detection of contradictory information and conduct dialogs to clarify false information. The impact of the suggested learning technique is measured over a longitudinal lifetime of the dialog system.

1.3 Contributions

This thesis introduces a dialog-based learning approach with unsupervised learning mechanisms for acquiring information and maintaining a knowledge base. In contrast to supervised learning, such a learning mechanism enables robots to learn and extend their knowledge autonomously without manual intervention by a human supervisor. Contributions of this work include

- A dialog-based learning approach for knowledge acquisition using multimodal data is introduced. The presented approach advances the state of the art with the proposed concept of a fully autonomous system which includes interactive knowledge acquisition and background knowledge acquisition, which is robust against recognition errors and facilitates maintenance of a knowledge base over a longer period of time. The approach includes a modular dialog approach to update a knowledge base of specific knowledge entities, with adding, updating and deletion of entities.

- The concept of dialog-based learning has been applied to and evaluated in several scenarios, such as learning of personal information, name learning, and object learning, in a more extensive way than other state of the art systems that facilitate interactive learning. Within these scenarios, the suggested knowledge model has been tested for learning of complex entities, object semantics, multimodal data, attributes, and social networks of persons.

- New approaches are presented for handling development of the knowledge base over time. The system's learning behavior is evaluated over time and long-time effects on quality are analyzed. Since such a learning system produces errors after some time, a knowledge mending approach is introduced that can effectively detect and solve problems of the knowledge base using interactive and non-interactive problem detection and resolution methods.

- A fully integrated system for knowledge acquisition in dialog is introduced. In contrast to existing work, the presented system can process a multitude of knowledge sources and is evaluated with expert users and naive users in a realistic (real-world) setting. This is facilitated by tight coupling and multimodal integration, which leads to improvements in recognition performance (e.g. speech recognition by contextual control and dynamic vocabulary) and dialog success. A new multimodal fusion approach is presented using confidence-based multimodal fusion and Bayes Net theory for integrating multimodal data and dialog features for open set person identification.[2]

- The high requirements for error tolerance and robustness posed to the dialog strategies by the scenarios are addressed by a combination of handcrafted dialog design and reinforcement learning for learning optimized dialog strategies. New multimodal user simulation methods are introduced for training multimodal dialog strategies. For combining different techniques as handcrafted dialog strategies and reinforcement learning in a runtime system, a modular dialog approach is utilized. Similar to an agent-based approach, this allows separating concerns, implementing modules independently and combining them within one runtime system.

1.4 Thesis Overview

This thesis is organized as follows. Part I presents a fully integrated framework for human-robot interaction and techniques for robust multimodal dialog processing. Part II extends this framework with dialog-based learning methods for knowledge acquisition.

Part I therefore is focused on error tolerant dialog strategies and integration aspects. Scenarios for experiments and evaluation in part I reflect common tasks of a humanoid robot in a household environment and combine ‘standard’ interaction and learning tasks. Effectiveness of the suggested methods in part I are measured by success rates of dialog tasks, subjective user feedback, and improvements of the suggested methods over a given baseline.

Part II focuses on analyzing different aspects of dialog-based learning. Scenarios for experiments and evaluation in part II are focused on tasks involving knowledge acquisition. So far, no standard for the evaluation of dialog-based learning exists and only parts of the techniques can be evaluated against a real baseline. To still be able to assess the success of the proposed methods, gold standards are introduced to measure the success of different techniques, measure relative improvements when combining different techniques, and to evaluate an overall system. Such an overall system evaluation was conducted for the interACT receptionist in a long-term study in a realistic environment. With the goal to set up evaluation metrics which are easy to understand by humans, i.e. one can personally estimate how well

[2] The dialog system integrates several components developed within SFB588, which are speech recognition, object recognition, person tracking, face identification, voice identification and pointing gesture recognition.

the system performs, an evaluation scenario was created for the interACT receptionist. Its performance is measured by how well it can model a group of persons working in the interACT lab and present the result on a "Who-is-Who" page. Within this scenario, the learning curves of the system are plotted over the already mentioned gold standards, subjective user feedback is assessed and different metrics have been applied to evaluate knowledge base quality.

Foundations and Related Work:

Chapter 2 includes foundations of the dialog system and related work for dialog processing, multimodal human-robot interaction and dialog-based learning.

Part I:

Chapter 3 presents the dialog manager TAPAS, which has been developed as a dialog system for a humanoid robot, with dialog architecture and tight coupling.
Chapter 4 presents a multimodal user model with multimodal fusion techniques on different levels (as sketched in figure 1.2) of the dialog system for multimodal person identification.
Chapter 5 presents machine learning techniques for training and optimization of dialog strategies for person identification with reinforcement learning and multimodal user simulation techniques.

Part II:

Chapter 6 presents a dialog-based learning approach with knowledge model and realization by dialog modules.
Chapter 7 presents application of the learning approach to multimodal object learning including semantic categories for unknown objects from the kitchen and household environment and new words in speech recognition.
Chapter 8 presents the interACT robot receptionist which is used for long-term evaluation of the dialog-based learning approach and social user models presented in the following chapters.
Chapter 9 presents acquisition of social user models, which includes offline information retrieval, information extraction and information acquisition in dialog.
Chapter 10 presents evaluation of the dialog-based learning approach for acquisition of personal information in a longitudinal study and evaluation of knowledge mending.
Chapter 11 concludes this thesis.

Hartwig Holzapfel

Chapter 2

Foundations and Related Work

This chapter first presents an overview of dialog management approaches in general and their application to human-robot interaction. Secondly this chapter describes foundations and work related to dialog-based learning.

In the first part of this chapter, a brief introduction is given into the field of dialog management for human-robot interaction. The presented methods for dialog management provide a foundation for the dialog approaches presented in this thesis, and address human-robot interaction, robustness and evaluation.

Section 2.4 then describes related work for dialog-based learning for humanoid robots and addresses the problems of learning of new words, learning of semantic categories, learning of multimodal models and cleaning errorful knowledge bases. These problems are described in the context of object learning and acquiring models of persons.

2.1 State of the Art of Dialog Systems for Human-Robot Interaction

2.1.1 Dialog Management Approaches

The term dialog manager is generally used for a specific component of a dialog system. Its basic functionality is to conduct a dialog strategy and execute dialog actions (often referred to as moves) in reaction to input events. Depending on the specific system, a few other components are integrated into the dialog manager, such as discourse management, abstract state modeling, belief update, and context management.

A categorization of dialog management approaches are given by McTear McTear (2004, 2002), who categorizes these approaches into three basic classes: the finite state based, the frame-based and the agent-based approach. Most of today's dialog systems can hardly be classified as one of these approaches. The agenda-based dialog manager Xu and Rudnicky (2000b) directly models the human's task, as an expectation of the dialog flow. It has been extended in Bohus and Rudnicky (2003, 2008), by separating task and discourse behavior. A frequently applied pattern is that of a state-based dialog manager which maintains a belief state of different observations and information about the dialog flow. This category includes the information-state update (ISU) approach Lemon et al. (2001) and abstract dialog state (ADS) based systems Denecke (2002). ADS and ISU approaches have the same concept of introducing different variables that characterize the current dialog state. Especially state-based dialog managers, e.g. the ISU approach, have shown suitability for combination with reinforcement learning algorithms, like conducted by Scheffler and Young (2002). Recently, there has been growing interest in optimizing dialog strategies with a partially observable Markov decision process (POMDP) Roy et al. (2000). The term partially observable refers to introducing "hidden variables" into the dialog manager, e.g. Young et al. (2007). Some of the variables cannot be observed directly, e.g. the real name of a person, for which the dialog manager only observes the speech recognition result.

Depending on the task complexity of the dialog system, a decomposition of the complete system into smaller parts can be of advantage. A decomposition of a dialog architecture is proposed by Turunen and Hakulinen (2003) or Nakano et al. (2006), who use agents for distributed interaction tasks, and Bohus and Rudnicky (2003), with a hierarchical task structure and handlers for specific states.

2.1.2 Dialog Systems for Human-Robot Interaction

Most traditional dialog approaches consider speech-only interactions, and have focused on phone based interaction, such as flight and train timetable information systems McTear (2002); Allen et al. (2000); Stallard (2000), call-routing systems Gorin et al. (2002), weather information systems, Zue et al. (2000), etc. The next generation of dialog systems has to cope with direct human-machine interaction from face to face, which exist for example in human robot dialogs or in smart room environments. This results in new challenges resulting from the physical environment shared by the user and the system, the situated and context-dependent communication, the changing environment, the multimodal interaction, etc.

To address these challenges, different approaches are taken, which often require the integration of interdisciplinary approaches. The applied dialog technologies for human-robot dialogs range from finite state systems to more complex models. Many robots use command-based speech input or simple dialog control. Some dialog systems for robots are based on finite-state automata e.g. the robots HERMES and BIRON Bischoff and Graefe (2002);

Toptsis et al. (2004). Finite-state automata are models, which are easy to design, and are sufficient in many scenarios. Later, the dialog model on BIRON has been extended to a more complex interaction management system with multimodal dialog capabilities Li (2007). Also Aoyama and Shimomura (2005) use a simple finite state dialog model. In the integrated system, ”Naturalness Support Behaviors” are studied, which include for example nodding, filler insertion, face tracking, and reactions to environmental stimuli during interaction. In complex environments, often more advanced strategy approaches are necessary, to deal with speech recognition errors, process multimodal information, and handle the manifold contextual states. Such advanced approaches are implemented e.g. for the robot Pearl Montemerlo et al. (2002), which uses a probabilistic approach to cope with recognition errors. The dialog system WITAS for unmanned vehicle control Lemon et al. (2001), adopts the information state update (ISU) approach. Also Bos et al. (2003) adopts the approach to dialog move engines with an information state model, developed within TRINDI Traum et al. (1999), in an agent-based architecture. An agent’s information state is updated on the basis of observed dialog moves, leading to the selection of a new dialog move to be performed by the agent.

2.2 Robustness and Learning of Dialog Strategies

In recent years, there have been two main approaches to create dialog strategies, either by manual writing of dialog strategies, or by applying learning mechanisms for dialog strategy training.

One shortcoming of handcrafting dialog strategies is that it is a time-consuming and non-trivial task, especially with increasing complexity of the dialog. Additional problems are robustness on unseen data. One promising approach to avoid these problems, is to use collected dialog data for automatic training of dialog strategies. For this task especially reinforcement learning has become popular. So far, reinforcement learning has successfully been applied in a couple of dialog scenarios Singh et al. (1999); Levin et al. (1998a, 2000); Walker and Shannon (2000).

One problem of applying this technique to dialog systems is the large number of data (i.e. dialogs) required for training of the system, so that training strategies on real data has usually been conducted with a limited state space and/or action space. More recently, there have also been approaches to training dialog strategies with a user simulation, which allows to generate a vast number of dialogs, which are necessary for training more complex dialog strategies Scheffler and Young (2002); Williams and Young

(2003); Schatzmann et al. (2006); Pietquin and Renals (2002); Schatzmann et al. (2007). So far, these approaches do not cover multimodal systems or approaches for dialog-based learning. Both aspects will be addressed in chapter 5.

Another aspect besides reducing manual labor by dialog strategy optimization, is to achieve robustness against errors that occur during the dialog. In addition to optimizing the dialog strategy itself to achieve robustness against errors, e.g. Lemon and Liu (2007), other approaches have been applied successfully, too. Approaches for robust dialog strategies include explicit error handling Gieselmann (2007); Skantze (2007a). Further approaches in the field derive robust strategies by using confidence measures from speech recognition Bohus (2007).

2.3 Evaluation

2.3.1 Evaluation of Human-Robot-Interaction

In the future, more and more robots can be found in environments of typical human everyday life, i.e. in hospitals, hotels, museums, schools, and households. The usage of robotic devices in the human world requires appropriate design of the robot's interface to the environment and of the robot's cognitive skills, thus enabling intuitive interaction between robot and people.

The goal of evaluation is to quantify different aspects of human-robot interaction (HRI) to improve robot design, increase acceptability and adapt robots more to the need of humans. As HRI is a relatively young field, there are still many areas that still need to be explored and many aspects of human-robot interaction still lack measures that can be quantified. In the following we want to give an overview over evaluation metrics and evaluation procedures which are current state of the art including recent achievements.

One common means to assess robot success are benchmarks. A great variety of benchmarks do exist: some recent examples with a great deal of public attention are robot soccer competitions in different leagues Bredenfeld et al. (2006), test course for rescue robots Jacoff et al. (2002), the DARPA Grand Challenge Thrun et al. (2006) and the DARPA Urban Grand Challenge DARPA (2007) for autonomous driving of cars. Human-robot interaction is contemplated by the RoboCup@Home league founded in 2007 Nardi and et al. (2007). The performance measurement is based on a score derived from competition rules and the evaluation by a jury. However, a transfer of such competition concepts and evaluation metrics to domains in the human everyday world can cover only a part of the necessary evaluation procedures.

Besides competitions, various metrics are used by international researchers like the preferred direction of approaching in a living room scenario Woods et al. (2006) or the distance a person feels most comfortable with when interacting with a robot Walters et al. (2007). Others, as proposed by Steinfeld et al. (2006) include success rates and number of operator interventions in tele-operated scenarios. Additionally, metrics for performance, world complexity and information quantification were established for autonomous mobile robots navigating in a corridor clotted by random obstacles Lampe and Chatila (2006). In the first category instantaneous velocity, traveled distance, mission duration, mission success rate and power usage were measured, whereas global complexity and the vicinity of the robot are taken into account in the second category. The last metric used is the conditional entropy measuring the information contained in the internal robot map compared to the world map.

As soon as communication forms an integral part of human-robot interaction additional objective metrics like WER: word error rate (the standard metric for automatic speech recognition - ASR), CER: concept error rate (error rate to measure understanding, based on recognized concepts) and TER: turn error rate (based on number of turns that cannot be transformed to the correct semantics) can be applied. Current research on spoken dialog system either uses objective metrics, subjective metrics, or both. The main advantages of subjective metrics over objective metrics are that the user's subjective perception of the system can be included in the evaluation. Most measurements are based on questionnaires with rating scales such as Likert-Scales. Approaches exist to build a unified framework for the evaluation of dialog systems and create comparable scores with the PARADISE framework Walker et al. (1997) for spoken dialog systems.

In contrast to metrics based on measurable characteristics and typically used in engineering, Kahn et al. (2006) suggest metrics for human-robot interaction devised from an psychologist's point of view which include autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. These contenders are attributed to a robot by the person interacting with it.

Coding of behaviors and deriving rules for interaction are another form of metrics adopted by some research groups. The problem when applying this procedure is the objective coding of behavior which actually is a subjective interpretation of an interaction scene as seen by an observer. In order to gain valid data the same interaction scenario should be coded by several independent observers of the experimental staff. So-called micro behaviors were used by Dautenhahn and Werry (2002) based on criteria like eye gaze, eye contact, operation and handling, movements, speech, attention, and repetitions. The length of eye gaze was used as a correlation to the subject's level of interest in a robot or toy truck. Behavior-level codes describing the adjustment of children to the setting of a communicative robot interacting with children in a primary school were used by Nabe et al. (2006); Kanda et al. (2004) to analyze the role of their robot.

So far many ideas, methodologies, metrics, and measurement criteria do exist in order to assess human-robot interaction, but most of the applied metrics consider mainly technical characteristics of the robot. Even success rates of interactions do not really picture the manifold ways of human behavior and the reasons for a failure of the interaction. The problem is that human behavior cannot be measured using simple scales. The assessment of interactions between naive persons and robots actually requires a framework of different metrics: a combination of objective metrics which can easily be measured and quantifiable subjective metrics characterizing human behavior. Here, undue influence of naive subjects as well as biased opinions of observers has to be taken into account by creating a set-up for sound experimentation and analysis.

2.3.2 Evaluation of Dialog

The list of different objective metrics which have been applied to dialog systems is relatively short. Most systems use some kind of recognition accuracy, dialog length, and dialog success. Recognition accuracy can be represented as Word-Error-Rate (WER) which is the simplest metric. It has the advantage that it is usually used for evaluation and comparison of speech recognition systems and can easily be computed when the transcription of speech input is given. However, WER is not necessarily the best metric to represent recognition accuracy. For example, it does not distinguish between content words and non-content words. Sentence-Error-Rate (SER) checks the correctness of complete sentences. Some evaluations measure correctly recognized semantic concepts, for example (semantic) Concept-Error-Rate (CER) is reported in Chotimongkol and Rudnicky (2001); Glass et al. (2000); Holzapfel and Waibel (2006). Differences exist whether CER is defined on fully correct semantic input or regarding the details used to measure correctness. CER is probably the metric, which is best suited to represent input understanding in a dialog system, because it is measured by the correctness of the input, which is actually used by the dialog manager. However, it requires semantic transcription of input, and is not as simple as word-error rate, since it depends on the type of semantic structure and details of semantic transcription.

Dialog length is usually measured in number of turns to achieve a certain goal. In task-oriented systems the number of turns is measured to achieve a predefined task. Some other metrics have been used, such as the total amount of time in seconds, or the number of syllables spoken, e.g. by Skantze (2007b). Glass et al. (2000) uses concept efficiency (CE) which quantifies the average number of turns necessary for each concept to be understood by the system, and query density (QD) which measures the mean number of new concepts introduced per user query. Both metrics relate to the length of the dialog with respect to how effectively information can be communicated without the necessity of a task definition.

A widely used metric is dialog success. However, the definition of dialog success varies among different systems. Most approaches use achievement rates of dialog goals, e.g. Schatzmann et al. (2005).

Besides the most commonly used metrics, there is a larger number of different features for detailed analysis of the interactions, e.g. Fraser (1997); Polifroni et al. (1992); Price et al. (1992); Simpson and Fraser (1993); Danieli and Gerbino (1995); Walker et al. (1998), most summarized in Möller (2005). These interactions parameters can be used for detailed analysis of the system, aspects of the interaction, and different quality aspects Möller et al. (2007).

As a framework for dialog system evaluation, PARADISE introduced by Walker et al. (1997, 2000) is best known. It offers a prediction model for quality judgments based on a regression model with interaction parameters as input. It serves two purposes, one part is the framework for prediction of quality judgments, and the second part is a set of questions and metrics for evaluation. The framework has been applied to a number of different systems, for example Hajdinjak and Mihelič (2006). Since PARADISE has initially been designed for speech-only interactions, a modified version, PROMISE, has been suggested by Beringer et al. (2002) to address aspects of multimodal systems.

Such frameworks apply both, objective and subjective metrics. Subjective evaluation is usually conducted with the help of questionnaires, which allow quantitative measurements based on Likert-Scales. A Likert-scale is a one-dimensional scale with a discrete set of response possibilities, usually a 5-point, or sometimes a 7-point scale to rate between Disagree and Agree. Questions are then formulated as statements. Some approaches use different opposites than agreement or disagreement, such as *'good'* vs. *'bad'*, *'very much'* vs. *'not at all'*. Questions are then formulated as real questions, such as “How is your overall impression of the interaction?”.

Recent work addresses aspects of how to successfully design systems, by introducing checklists or design principles, e.g. Niels Ole Bernsen (1996); Bernsen and Dybkjaer (1997); Dybkjaer and Bernsen (2000); Suhm (2003). Such design principles are important in the design of spoken dialog systems, and many design errors can be prevented by following such design principles.

Similar as there are guidelines for designing a system, there also exist guidelines for evaluating systems. A good overview over subjective evaluation according to de-facto standards is given by Möller et al. (2007). Moeller analyzes different de-facto standards with respect to which aspects of quality can be measured by these standards, and how reliable the measures are. He refers to de-facto "standards" as guidelines, which are formulated in terms of recommendations to the evaluator. A limited number of practical guidelines are defined, by such a de-facto standard, on how to carry out assessment and evaluation experiments. Often these guidelines are universally valid, some are restricted to a specific domain and need to be adapted for other systems.

Guidelines, which are relevant for spoken dialog system, are first general recommendations on assessment and evaluation methods, e.g. Fraser (1997) in the "EAGLES Handbook". Guidelines for Wizard-of-Oz experiments, which are important parts in the design of dialog systems, are presented e.g. in Fraser and Gilbert (1991); Dahlbäck et al. (1993). As introduced above, subjective user feedback is important to assess system quality. Guidelines for collecting quality judgments are presented with the introduction of the SASSI questionnaire, developed by Hone and Graham (2000, 2001), which has been adapted for usage in this thesis.

2.4 Dialog-Based Learning for Humanoid Robots

Dialog systems usually have static knowledge about the environment, and static interaction strategies. This can be desirable, and robust dialog strategies can be trained, e.g. by reinforcement learning, to obtain robust and predictable behavior in a predefined environment. However, some situations require adaptation of the system during runtime and require the system to maintain its knowledge over a longer period of time. An autonomous system has to achieve this without supervision by a system expert.

The field of work which addresses interactive learning in human-robot interaction can be categorized into the three categories

- acquiring personal information
- object learning
- programming by demonstration for learning of actions

In the following, we first introduce architectures for learning, and then address different learning approaches. Programming by demonstration can be considered as the most advanced area in the above list. Despite the fact that programming by demonstration is a kind of interactive learning, most approaches use no or only very limited dialog capabilities. Therefore, we concentrate on acquiring personal information and object learning.

When looking at the field of related work it can be seen that most approaches go into different directions and have different main focuses. However, when looking more closely at the learning approaches it becomes clear why this is the case. Learning in the context of human-robot interaction includes a multi-disciplinary field. Work conducted in this field covers different areas, such as robot control, visual processing, speech recognition, dialog management, and semantics. When reading the work presented in the following, this should be kept in mind. These areas also help to classify the presented approaches. For example, object learning approaches can be distinguished by whether the main focus is on visual processing, semantic category acquisition, dialog management; as well as a combination of these areas. For example, by combining visual processing and speech recognition, labels can be associated to new objects, while in another scenario, deep semantics are acquired for a new object.

2.4.1 Object Learning

The task of learning and memorizing objects is a complex task which includes different learning approaches. Some learning approaches are similar to functionality required by other learning tasks, e.g. acquisition of personal information. Learning of objects in the context of human-robot interaction has recently seen increasing attention.

Work in the area of object learning for humanoid robots can be classified by the subtasks of learning that are addressed. These categories are (i) new words learning in speech recognition (ii) learning visual features of real world objects (iii) learning of semantics. These categories define a useful classification scheme since they represent the main categories that form a typical object learning task for a humanoid robot.

Table 2.1 gives an overview over existing approaches and a classification according to the categorization scheme, including relevant work which addresses single categories of the classification scheme without explicitly addressing object learning. The first entries directly address object learning, and the following entries, starting with Schaaf (2004) and Gavalda (2000) address aspects relevant to object learning. For example, Schaaf (2004) addresses unknown word detection and new words learning and Gavalda (2000) addresses learning of semantic grammars.

As can be seen from the classification table, only few approaches address multiple categories, most addresses only a single category of object learning.

Authors/Publications	Speech	Vision	Semantic
Roy (1999, 2003)	X	X	X
Dusan and Flanagan (2003, 2002)	X	X	X
Carbonell (1979)	-	-	X
Azad et al. (2007)	-	X	-
Lömker (2004)	-	X	-
Kirstein et al. (2005); Wersing et al. (2006)	-	X	-
Becher et al. (2006)	-	X	X
Schaaf (2004)	X	-	-
Gavalda (2000)	-	-	X
Kaiser (2006)	X	-	-
Choueiter et al. (2007)	X	-	-
Scharenborg and Seneff (2005)	X	-	-

Table 2.1: Work addressing object learning, and coverage of tasks

During the last few years, several approaches have been presented for learning of unknown objects in the field of visual processing, e.g. Lömker (2004); Kirstein et al. (2005); Wersing et al. (2006); Kirstein et al. (2008); Azad et al. (2007). These works present approaches for visual feature extraction, learning and recognition of objects, and allow the robot to recognize an object again by vision after it has been learned. The presented approaches should not be seen as fully integrated learning systems of a robot, i.e. they either do not work as autonomous systems without a supervisor, who provides labels via keyboard, or the system is able to process speech input but is restricted to a fixed set of predefined categories. Such approaches rather fit into more complex systems, as presented in this thesis and in following descriptions. In this thesis we present experiments on object learning, which have been conducted with the integration of the system presented in Azad et al. (2007) visual object recognition and learning of visual features.

An integrated system for the robot BIRON, which is able to acquire new information, is published in Wrede et al. (2006), presenting work which has been conducted within the COGNIRON[1] project. A later version of the system is presented in Hanheide and Sagerer (2008). The system integrates different behaviors and is able to switch between standard interaction and learning mode. They also follow the approach of "learning by interacting", by abstracting from particular machine learning techniques and excepting learning as a general and systemic challenge. The system's ability to learn covers interactive learning for objects and locations in an apartment within

[1] http://www.cogniron.org/

the so-called “home-tour” scenario. Objects and the environment are represented in a so-called active memory, which integrates percepts from different modalities. The active memory can be extended during runtime, and integrates anchors obtained from visual processing with Learning of new objects and locations is conducted with dialogs, during which labels are provided by a human Spexard et al. (2007). Learning is restricted to a finite set of semantic labels and a small speech recognition vocabulary, but it is combined with flexible representation of visual features. To show that the system is accepted by humans as a social actor, experiments have been conducted with naive users in an apartment that has been permanently rented. Results show that a given task of taking the robot around in the apartment (requesting the follow behavior), teaching two rooms (using the location learning with autonomous exploration), and showing two objects, could be completed by 22 out of 24 participants within 15 minutes. However, only about 33% of all tries to teach a room and only about 42% to initiate a follow behavior were successful at the first try. These numbers show that especially for naive users, robust and flexible dialog strategies, including error treatment, exception handling, and informative feedback, are crucial.

The “home-tour” scenario has also been adopted by Kruijff et al. (2007), and in a similar version by the RoboCup@Home competition[2] Nardi and et al. (2007). Kruijff et al. (2007) describe their approach as human-augmented mapping, which allows a robot to augment its map, which has been acquired autonomously with a 3D laser scanner (SLAM), with ontological information obtained through dialog. The approach extends previous work by introducing a structured, non-flat ontology, which models locations and objects. Topp et al. (2006) describe an experimental setup with a Wizard-of-Oz study of the human-augmented mapping task. Their findings are that individual differences in teaching an environment exist, and that the observed diversity in strategies was quite large, e.g., when introducing the kitchen. Different labels were given by different persons, such as “this is the coffee machine” versus “this is the kitchen” – in one case describing important objects that mark a location, in the other case describing the location itself. One of our conclusions from these experiments is that labels obtained during such dialog interactions need to be treated as indirect labels.

Wu and Nevatia (2007) presents an incremental object learning approach using general shape based part detectors to reduce manual labeling. The system uses a two-stage process with an oracle for unsupervised learning, which is based on a combination of shape based part detectors learned by off-line boosting. The oracle provides the basis for online learning. In this applica-

[2] www.robocupathome.org

tion, the oracle is trained to have high precision to achieve good performance, while detection rate can be low.

A different approach has been taken by Roy (1999); Roy and Pentland (2002); Roy (2005), who combines speech processing and visual processing in a visual grounding mechanism, as a computational model for early lexical learning in infants, with minimal prior knowledge. The system was able to observe words which describe shape and color of objects and grounding of these words in perception. Perceptual input was provided by speech recognition and visual perception from a video camera. In the learning phase, simultaneous occurrences of those words and observations in video were used to train the models, which allows to learn corresponding words and visual observations even in unsegmented video. In contrast to other work, semantic categories are not predefined, but are learned implicitly. Semantic structure is shallow and syntax and grammar were not studied in this approach. The system does not represent words as abstract symbols. Instead, words are represented in terms of audio-visual associations. This allows the machine to represent and use relations between words and their physical referents. An important feature of the word learning system is that it is trained solely from untranscribed microphone and camera input.

Later work in Roy (2003) presents experiments with the approach in interactive robotics, and a small robot that can learn objects from a teacher who describes objects in front of the robot. Also here, the robot starts with no knowledge, and lexicon and language understanding are learned from scratch from observations.

Also the approach of Steels and Kaplan (2001) analyzes very early stage language acquisition, with social learning and grounded communication, for the robot pet Aibo. Further work on grounded natural language communication includes event descriptions by Steels and Baillie (2003), in a system for open-ended communication by autonomous robots about event descriptions anchored in reality through the robot's sensori-motor apparatus.

Shibata et al. (2007) presents an approach to also find corresponding objects and reference in speech. Here, object models are learned automatically from TV cooking shows, given keywords for speech recognition. Their model also captures the aspect that objects can change their shape over time, and so different snapshots are extracted from different images in the video. In a similar context, the work from Fritz et al. (2007) addresses cross-modal learning of visual categories. Here, spatial reasoning is applied to associate visual categories to different objects in one image, for which a description is given by a human tutor.

One of the first approaches to language acquisition was published by Gorin et al. (1991), who presented experiments for automated call routing based on text input. In this study associations were discovered between words and meaningful machine actions. Mutual information was used as a measure to represent the weights in the employed information-theoretic networks. The extension of this work to spoken input was published later by Gorin et al. (1994). In these experiments, the acquired linguistic units (words) were associated with fixed semantics represented by a list of preprogrammed machine actions, resulting in single-layer information-theoretic networks. An extension of these experimental studies was later made to a system for automatically routing telephone calls using the caller's natural spoken inquiry, Gorin et al. (1997).

Without considering semantics, a new object which is learned by a robot is first of all simply a 'thing'. However, for the robot to make use of an object, or to discover completely new objects, the robot must be able to 'understand' the meaning of an object and ways to use the object. Such meaning is modeled by semantics. The semantics of an object also determine how it is referenced in speech, and especially the type of grammatical constructs. A robot who learns new objects thus should also be able to acquire semantics associated to this object. In contrast to Roy (1999, 2003), where semantics are learned implicitly, the following approaches use an explicit model of semantics. Learning of semantics, but not directly addressing learning of objects, has already been studied outside the context of human-robot interaction, for example in early work by Carbonell Carbonell (1979) with the systems FOUL-UP and POLITICS. Both systems were created as dialog systems for learning semantics of language.

Recent work for learning semantics also in dialog has been presented by Dusan and Flanagan (2002, 2003) with the spoken dialog system ABILITY, which is capable of learning new words and phrases during the interaction with users. After learning, users could use these new words during their future interactions with the system. They present an adaptive dialog system, which can lean new words, phrases, sentences and their meaning. Possible input modalities are speech, drawings, pointing gestures (pen), and video. The speech understanding component consists in part of a semantic grammar, and a semantic database, which contains assignments of objects to semantic attributes from a shallow ontological structure. For recognition of new words, first, the speech recognition hypothesis must be rejected by the speech understanding component, then, a second recognition pass is computed with a larger grammar. With this system it is possible to extend the understanding component, and learn new words and their association to simple predefined semantic meanings. Besides objects, also actions can be learned, e.g., by saying "remove is the same as delete", which results in

With application of interactive learning to the scenario of a humanoid robot in a household environment, it can be necessary to acquire rich semantic information for object models. Acquisition of rich semantic information is described in a robot training center by Becher et al. (2006); Kasper et al. (2007). The approach presents a fully integrated system for teaching a robot in an interactive manner with supervised data annotation in a training center. Among various other aspects of teaching the robot, an ontology of objects can be developed interactively. The ontology describes an object hierarchy; properties, attributes and actions can be associated with an object. Sensor feedback is used to detect selected features automatically, these can then be confirmed by the user. In contrast to the focus of this thesis, here, training requires a system expert, who conducts training in a training center, which is not available in the household environment during runtime of the robot. The approach rather intends to create high-quality initial knowledge models.

The approach presented in this work is rather intended to easily acquire information about a previously unknown object with comparable short dialogs and interaction with 'end-users' instead of system developers. A decision for training center or dialog-based learning does not need to be exclusive. It seems reasonable to start with initial training as described by Becher et al. (2006); Kasper et al. (2007) and extend the ontology during runtime with dialog-based learning methods. In addition, while most of the approaches presented above address only some aspects of object learning, our approach combines the aspects presented above within one system, namely learning new words, semantic concepts and properties, and integrates with visual object recognition for grounding of the objects in the real world.

2.4.2 Acquiring Personal Information

The main difference of acquisition of personal information to previously introduced approaches, is that usually these dialogs are conducted with persons about themselves. Some learning can only be conducted if the robots talks to the target person directly, e.g. for obtaining visual data, and for obtaining social information, where the person's opinion is required.

In contrast to object learning, grounding of persons can be conducted with models that integrate more prior knowledge. For example, state-of-the-art algorithms can detect faces and their poses in cluttered backgrounds (Viola and Jones, 2001; Gu et al., 2001; Schneiderman and Kanade, 2000). Grounding algorithms have been successful in detecting which person is speaking when, by multimodal integration (Lang et al., 2003), and later with the focus on robust tracking algorithms (Nickel and Stiefelhagen, 2007). These perceptual technologies provide basic support for the integration of userID in a dialog system, which needs to distinguish between speakers or localize the speaker in the environment.

Modeling user ID in a dialog system in the past has not necessarily been conducted with the additional challenge of grounding. Earlier work integrates for example face identification, voice identification, or pure spoken person registration in static settings and fixed environments. Later it was shown that the time of recording images has significant impact on face ID classification rate. Images of a single person, recorded during a single session, are significantly easier to classify than images recorded a few hours later, or even images recorded months later.

A faceID system that deals with real-life problem has been demonstrated in Sakaue et al. (2006), where experiments are presented that have been conducted during 15 days in an apartment rented specifically for this purpose. During that time, the apartment was inhabited by a 3 person family, and five robots were placed in the apartment to identify the person at different locations during any time of the day. Another real world challenge is presented in Ekenel and Stiefelhagen (2007), which describes a publicly available database, with over 100 persons recorded during February 2005 and from August to December 2005.

A dialog-based multimodal user registration system is described in Huang et al. (2000). Goal of the system is to register a user in a multimodal room during dialog-controlled interactions, with face identification, speech input, and information fusion in a Bayesian network. After registration, users can update their profile such as email address and phone number.

A humanoid robot with a person memory is the robot REEM-A and REEM B. Both robots have been demonstrated in videos[3],[4] distributed on internet platforms. The demonstrated capability shows capabilities of open set face identification. That means a person is identified if stored in the database. If no models exist for the person, the class *unknown* should be detected. Its learning mechanism is quite simple. If faceID reports the *unknown* class, a name can be said, which is then associated with the newly collected data.

Kim et al. (2006) describes a face identification method for a humanoid robot that uses a combination strategy of several features. Their architecture also supports interactive enrollment of users. The detection of the *unknown* class is based on visual features only, i.e. no dialog strategy is used to identify persons or train new persons.

[3] http://www.youtube.com/watch?v=B_jllEvrOZQ

[4] http://www.youtube.com/watch?v=Rb7oU8J-ZV0

2.4.3 Names and New Words Learning

Another important aspect of knowledge acquisition for a humanoid robot is to learn the word itself, i.e. to deal with unknown words in speech recognition, including detection and learning of new words. This is even a crucial aspect, if object learning is extended to an unrestricted set of objects, which we assume to be a criterion of real-world environments. Not all words in an open domain can be covered by a speech recognizer beforehand. For example, the number of all person names is too large for a standard speech recognizer to contain all names in its vocabulary. The problem of unknown words is also referred to as out-of-vocabulary (OOV) problem and it exists in several domains. So far, different approaches exist that address the OOV problem and the dealing with large vocabulary sizes, e.g. Young (1993); Slobada and Waibel (1996); Hetherington (1995); Schaaf (2001); Park and Glass (2006).

The problem of a large vocabulary is that recognition with a large vocabulary is basically slower than recognition with a small vocabulary. Especially, a problem exists in an interactive system if recognition is significantly slower than real-time. In addition, recognition accuracy decreases, if the vocabulary contains too many words. This problem is described in detail, for example, in Schaaf (2004). Thus, a tradeoff has to be found between a small, efficient vocabulary, with coverage of the most frequent words, and a large vocabulary which contains additional words.

Promising approaches, to efficiently deal with a large vocabulary, are dynamic vocabulary approaches. They basically follow the idea of name recognition on moderately large vocabularies, which provide good recognition accuracy, and to use a larger vocabulary when an OOV word occurs. Approaches that allow dynamic adaptation of the vocabulary are presented in Chung et al. Chung et al. (2004); Chung (2001), who describe a system with a dynamic vocabulary that can be updated according to the given context. The approach from Scharenborg and Seneff (2005) runs multiple recognition passes on speech input with a phone-based OOV word-model in the first step, which is used to constrain the vocabulary in the second step that best matches the resulting phone graph. Choueiter et al. (2007) present experiments with a subword modeling approach in a multi stage recognizer, to obtain spelling and pronunciation of new words. Their experiments show improvements over a large vocabulary isolated word recognizer. Kaiser (2006) present an approach for word acquisition in meetings by combining speech recognition and handwriting from redundant multimodal data, captured by microphones and whiteboard drawing.

Besides detection of unknown words special attention is required for obtaining a phonetic representation of a name which can be used to understand the user's name as well as to allow the system to pronounce the name. Chung and Seneff (2002) combine phoneme recognition of spoken input to obtain phonetic representation of names with telephone keypad input to obtain textual representation of names. Hild et al. combine spelling recognition (Hild and Waibel, 1995) with spoken names (Meier and Hild, 1997) for fusion of spoken and spelled names on large vocabularies.

Work presented in this thesis integrates the approach described in Schaaf (2001, 2004), which uses so-called Head-Tail models for acoustic modeling of unknown words. It has the advantage that it can be integrated with our speech recognition grammar, which also gives information about a possible semantic meaning of the OOV, based on grammatical construction of the utterance. It allows unknown word detection with a restricted vocabulary in a first step for efficient decoding in real-time, so that a second recognition can be performed in a second step with a broader vocabulary, only on utterances where an unknown word has been detected.

2.4.4 Active Learning and Data Cleansing

Another important aspect of system that can learn over time is to be able to correct information that has been acquired at some time, for which in part unsupervised learning mechanisms are relevant to detect problems or contradictory information. A field that slightly relates to this idea is an area called 'active learning'. The idea of active learning is to efficiently select samples, which are annotated by a human, for training of a classifier. Efficiency means to reduce the amount of workload of the human as much as possible. The difference to the dialog-based learning approach is that in active learning, the human can in fact 'annotate' samples, and this kind of notion does not exist in the dialog-based learning approach, as the 'samples' exist only internally and cannot be observed by the human. However, a series of techniques from active learning are also applicable to the unsupervised learning approach for problem detection in our approach. The idea which can be formulated as "samples in the same cluster are likely to have the same label" is also followed by Chapelle et al. (2002). Data therefore is pre-clustered and successively the best samples are selected. This approach will be adopted in this work, namely by pre-clustering data, and successively selecting the next label which needs clarification. Obviously the selection algorithm and the dialog interactions are conducted differently.

A second related field is work on deduplication in databases and data cleansing Maletic and Marcus (2000). The problems can be formulated in a similar fashion, while deduplication in databases targets at detecting double occurrences of entities that have been specified by different entries, e.g. due to typing errors, different formats, etc., this problem also exists in dialog-based learning, as objects can be learned twice with different spellings. Differences are in the treatment of inconsistencies, e.g. one problem in dialog-based learning are different persons that are stored with the same label, and in the kind of features that are used, as deduplication in databases is generally based on textual entries. However, similar algorithms can be applied. Maletic and Marcus (2000) describes data cleansing as an approach for error resolution in databases and suggests the following three-stage model:

- define and determine error types
- search and identify error instances
- correct the uncovered errors

All three steps apply also to our problem, though different algorithms might be applied. When looking more closely at methods applied to step 2, i.e. searching and identifying error instances, one can find out that among other approaches, such as pattern-based and association rules, in deed similar algorithms are applied here, again including clustering and statistical approaches (e.g. Yang et al. (1999)). The approach presented here cannot use either approach, active learning or deduplication directly without modification. Rather, we have adopted the same ideas and process model and have applied problem-adequate metrics for clustering, problem detection and resolution. Especially the steps of detecting problems and resolve errors require new methods than what we have seen so far, as the error resolution dialogs does not follow the idea of annotation of samples, but rather information must be presented in a way to effectively resolve these errors in dialog.

2.4.5 Conclusion

Several approaches for interactive learning have been compared in this chapter. The different approaches can be distinguished by the main focus of each approach. What is missing so far, is learning over a longer period of time, including mechanisms to correct stored information or knowledge. Some dialog approaches apply only simple predefined dialog strategies or are restricted to labels from a limited vocabulary. One contribution of this thesis therefore is optimization of dialog strategies for such a learning task. Most approaches have different main focus, and only few integrated approaches exist. Therefore, this thesis presents a new integrated approach, with a multimodal knowledge base and integration in a dialog system, which can learn during runtime and runs completely autonomous over longer period of time in social setting. Each aspect of the knowledge base builds on approaches presented so far, e.g., for modeling of semantics, or learning of new words in speech recognition. In contrast to work which focuses on learning from scratch, e.g. Gorniak and Roy (2005), work presented in this thesis can be understood as an approach which extends existing knowledge structures during runtime. For example, the system basically uses predefined knowledge about how dialogs are conducted. To give a second example, in the case of object learning, the robot uses an existing ontology of objects which is extended during runtime. Such a knowledge structure is typically defined manually or created during design time of the robot in a training center, as described by Becher et al. (2006).

Part I

A Dialog System for Multimodal Human-Robot-Interaction

Chapter 3

Dialog System Architecture and Speech Processing

This chapter first introduces the architecture of the dialog toolkit TAPAS which has been designed specifically for the task of multimodal human-robot interaction and learning. It is the foundation for the remainder of this thesis, and experiments presented in the following have been conducted with this toolkit. Following the description of the architecture, this chapter describes how speech is processed within this architecture with a tight coupling scheme, which allows contextual control of speech recognition and contextual interpretation.

The following chapters seamlessly continue the description of the dialog system. Further aspects of the system are the integration of a multimodal user ID model in the dialog architecture (chapter 4), optimization of dialog strategies such as reinforcement learning (chapter 5) for creating robust error-tolerant dialog strategies, and a modular dialog architecture (chapter 6), which allows mixing handcrafted dialog strategies and dialog strategies trained by reinforcement learning, e.g. as used in the receptionist robot (chapter 8).

The dialog toolkit TAPAS has been used to develop various dialog systems including multimodal multimedia access (Metze et al., 2005; Gieselmann and Holzapfel, 2005; Holzapfel, 2005) within the EU-project FAME, and several human-robot interaction scenarios (Stiefelhagen et al., 2007) within the Collaborative Research Center SFB588 on humanoid robots, such as a barkeeper robot (Prommer et al., 2006), a receptionist robot (Holzapfel and Waibel, 2007), dialogs for basic robot control tasks and robot services (Holzapfel, 2008; Fügen et al., 2006). In Gieselmann (2007) the dialog manager was used for studying and developing error-recovery strategies in multimodal interaction with a humanoid robot and through a web interface.

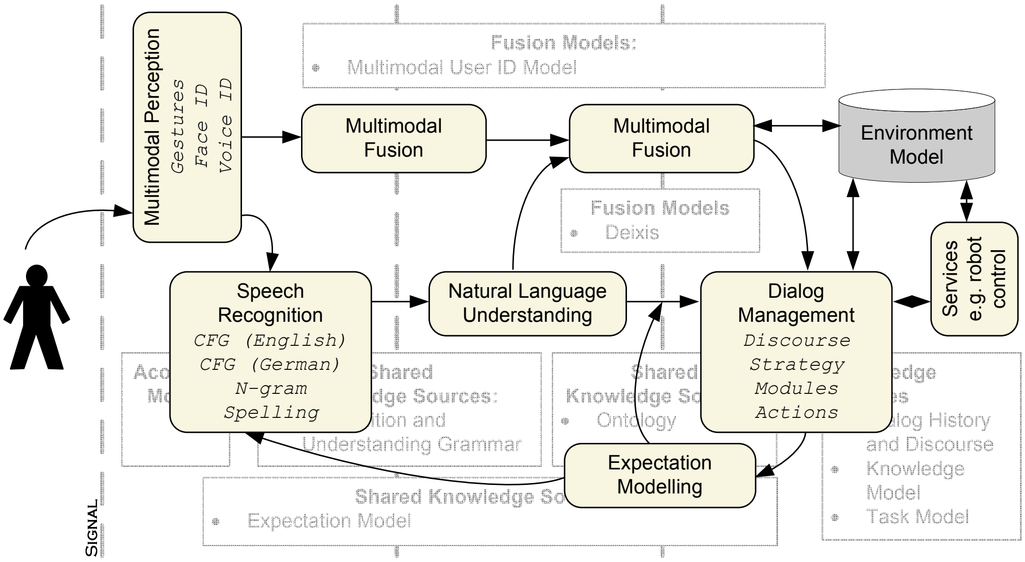


Figure 3.1: TAPAS dialog system architecture with tight coupling and multimodal fusion

3.1 System Architecture

The presented scenario of human-robot interaction and learning necessitates perceptual technologies for different modalities and their integration in the dialog system. Enabling such kind of interaction requires to overcome various challenges, including scientific problems and engineering tasks. A dialog system to enable such kind of interaction must be implemented efficiently so that the system can react in real time, must be robust against recognition errors, must be flexible to adapt to changing environments, and finally it must be scalable and extensible. We address these challenges and present results on different levels of the system hierarchy, including tight coupling of recognition components as presented in this chapter.

3.1.1 System Overview

Figure 3.1 shows this architecture. It highlights the major components and the data-flow between these components. The diagram is based on figure 1.2, which has already been shown in the introduction, and which highlights the knowledge sources. The system shows an integrated architecture and an extended tight coupling scheme with multimodal integration. The general components on the perceptual side of a dialog system are the recognition and understanding components, multimodal fusion, and interpretation. The output of the contextual interpretation component is fed into the discourse model, which delivers state information to the dialog strategy. The dialog strategy decides which actions to take. These actions generate sytem output with text generation and text-to-speech. The dialog actions can virtually execute any kind of functions to communicate with other system components, for example to update the database or trigger robot actions. But almost every dialog action also generates text-to-speech output and updates the expectations model. The dialog manager is the central decision-making component, controls the strategy and directs all system actions including learning tasks.

Tight coupling is achieved by the generic expectaions model for speech processing, by multimodal integration and by sharing of knowledge sources. A generic contextual expectations model supports understanding of words from a very large vocabulary and learning of new terms in speech recognition, facilitates contextual weighting of grammar rules, discourse update, resolution of elliptical utterances, reference resolution and multimodal fusion, and has access to knowledge sources of different components. The expectations model for speech recognition has already been published in Holzapfel and Waibel (2006) and is presented in the following section with slight modifications to the original publication. Multimodal integration includes multimodal fusion of speech and 3D pointing gestures, as published in Holzapfel et al. (2004), and multimodal fusion on different system levels for person identification with a user model, as described in chapter 4. The dialog system furthermore is coupled loosely with visual object recognition (chapter 7) and the robot control architecture for the humanoid robot Armar III.

The traditional dialog architecture is created by sequential processing of speech recognition, natural language understanding and dialog management. However, recent approaches show that interconnection of these components improves system performance in several ways. On the recognition side, valuable information is given by the dialog state that can be used to improve speech recognition in dialog context. Sharing linguistic knowledge sources, i.e. recognition and understanding grammars, improves processing speed and robustness, less knowledge sources need to be maintained. Both approaches are compared in figure 3.2 In previous experiments (Fügen et al. (2004)) we have shown improvements in recognition accuracy over loose coupling especially for contextual utterances and distant speech recognition, which confirms previous work with contextual weighting schemes such as Stent et al. (1999), and Lemon (2004), who show improvements by using subgrammars.

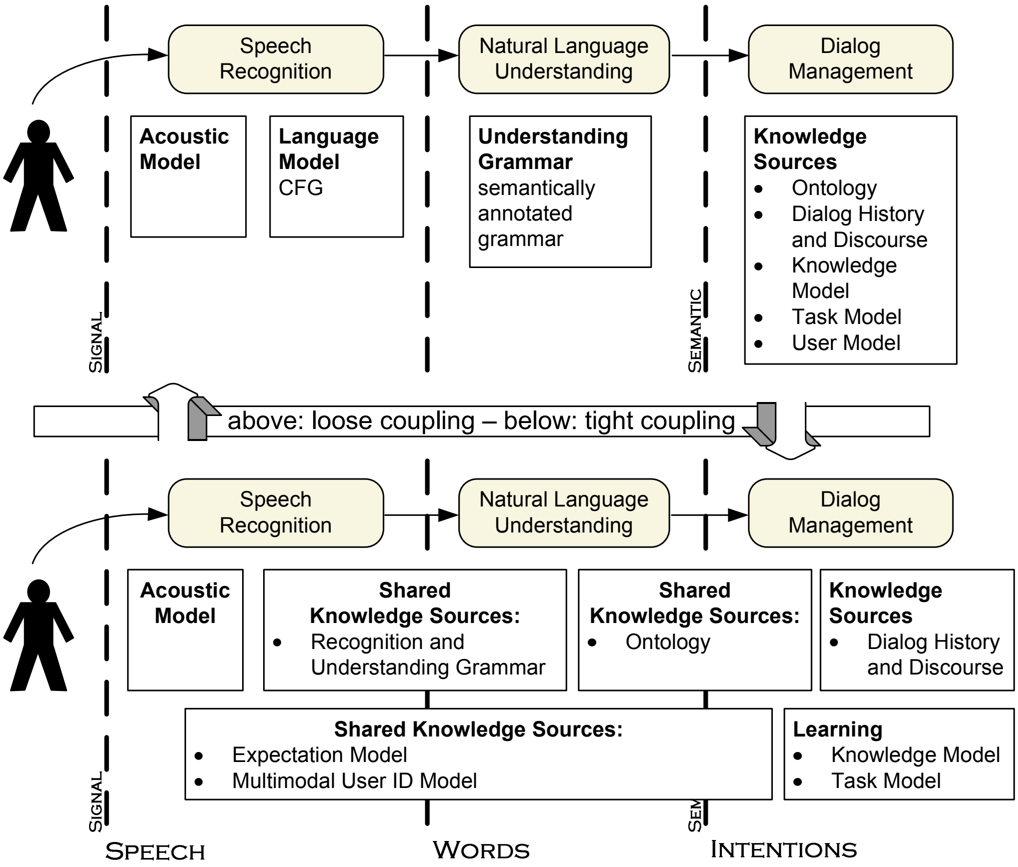

Speech Recognition
Natural Language Understanding
Dialog Management
Acoustic Model
Language Model CFG
Understanding Grammar semantically annotated grammar
Knowledge Sources
• Ontology
• Dialog History and Discourse
• Knowledge Model
• Task Model
• User Model
SIGNAL
SEMANTIC
above: loose coupling – below: tight coupling
Speech Recognition
Natural Language Understanding
Dialog Management
Acoustic Model
Shared Knowledge Sources:
• Recognition and Understanding Grammar
Shared Knowledge Sources:
• Ontology
Knowledge Sources
• Dialog History and Discourse
Shared Knowledge Sources:
• Expectation Model
• Multimodal User ID Model
Learning
• Knowledge Model
• Task Model
SIGNAL
SEM
SPEECH
WORDS
INTENTIONS

Figure 3.2: Dialog systems architecture with comparison of loose and tight coupling

3.1.2 Speech Processing with Multilingual Grammars

The interface between dialog manager and speech recognizer is by text and semantic information in the one way, and context-dependent adaptation of the speech recognizer in the other way.

Generally, a speech recognizer produces text, which is bound to its language model. In up-to-date speech recognition systems the language model is either an n-gram model, a context free grammar, or recently a hybrid model. The advantage of the n-gram model is its flexibility to recognize any kind of input, but a requires large amounts of training data. The output of an n-gram model is text, and an additional stage is required to convert the text to semantics with a natural language understanding compontent. The advantage of the context free grammar (CFG) is that it requires no or only little training data as it is a handcrafted model, and that the ouput of the context free grammar is closer to a semantics, which can be encoded in the parse tree. A drawback of context free grammars is that the model can only understand utterances that are covered by the grammar. Both models are used in current dialog systems, n-gram language models often are preferred where unrestricted speech is necessary or when large amounts of in-domain training data are available. Context free grammars are used in restricted scenarios or where an initial model has to be created with no or only little in-domain data.

Most dialog systems implemented with TAPAS use context free grammars, and some applications exist which use n-gram models, where recognition of unrestricted speech input is necessary, for example in chapter 9. The context free grammars used in our system already encode semantic information and are therefore called semantic context free grammars (Gavalda (2000)). The grammars are shared by the speech recognizer and the dialog system. In fact, the dialog manager generates runtime grammar models, which are loaded by the speech recognizer, so that both components share the same models. As described in detail in section 6.3.2, the semantic grammars define rules for converting the recognition input to semantics. Furthermore, since the models are shared, the dialog system can update the models during runtime, e.g. by adding new words or changing rule probabilities depending on the context. Also the speech recognizer directly outputs a parse-tree, so that no additional and possibly ambiguous parsing is necessary. The dialog manager's grammars are syntactically specified in JSGF. The generated grammars can be adapted to the speech recognizer that is being used, and include SOUP, PHOENIX, JSGF, and Microsoft SAPI grammar formats.

The design of the complete dialog manager is language independent. When implementing a dialog system for a specific language or porting a dialog application to a new language, recognition grammars and speech generation grammars need to be implemented as language specific parts. The different recognition grammars for language-specific recognition modules, as shown in figure 3.1, are generated from a joint multilingual grammar resource using the environment model and database information and will be described further in chapter 6 in the context of knowledge-acquisition. By using the language independent design of semantic structures, it is possible to work with a generic dialog context and expectations model on a semantic level.

3.1.3 Speech Recognition

For speech recognition, we use the Janus Recognition Toolkit (JRTk) presented in Finke et al. (1997) with the Ibis single pass-decoder by Soltau et al. (2001). In combination with the context free grammars generated by the dialog manager, the Ibis decoder directly uses these grammars as the language model for speech recognition. Additional speed up for the interface between Janus and Tapas is achieved by sending n-best lists of parse-trees instead of text hypotheses, so that the language understanding component can skip the parsing step and directly use the result of the speech recognizer. This is furthermore of advantage, as in case of ambiguous parses, the parse result of the speech recognizer depends on contextual weighting from the dialog.

For detection of unknown words (OOVs), the speech recognizer provides head-tail models to detect unknown words on the phonetic level, which have been introduced by Schaaf (2004). We use these unknown-word models within a dynamic vocabulary recognition approach for detection of unknown words (OOVs), adding of new words during runtime, and multi-stage decoding. With multi-stage decoding the recognizer can run with an efficiently small name vocabulary and if necessary, re-decode speech input with a larger but slower vocabulary for less frequent names.

3.2 Context Modeling and Tight Coupling with Speech Recognition

This section presents a generic expectations model for contextual weighting of speech recognition, which is also used for interpretation of contextual utterances (e.g. resolution of elliptical expressions). The model is designed as a domain- and language-independent approach. It is integrated in the Tapas architecture as a generic construct and is applied in all applications in this thesis. The generic expectations model is part of the tight-coupling scheme, and it can be shown that the contextual model improves recognition accuracy. Part of the work presented this section have already been published in Holzapfel and Waibel (2006). In contrast to related work, the approach presented here is generic in a way that for a new system no training is required to enable contextual weighting. Rather, grammar weights are determined by correlations of expected information types with grammar rules using ontological information. Experiments in a robot-barkeeper scenario as described in the following with generic contextual weighting show improvements of 33% (relative) on close-speech and 21% (relative) on distant-speech recordings. Recognition rates have been measured on semantic concepts at 5.2% error rate for close-speech and 15.7% error rate for distant-speech.

3.2.1 Foundations of the Expectations Model

Existing work already makes use of contextual control of the speech recognizer. For example the information state update (ISU) dialog manager Lemon (2004) uses grammar-switching, based on the assumption that dialogs consist of adjacency pairs where that answers follow questions, commands are acknowledged, etc. In grammar-switching, the sub-grammars are determined by the previous dialog action of the system. Most researchers working on contextual control of a speech recognizer by means of a dialog manager use different stages and language models: Xu and Rudnicky (2000a); Fosler-Lusier and Kuo (2001) use a general n-gram language model which is used at the beginning and in underspecified situations and a specialized language model which can be an n-gram language model or a grammar-based one and is used in specific situations based on the preceding system prompt. In Solsona et al. (2002), the state-independent n-gram language model is also combined with a state-dependent finite state grammar by comparing the acoustic confidence scores. In this way, perplexity and word error rates can be reduced significantly.

As in Lemon (2004), our approach makes use of the assumption of adjacency pairs, and it extends the context switching model with more detailed utterance categorization, and the type of requested information from ontological information. The approach can be applied to any semantic grammar and in contrast to generating different sub-grammars, only a single grammar is used where the probability of different grammar rules is adapted. Speech act theory has become common to model and categorize specific actions in dialog systems (Traum (1999)). Beyond speech acts, Traum and Hinkelman (1992) describe conversation acts that cover additional actions in dialog such as turn taking and grounding. They define four speech act categories, 'turn-taking', 'grounding', 'core speech acts', and 'argumentation'. Different annotation and labeling schemes have been developed for speech acts like DAMSL[1], or SWBD-DAMSL. Our dialog system uses a specific speech act called 'information request' that models almost any action or utterance that expects an answer from the conversation partner. For our analysis a more detailed classification of information requests is required, e.g. as used in CLARITY Levin et al. (1998b). The CLARITY annotation scheme is based on DAMSL and SWBD-DAMSL but provides more details especially for information requests. The categories for system utterances are similar to those used in CLARITY. They are described later in this section.

[1] http://www.cs.rochester.edu/research/cisd/resources/damsl/

3.2.2 Speech Acts and Utterance Classification

A dialog move is selected by the dialog strategy based on the purpose that it serves, for example to request new information, generate clarification questions, give information, or generate confirmations. Each purpose leads to a different response by the user, which can be predicted if each system utterance is associated to an utterance class, and the classes are well chosen. The classes that we have analyzed are shown in figure 3.3.

The speech acts shown in figure 3.3 all inherit from a general node 'info-request'. The 'info-request' is the most general element to describe a question (or any other kind of action) that expects an answer relating to this question. The top level node corresponds to the speech act category describing an information request in DAMSL. However, for our purposes the DAMSL tagging scheme is not detailed enough, so we extended the scheme to the following speech acts. 'qst_yesno' expects 'yes' or 'no' as answer; 'qst_wh' is a category for all 'wh'-questions such as who, what, when, where, and questions asking for numbers, which represent the subcategories of 'qst_wh'; 'qst_or' is a question, where the user can select one of the presented alternatives, e.g. "do you want x or y?"; 'qst_open' is an open question where the user is free to answer, and no explicit expectation can be generated based on the speech act. Here, only the type of requested information determines an expectations context. The last type 'qst_open' cannot restrict the expectations, whereas all others can. Core and Allen (1997) use a category that combines different actions that influence the addressee's future action. This category contains 'open option' and 'directive'. Subtypes of 'directive' are 'info-request' and 'action-directive'. Our approach goes in-line with this description and refines the information request category to do more detailed analyses.

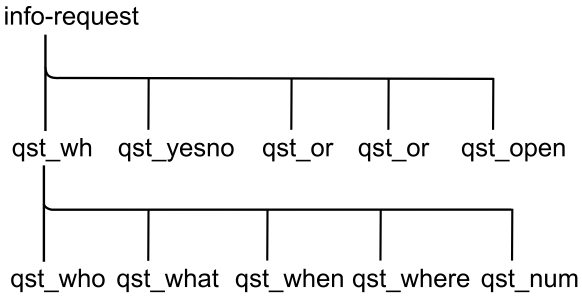


Figure 3.3: Utterance classification with inheritance model for subtypes of the 'question' speech act category used to generate information requests

3.2.3 Target Types and Contextual Weighting

The system's utterance class already provides relevant information to predict the next user utterance. Additional information from the dialog context, the type of information requested and other information to achieve the active dialog goal can further specify the expectations more precisely. The the order of importance of expected information is (i) direct response to question (ii) indirect response to question that implicitly answers the question (iii) response to question in combination with repeating information (iv) repairing previously given information (v) giving information for one of the active discourse segments.

We call the requested type of information a target. A target is a piece of information that is described by its semantic type, a reference to the dialog goal that defines the frame for the target and the TFS path to the desired information within the specified goal. As dialog goals, as well as the discourse representation, are modeled with typed feature structures, their structures define the context for the targeted information. Information required by the dialog goal, which is not given in discourse is expected to be delivered by the user. Information that is already given in discourse is either expected to be repeated, to be confirmed or to be repaired.

When asking for missing information, we expect to be able to extract this from the user's answer. The answer can be elliptic, giving directly the desired information, such as 'two' in reply to asking 'how many persons?'. Or, the answer can be embedded within a complete sentence. The construction algorithm for generating expectations based on the target information first picks the target type and then walking up in the TFS path, picks all parents recursively. This results in a list of TFS nodes describing the targeted information within more or less context of the dialog goal.

A small example illustrates the algorithm. Figure 3.4 shows the required information for a dialog goal. When executed, it instructs the robot to serve a cup of coffee, with the options of adding milk or sugar. The path '$OBJ|MILK$' references the type 'att_milk' with its sub-feature. To get information about the type 'att_milk', the system generates an information request. The target is defined by the dialog goal and the path '$OBJ|MILK$' that references the type 'att_milk'. The expected response can be 'yes' or 'no', which both directly respond to the given question. The answer 'yes' is converted to the following TFS and is then unified with the discourse representation with the prefix path '$OBJ|MILK$'.

```
{att_milk BOOL [base:boolean]}
```

Note that the expectations model also covers formulations like 'with milk please' or 'I would like my coffee with milk and sugar'. The answer 'with milk' is first converted to the above TFS and is then unified with the discourse with the prefix path '$OBJ|MILK$'. The same applies to the response 'with milk and sugar' which describes a more complex construct than 'with milk', but matches the expected information as well.

```
[act_bring
OBJ [obj_coffee
     MILK  [ att_milk
             BOOL [base:boolean] ]
     SUGAR [ att_sugar
             BOOL [base:boolean] ]
    ]
]
```

Figure 3.4: A TFS describing the precondition of the 'make-coffee' goal.

When a list of possible TFS nodes has been determined, the next step ist to select grammar rules that can generate the desired semantic representation from spoken input. The algorithm to find these grammar rules is constructive and uses induction to search all conversions of grammar nodes to a given semantic representation, where the semantic type of the grammar node matches the desired semantic type.

3.2.4 Experimental Results

We compared the speech recognition results of a system which uses the context dependent weighting of rules to one without it, on human-robot dialogs in the domain of a household robot. We evaluated the approach on two different interaction sets. Both sets were recorded with close talk and distant speech microphones.

Set 1 consists of requests for actions by the user (User Commands), responses by the system including clarification requests or queries for missing information where necessary, and user replies (Response Set). It contains eight speakers, all interaction are in English.

Set 2 was recorded in a different setup with different users in multimodal human-robot interaction, where the robot plays the part of a bartender to serve different objects from the table in front of him. The user responded to questions from the robot asking for object properties Prommer et al. (2006).

set	condition	baseline		improvement	
		WER	SER	WER	SER
Responses	close	29.11%	30.00%	8.87%	8.89%
Overall	close	22.74%	31.89%	3.56%	3.88%
Responses	distant	36.77%	39.60%	16.45%	11.86%
Overall	distant	31.41%	45.33%	6.66%	5.15%

Table 3.1: Set 1: Close and distant talking word and sentence error rates together with their relative improvements

The full set contains 314 utterances for English and 171 utterances for German, each including some segmentation errors (e.g. utterance was recognized though nothing was said) and out-of-domain utterances that are not covered by the system. The constrained set excludes out-of-domain utterances and segmentation errors, which results in a set size of 267 utterances for English and 152 utterances for German.

Three categories of weights have been used: unexpected, normal and expected. The grammar weights for these categories have already been trained in previous work Fügen et al. (2004).

Evaluation details on Set 1 with handcrafted weighting have already been presented in Fügen et al. (2004), in our experiment the selected rules offer a marginally broader selection of rules that however, did not have any effect in word-error rate, presumably because the hand-crafted selection was already very good. Table 3.1 shows the baseline (no rule weighting) and the relative improvements achieved on Set 1, measured with word error rate (WER) and sentence error rate (SER). The evaluation on the Set 2 is shown in table 3.2, where the figures for German (close-talk) and English (close-talk and distant-speech) are given. Here, we show the numbers for word-error rate (WER) and semantic concept error rate (CER) for both close-talk and distant-speech on the full set ('Overall') and a constrained set ('In Domain'). The relative improvements for the numbers are computed in table 3.3. We have chosen the concept error rate (CER) since it is useful to measure the effects on a dialog system. It is more informative than word error rate and also ignores semantically irrelevant errors. It is computed similar to the common word error rate by simply comparing IDs of semantic concepts. The results for the German baseline (without rule weighting) are already very good, which we attribute to the recording conditions with users that are familiar with ASR systems, which is not the case for the English recordings. As only few errors remain in the German set which are not due to segmentation errors or noises,

set	condition	baseline		improved	
		WER	CER	WER	CER
In Domain	close - English	12.8%	7.8%	10.1%	5.2%
Overall	close - English	28.3%	15.9%	26.2%	13.7%
In Domain	distant - English	32.1%	19.9%	29.2%	15.7%
Overall	distant - English	41.9%	26.4%	39.7%	21.7%
In Domain	close - German	9.8%	4.6%	9.1%	3.9%
Overall	close - German	21.3%	13.1%	20.9%	12.0%

Table 3.2: Set 2: all utterances and in domain utterances (parsable input) for close and distance talking conditions for English and close talk for German. Evaluated on word error rates (WER) and semantic concept error rates (CER).

set	condition	impr. WER	impr. CER	rel.impr. WER	rel.impr. CER
In Domain	close - English	2.7%	2.6%	21.1%	33.3%
Overall	close - English	2.1%	2.2%	7.4%	13.8%
In Domain	distant - English	2.9%	4.2%	9.0%	21.1%
Overall	distant - English	2.2%	4.7%	5.3%	17.8%
In Domain	close - German	0.7%	0.7%	7.1%	15.2%
Overall	close - German	0.4%	1.1%	1.9%	8.4%

Table 3.3: Absolute and relative improvements on Set 2

the improvements are smaller than for the English system. It is interesting to see that the improvements for the concept-error rate, which is more important for the dialog system are more significant than the improvement for the word-error rate.

As the general approach is language independent and operates on a semantic level, uses speech act theory and ontological information, it can be used in multilingual applications of TAPAS with multilingual or multiple monolingual speech recognizers, which use different language models, which is according to Schultz et al. (2003) a typical setting for multilingual speech recognition. In the experimented multilingual settings, the rule set is the same for the different languages. Language-specific information are automatically computed by contextual weighting of gammar nodes.

The experimental data set reflects aspects of a typical dialog-based learning scenario, which regularly requires confirmation questions, which expect confirmation or rejection by the user. At the same time the system also needs to recognize speech input that is not expected by the system, e.g. when the user reacts with more complex utterances.

3.3 Conclusion

This chapter has presented an overview of how the approaches introduced in this thesis fit into the dialog system architecture. The chapter also gave an overview of the interconnection of the dialog system components and presented experiments for an expectations model with tight coupling of speech recognition and dialog manager. The results show that the generic approach for contextual control of the language model improves speech recognition and especially semantic understanding rates. In a learning system this is an important achievement, as learning of new words is a hard task. By contextual weighting and by restricting the vocabulary, confirmations by the user can be recognized with high confidence and on the other hand understanding new words from a large vocabulary can be accomplished. As grammars are shared by dialog manager and speech recognizer, this model provides a basis for learning of new words during runtime and for dynamic vocabulary switching in dialog context. The general approach is language independent, as it operates on a semantic level, uses speech act theory and ontological information.

Chapter 4

Multimodal Person Identification in Dialog

4.1 Overview

This chapter presents a multimodal user identification model which keeps track of the user ID during a dialog session, including multi-layer information fusion of different modalities and integration of dialog state features. The approach extends existing work with an integrated method for identifying and learning persons including unknown person identification with multi-layer information fusion. As such, the approach extends existing work with an integrated method for identifying known and unknown persons in dialog. In this chapter, the following problems are addressed: to achieve better recognition rates than e.g. single image identification during natural interaction, to deal with unknown persons for new person learning, and to provide an estimation of classification reliability for better dialog decisions. The challenges are addressed by multimodal fusion of recognition results, fusion over time, confidence estimation, integration of dialog state information, and open set person identification. As the user ID model keeps track of the user ID in a natural interaction, the method is clearly distinguished from interactions like “please look directly into camera while I'm taking a picture of you”.

For the task of identification of a person, the literature distinguishes closed set identification, where the person to be identified belongs to a set of known persons for which training data exists, and open set identification, where the person to be identified is not always represented in the training data. A humanoid robot in a real world environment needs to deal with both, known and unknown persons. Identifying unknown persons is a prerequisite for learning persons. This model therefore provides the basis for learning persons, which includes identifying known and unknown persons, automatically collecting training data, adding the user to the set of known persons and updating the classification models. To achieve this learning functionality, appropriate identification methods must be applied, which are presented in this chapter.

Figure 4.1 gives an overview over the components integrated in the ap-

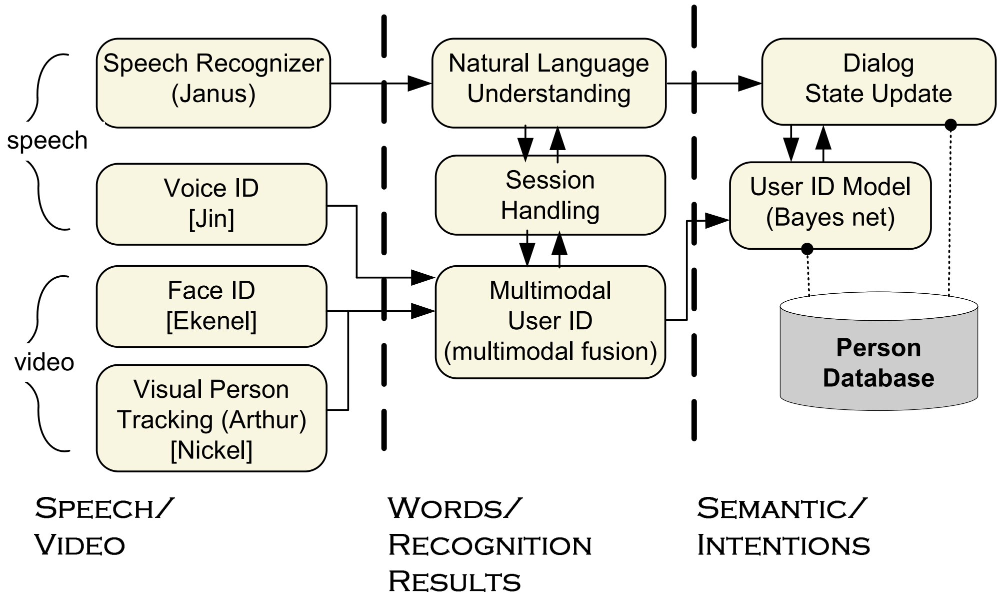

speech
video
Speech Recognizer (Janus)
Voice ID [Jin]
Face ID [Ekenel]
Visual Person Tracking (Arthur) [Nickel]
Natural Language Understanding
Session Handling
Multimodal User ID (multimodal fusion)
Dialog State Update
User ID Model (Bayes net)
Person Database
Speech/ Video
Words/ Recognition Results
Semantic/ Intentions

Figure 4.1: Components involved in multimodal user identification

proach for person identification in a multimodal user model. The components on the speech/video layer are not part of this thesis and have kindly been provided by colleagues. Each component is introduced briefly in the following section, and integration of speech recognition has already been described in detail. The separation of the used components and work conducted for this thesis is also shown in figure 4.2, which is also presented in more detail in the next section.

The user ID model keeps track of the person's ID to represent the person interacting with the robot, and provides ID information for the dialog manager. As learning of personal data requires high reliability in the user model, the system's confidence in its hypotheses can be increased by acquiring additional information from the user, e.g. by initiated an identification dialog, and let the user confirm the system's assumption. During the course of this work, we have designed several dialog strategies to implement this module's function, some of which are described in the following chapter(s). The scenario also requires information fusion on several levels. For example, information to identify a person during a dialog session is provided from speech input, from dialog features (e.g. confirmations), context, face ID and voice ID. A special consideration is that identification is conducted as an 'online' algorithm. 'Online' means that classification is performed during runtime. For example, as face identification can benefit from the observation of multiple images, we distinguish between single-image faceID and sequence faceID, the latter of which is computed from a sequence of images that have been recorded during the interaction up to the time of classification.

Evaluation of the approach for person identification in dialog with real data shows high recognition rates in both, closed and open set identification, and supports the hypothesis that the approach is suitable for the task of learning and identifying persons. To show effectiveness of the approach, different experiments are presented to evaluate the parts in isolation, and different dialog strategies are evaluated to optimize overall identification rate in dialog. A key aspect of user identification is the use of confidence measures, which are used in fusion of multimodal ID hypotheses, and also to compute a correct-classification measure that allows dynamic weighting of classification results in a belief network. The belief network computes a posterior probability for each person, including the 'unknown' model, which is then used as a belief state in the dialog system.

This chapter is based on a series of already published articles, with the initial system design of the author in Holzapfel et al. (2007), and subsequent experiments for confidence-based fusion in Könn et al. (2007) and Grosse et al. (2008), and information fusion on the dialog level in Holzapfel and Waibel (2008b). It should be noted that the latter three publications include results from three Studienarbeiten at the Universität Karlsruhe (TH) by Stephan Könn, Philipp Große, and Philipp Hüthwohl, supervised of the author of this thesis. Section 4.2.1, i.e. description of the faceID component, has been taken from Holzapfel et al. (2007), and has been written by Hazim Kemal Ekenel. Section 4.2.3 is taken from Grosse et al. (2008) with minor modifications, and sections 4.4 and 4.5 are taken from Holzapfel and Waibel (2008b) with minor modifications.

4.2 Confidence-Based Multimodal User ID

As has already been motivated, confidence measures are an important aspect of integrating faceID in the dialog. This section presents the approach for multimodal user ID modeling (with faceID and voiceID), which can be summarized as confidence-based multimodal user ID. The section begins with a brief overview of the face identification and voice identification components with references to the original work of these components.

The dialog system is designed as a one-to-one communication system assuming that one person talks to the robot. Grounding of the speaker in the real world is done by a (mostly vision-based) person tracking library (Arthur) published by Nickel and Stiefelhagen (2007). The system is designed for one-to-one communication (in contrast to multi-party communication), but in a real-world scenario additional persons might appear interrupt the visual field of the robot. Therefore, Arthur provides person tracks, and image data for face recognition is only extracted from the relevant track.

4.2.1 Face ID Component

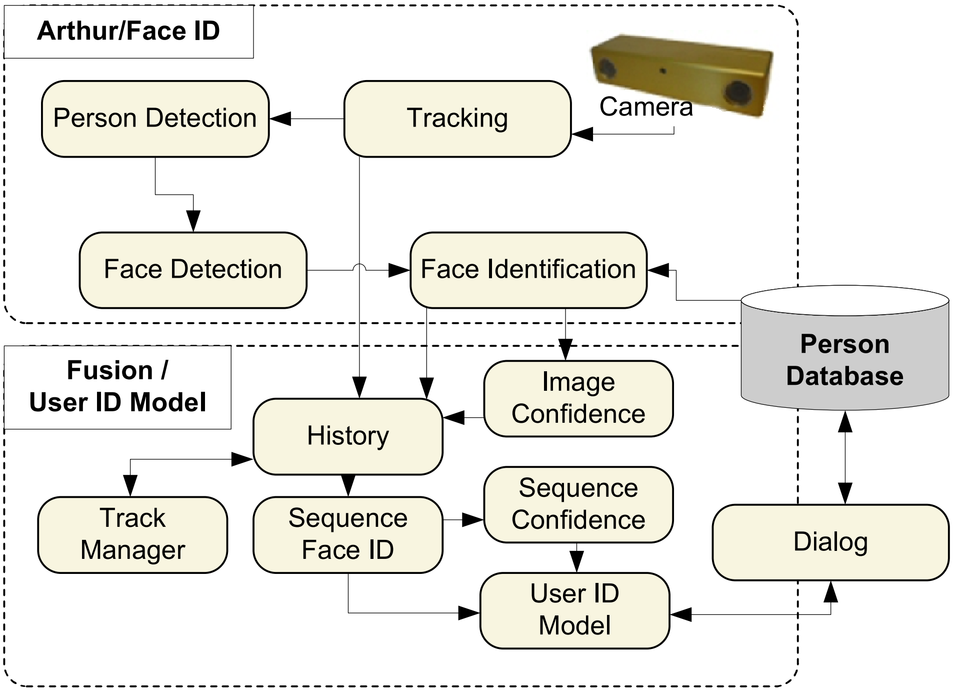


Figure 4.2: System architecture with face recognition system

In order to create an assumption about the person in front of the robot, several processes have to be passed that interact with and depend on each other. Figure 4.2 gives an overview of the components involved and logical categorization of processing units. The box "Arthur/Face ID" corresponds to the integrated face identification and tracking components which are integrated through the Arthur tracking library. As the first component of the chain, the robot camera takes images in regular intervals, which is a process controlled (including head movement and multi-person tracking) by the Arthur tracking library. Face identification is then conducted by subsequent face detection, eye detection, face alignment and classification. The user ID model then integrates these face identification results over time by constrained-based fusion, track selection and by multimodal fusion and produces a multimodal user ID hypothesis and a confidence value estimating the probability of a correct hypothesis. The whole process is performed for each new image recorded by the robot camera, i.e. sequence hypotheses and confidences will be updated online for each new frame.

Face classification is conducted with a local appearance-based approach Ekenel et al. (2007) which first detects face regions in a single image, aligns and normalizes these segments and then extracts discrete cosine transform (DCT) features from 8x8 pixel blocks. The extracted feature vector is compared against prototypes stored in a database using a nearest neighbor classifier. The approach is extensively tested on the publicly available face databases and compared with the other well known face recognition approaches. The experimental results showed that the proposed local appearance based approach performs significantly better than the traditional face recognition approaches. Recognition results from the face recognition grand challenge (FRGC) version 1 data set for face verification, described in Ekenel and Stiefelhagen (2005a), and recognition results from the FRGC version 2 data set for face recognition, described in Ekenel and Stiefelhagen (2006), give an understanding what performance can be expected in the best case in a scenario as descibed in this thesis. For example, in the conducted experiments on the FRGC version 2 data set with 120 individuals and ten training and testing images each, 96.8% correct recognition rate is obtained under controlled conditions and 80.5% correct recognition rate is obtained under uncontrolled conditions. Further details of the component and video-based face recognition can be found in Ekenel and Pnevmatikakis (2006); Ekenel and Jin (2006); Ekenel and Stiefelhagen (2005b).

To support learning of new persons, the face identifier supports extending the models during runtime, more or less in real-time, by adding new feature vectors to the model. In this regard, the DCT- and nearest-neighbor-based model provides another benefit, since new samples can easily be incorporated into the database, while other methods demand a complete revision of data, e.g. the PCA approach described by Turk and Pentland (1991). As more and more feature vectors are added to the model by constant recording, the nearest neighbor classifier tends to develop a bias towards classes with over-represented classes. As also the computational efforts increase significantly and affect the real-time behavior of the system, regular automatic data reduction steps have been conducted. In the experimental system, k-means and random selection, after 10% outlier removal, with a limit of 400 entries per class, provided good results to maintain a constant quality. Such updates were conducted during ‘sleep’ phases of the robot, e.g. at night.

4.2.2 Voice ID Component

Voice identification in this thesis is based on the approach presented in Jin et al. (2007), with Gaussian Mixture Models (GMM). Voice ID is computed on close talk microphone input, which is also used for standard dialog interaction. No modifications have been made to the component itself, which is integrated as a block-box component. Segmentation of utterances is conducted by automatic speech segmentation, a component which is part of the speech recognition toolkit. Similar as in the face identification approach, voice identification integrates speech segments iteratively while the dialog session continues. Experiments in the following integrate the component and additionally include confidence measures for multimodal fusion.

4.2.3 Confidence-Based Multimodal Fusion

Architecture

The approach to multimodal fusion extends the approach from FaceID and conforms to the turn-based dialog system architecture. Each turn corresponds to a dialog utterance by the user during which a single audio file is recorded. Video images are recorded continuously during the dialog. In order to create an assumption about the person in front of the robot several processes have to be passed. Figure 4.3 gives an overview over the system architecture and its components. The system is divided into six separate subsystems (illustrated through dashed lines): the single image layer (1), the image sequence layer (2), the single turn layer (6), the concatenation turn layer (5), the audio sequence layer (4) and finally the central multimodal layer (3). Apart from the single turn layer all other subsystems provide hypotheses in the form of n-best hypothesis list and a corresponding confidence. The additional concatenation turn layer has been introduced to provide audio utterances adequate for voice ID. Since data collected with the dialog system includes many short utterances, most of which are shorter than 1 second, a *concat turn* is simply the concatenation of the single audio utterances, which have been recorded during the dialog.

Confidence-Based Fusion

We use confidence measures as basis of an adaptive weighting for fusion, which includes fusion of different modalities and per-modality fusion of single hypotheses to obtain sequence hypotheses. A “single hypothesis” represents an n-best hypothesis list, i.e. a list of hypotheses with probabilities that

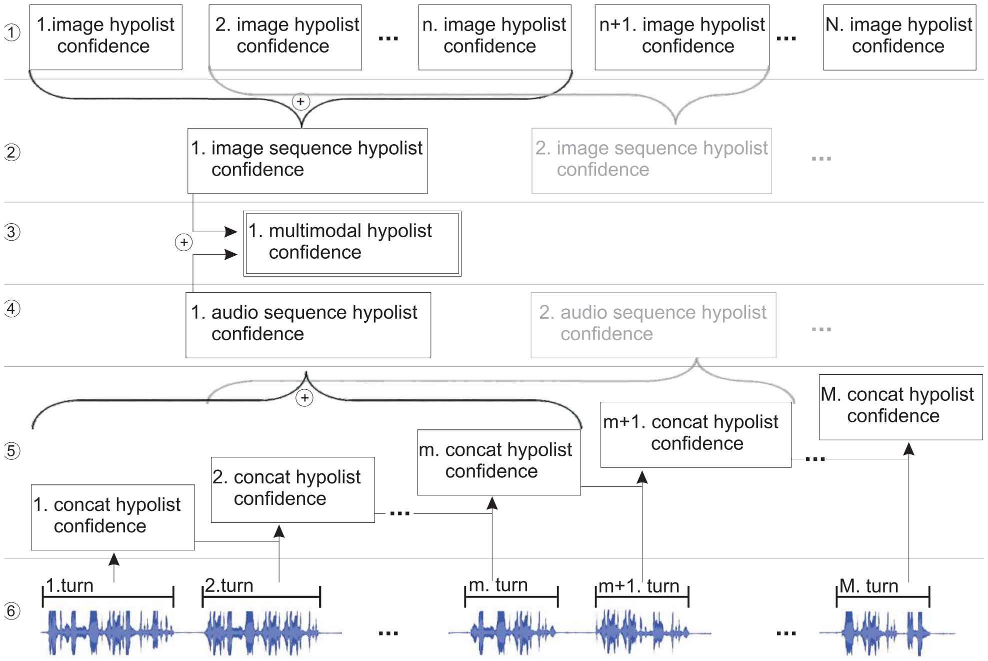

1
1.image hypolist confidence
2. image hypolist confidence
...
n. image hypolist confidence
n+1. image hypolist confidence
...
N. image hypolist confidence
+
2
1. image sequence hypolist confidence
2. image sequence hypolist confidence
...
3
+
1. multimodal hypolist confidence
4
1. audio sequence hypolist confidence
2. audio sequence hypolist confidence
...
+
5
1. concat hypolist confidence
2. concat hypolist confidence
...
m. concat hypolist confidence
m+1. concat hypolist confidence
...
M. concat hypolist confidence
6
1.turn
2.turn
...
m. turn
m+1. turn
...
M. turn

Figure 4.3: Structure of the multimodal classifier with different classification layers.

sum up to one. Fusion is realized as summation over n-best lists which are weighted by confidence estimates. This approach is illustrated in figure 4.3 on several layers, fusion is marked by $\oplus$. In mathematical terms, calculating the hypothesis of the next higher layer is described as:

$$H_{new} = \sum_{i=1}^{W} conf(H_i) \cdot H_i \tag{4.1}$$

where H respectively denotes an n-best list, and $conf(H_i)$ represents the confidence for this hypothesis. W refers to the *sliding windows* size, i.e. the number of accounted hypothesis lists. The fusion method is applied to each fusion step, including the multimodal integration stage, where only two hypothesis lists are merged. In the multimodal case, the fusion takes over the part of dynamic wheighting of the two modalities. In general, W denotes the maximum number of accounted single hypotheses. Since the classification approach is designed for a live (online) system, the system initially deals with small numbers of W which grows until the maximum sequence length is reached. From there on the sequence is shifted as a sliding window over the single hypotheses with a fixed size.

Each fusion step requires that hypothesis lists are normalized. In literature, different normalization techniques can be found. For the presented system we have been using a technique that is robust against different classifier types, i.e. k-Nearest-Neighbor and GMMs. It preserves the n best scores and distributes the probability mass over these results as shown in equation 4.2.

$$\bar{s}_i = \frac{s_i - min}{\sum_{i=1}^{n} (s_i - min)} \tag{4.2}$$

s_i denotes the score of the i-th best hypothesis, n denotes the length of the new hypothesis list, and min denotes the smallest boundary score, i.e. score with index $n + 1$ which defaults to 0 if less than $n + 1$ values exist. In the presented system we have been working with a list size of ten. If the list size remains constant during fusion, the normalization function decomposes to a scalar value for the whole list which only depends on the confidence values and the number of hypotheses W. The normalization method is also applied to the outcome of equation 4.1 when a constant list size, e.g. ten, is required.

4.2.4 Confidence Estimation and Confidence Features

The term *confidence* refers to the reliability of the classification, i.e. the likelihood that a given classification result is correct. This confidence is approximated by a logistic regression model (Hosmer and Lemeshow, 1989), which is used to model the likelihood of an event as a function of predictor variables. In this case, the binary, dependent variable Y models the event *classification correct* ($Y = 1$) and *classification incorrect* ($Y = 0$). In the multi-tier architecture, for each level, i.e. for each classifier, a separate regression model needs to be trained. Correspondingly, for each regression model the features are selected separately.

For face identification we have analyzed features from the vision system, such as *Image*: the mean gray value of an image, and *Dist*: the observed distance between the subject and the camera, stability measure of the eye-detection component, which we have presented in Könn et al. (2007). Both are - more or less directly - derived from the image data, and have been shown to be effective for confidence estimation.

Since the output of a classifier is an n-best list of hypotheses, another idea is to use features that exclusively exploit the distribution of the n-best list. Such features are the entropy of the n-best list (*Ent*), the difference between the two highest scores of the n-best list (*Diff0*), as well as two more fine grained difference-based features (*Diff1*, *Diff2*). The four n-best list-based features are suitable as confidence features, because they are directly related to the structure of the n-best list, i.e. distance of hypotheses, and reflect the probability of confusion. They are calculated as follows:

$$\begin{aligned} Ent &= -\sum_{i=1}^{n} k_i \cdot log_2(k_i) && (4.3) \\ Diff0 &= k_1 - k_2 && (4.4) \\ Diff1 &= \sum_{i=1}^{n} \frac{k_i - k_{i+1}}{i} && (4.5) \\ Diff2 &= \sum_{i=1}^{n} \frac{k_i - k_{i+1}}{e^{i-1}} && (4.6) \end{aligned}$$

where k_i denotes the score of the i-th best hypothesis, and N denotes the length of the n-best list. It can be seen that the two confidence features *Diff1* and *Diff2* are closely related, and their values are not statistically independent.

Additional confidence features, which make only sense for sequence hypotheses, are agreement (*Agre*) and stability (*Stab*), again introduced in Könn et al. (2007). *Agre* denotes the number of single image hypotheses, which are equal to the best hypothesis, divided by the total number of accounted hypotheses (corresponding to the sliding window size). *Stab* denotes the number of hypothesis changes relative to the total number of accounted hypotheses within the sequence.

4.2.5 Experimental Data Corpus

For evaluation of the approach, data was collected during dialog experiments in the receptionist robot scenario, including audio data and video data for multimodal person identification, with the tracking library Arthur[1]. Earlier experiments, which compare the confidence-based fusion approach for face ID against other video-based fusion techniques have been presented in Könn et al. (2007), with an earlier version of the system.

The data has been recorded from a dialog corpus of 38 subjects in 85 sessions. It comprises a collection of single images (recorded at 8 to 15 frames per second) and single audio utterances. The length of recorded sessions varies depending on the dialog length. A session on average contains 1019 single images with 378 face detections and 14 single turns, with a total audio length of 14 seconds. From the perspective of our proposed multimodal system a session on average generated 378 single image hypotheses, 377 image squence

[1] http://isl.ira.uka.de/˜nickel/arthur/

hypotheses, 14 concatenated turns hypotheses, 13 audio sequence hypotheses and 390 multimodal hypotheses. Due to practical reasons, additional audio data was recorded for 11 speakers, as the majority of audio segments in the initial recordings where 'yes' and 'no' utterances, and the voice identifier was not designed for one-word speech utterances. Training of person-specific voice models was conducted with 10 seconds of audio, which is the amount of audio data that can be recorded during one or two sessions with average length.

As for a clean evaluation, independent data sets for training and evaluation of the different layers must be used, the data was split into 5 different sets, which are obtained by dispersing the video and audio sessions among the sets. Set 1 was used for face ID and voice ID training, where the face ID was trained on 25 persons and the voice ID was trained on 8 persons. All other sets are used as evaluation data for face ID and voice ID and exist as closed set versions and open set versions. A closed set version contains only sessions where all subjects are also contained within set 1 and thus belong to the training set. An open set version contains all sessions from the closed set version plus further sessions with 'unknown' persons. For training and evaluating of logit-coefficients, we used the open set versions, since the recognition rate is fairly high and we wanted our system to cope with unknown persons as well. Set 2 is used to train logit-coefficients for single image face ID and voice ID. Set 3 is used to train logit-coefficients of sequence hypotheses, set 4 is used to train logit-coefficients of multimodal person ID and set 5 finally is used to evaluate person ID classification on unseen data.

4.2.6 Selection of Confidence Features

To be able to provide confidences within each layer of the system, suitable confidence features must be selected and logit-coefficients are calculated for each subsystem.

The confidence classifier of the single image layer is obtained by training logit-coefficients for different combinations of the confidence features (*Diff0, Diff1, Diff2, Ent, Image* and *Dist*), which are computed on the open set training data of set 2. Figure 4.4 shows a detailed section of the corresponding ROC graph. ROC graphs are discussed by Fawcett (2003) in detail as a good way to compare classifiers which can be evaluated with true positive and false positive rates.

Most confidence features are clustered within the same true positive/false positive area[2], with a rather low false positive rate (<0.1) and a rather high true positive rate (>0.8), except pure Entropy which has a true positive

[2] TP: true positive, FP: false positive

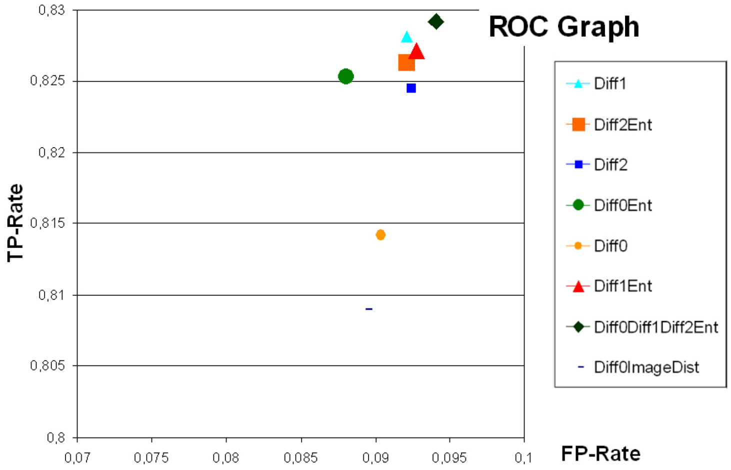


Figure 4.4: ROC graph showing different confidence features for single image hypotheses (face ID), trained and evaluated on open set data

rate of 0.71. Among the minor differences between the feature combinations, *Diff0Ent* produces the lowest false positive rate, while being only slightly worse than the best feature combination regarding true positive rate, and thus was used in the final setup. Combinations of different *Diff* features, e.g. *Diff0Diff1Diff2Ent* are problematic for logistic regression, as they are not statistically independent. The combination of *Diff0* and *Ent* also shows slightly better results than the single *Diff0* (which was used for the CRCM approach in Ekenel and Jin (2006)). The features used in our previous experiments, presented in Könn et al. (2007), could not fully be transferred to this approach since tracking and face identification methods differ.

The confidence classifier of the image sequence hypotheses is obtained by training logit-coefficients for different combinations of the confidence features (*Agre, Stab, Diff0, Diff1, Diff2, Ent*) which are computed on the open set training data of set 3. Figure 4.5 shows a detailed section of the corresponding ROC graph. In this comparison those feature combinations, which took *Agre* and *Stab* into account performed best. Before deciding on the best feature combination one has to consider different sequence lengths which is an important aspect of the online system. While it is obvious that with increasing sequence length, the quality of the hypotheses increases, this is not necessarily true for confidence classification. We have calculated possible confidence feature combinations according to the sequence length of 4, 15, 50, 100 and 200, whereof 200 was used in the evaluation. Figure 4.6 shows the development of the four best feature combinations over these sequence lengths. On the given data, *AgreStabDiff0Ent* shows the highest stability

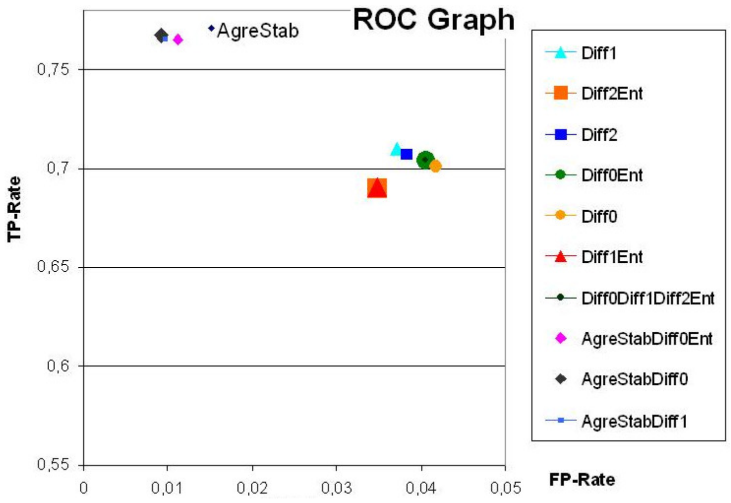


Figure 4.5: ROC graph showing different confidence features for sequence image hypotheses (sequence face ID), trained and evaluated on open set data

concerning different sequence lengths.

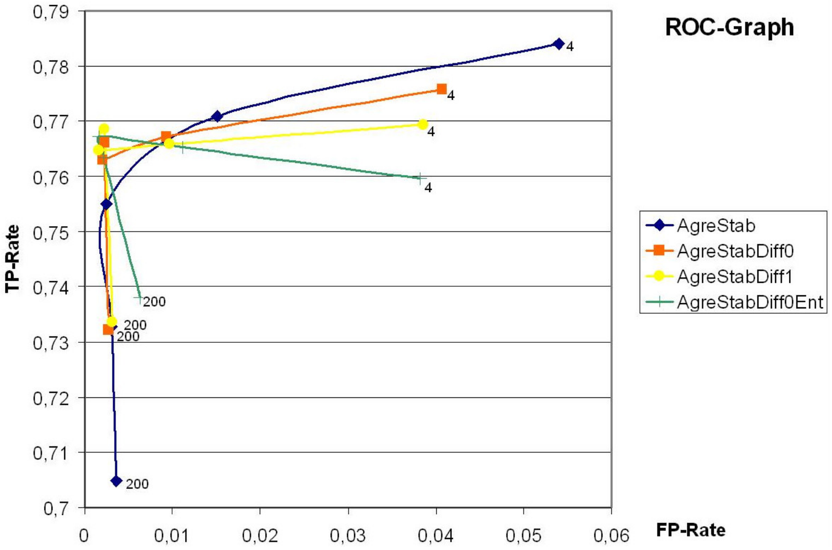


Figure 4.6: ROC graph of sequence face ID with different sequence lengths

To evaluate the result of the confidence estimation, figure 4.7 shows a plot of the recognition rate per confidence bin. The confidence bins are denoted over the y-axis with the intervals 0-5%, 5-15%, ..., 85-95%, 95-100%. The confidence classification rates of each bin are denoted over the x-axis. Both plots, single image confidence and sequence face ID confidence, are relatively close to the linear regression line of the recognition rates per bin. The plot shows that the actual correct recognition rate is slightly better than the estimated confidence, i.e. the plot points are below the diagonal line from 0,0 to 1,1. This is a desirable effect, as this way, the dialog system does not over-estimate incorrect hypotheses.

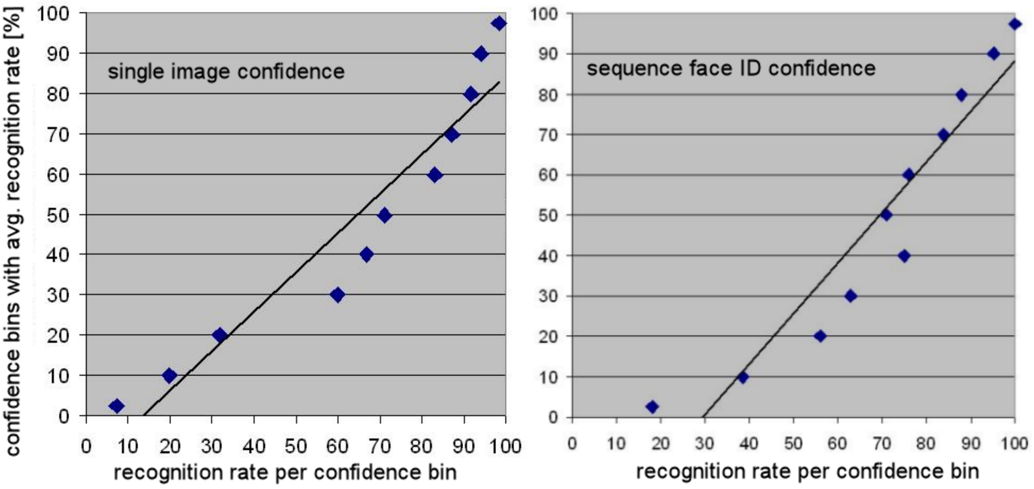
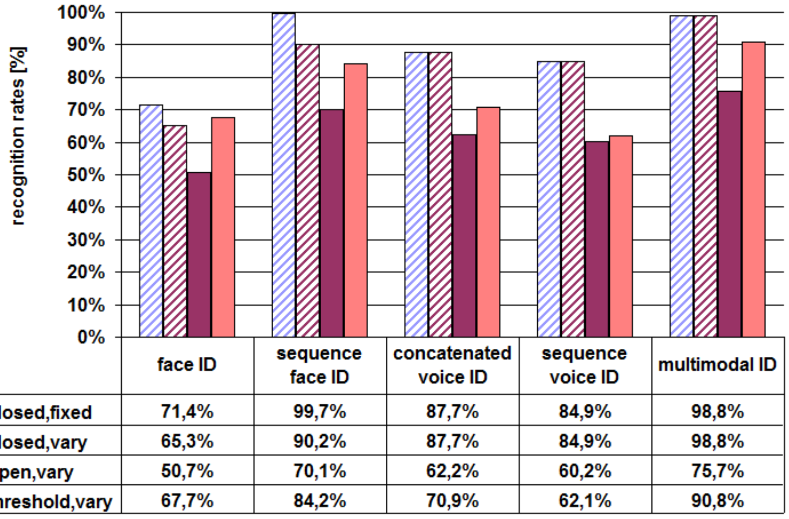


Figure 4.7: Confidence classification results per confidence bin

In the following, the results of the system and its subsystems are subsumed based on the evaluations of set 5. Figure 4.8 shows the recognition rates for the different layers and different subsets. Again, the set was evaluated with open and closed set conditions. Additionally, it was distinguished between different recording positions in the lab. The 'fixed' recording position has mostly light from the side, but varying light conditions during the day. The 'vary' data set contain additional data recorded at a second position, where the general direction of light was from the front for testing of robustness. The category 'threshold,vary' additionally conducts unknown person detection by assigning the unknown category to each hypothesis with a confidence below 30%.

The figure shows significant improvements at several layers for closed set person identification (1st and 2nd bars). Classification of the sequence face ID benefits from competing hypotheses, which in case of false classifications are spread in the Nearest-Neighbor feature space. In case of voice ID, the GMM-based classifier tends to produce similar (incorrect) hypotheses. This



	face ID	sequence face ID	concatenated voice ID	sequence voice ID	multimodal ID
▨ closed,fixed	71,4%	99,7%	87,7%	84,9%	98,8%
▨ closed,vary	65,3%	90,2%	87,7%	84,9%	98,8%
■ open,vary	50,7%	70,1%	62,2%	60,2%	75,7%
■ threshold,vary	67,7%	84,2%	70,9%	62,1%	90,8%

Figure 4.8: Recognition rates of the multimodal identification layers with closed and open set conditions

can also be seen by the best confidence feature which is *Diff0* and does not include *Agre* and *Stab*. Thus, sequence voice ID does not produce better hypotheses than the concatenated voice ID, but produces better confidence estimation, which is of great importance to the multimodal fusion.

Improvement of the multimodal ID can be seen best on set 'closed,vary', which has been recorded at more difficult conditions for face identification. Detailed numbers are shown in table 4.1. On average, the confidences distinguish between correct and incorrect classifications. An exception is seqFaceID, were too few incorrect hypotheses have been seen during confidence training (>99% correct). It can also be seen that unknown persons receive very low confidences, which suggests that unknown classification is possible. The challenge here is to distinguish unknown form incorrect recognition, which can be addressed e.g. by calculating average confidences over a sequence of hypotheses and then applying a threshold. The numbers in figure 4.8 have been computed with an optimal threshold of 0.3 for the multimodal ID. At this threshold level, 70% unknown was detected correctly, and <0.8% new errors (known vs. unknown) are made.

4.2.7 Conclusion and Discussion

The presented approach for multimodal ID fusion uses confidence measures on several layers. Different features and feature combinations have been

face ID *(Diff0Ent)*			
	number	mean	std. deviation
Set ‘closed,vary’ hypothesis true	4540	0.487	0.36
Set ‘closed,vary’ hypothesis false	2416	0.154	0.212
Unknown	1993	0.119	0.166
sequence face ID *(AgreStabDiff0Ent)*			
	number	mean	std. deviation
Set ‘closed,vary’ hypothesis true	6261	0.491	0.396
Set ‘closed,vary’ hypothesis false	684	0.432	0.254
Unknown	1988	0.029	0.046
voice ID *(Diff0)*			
	number	mean	std. deviation
Set ‘closed,vary’ hypothesis true	214	0.452	0.164
Set ‘closed,vary’ hypothesis false	30	0.336	0.073
Unknown	100	0.315	0.049
sequence voice ID *(Diff0)*			
	number	mean	std. deviation
Set ‘closed,vary’ hypothesis true	197	0.601	0.284
Set ‘closed,vary’ hypothesis false	35	0.167	0.145
Unknown	95	0.265	0.196
multimodal ID *(AgreStabDiff0Ent)*			
	number	mean	std. deviation
Set ‘closed,vary’ hypothesis true	7006	0.804	0.32
Set ‘closed,vary’ hypothesis false	171	0.326	0.452
Unknown	2083	0.094	0.274

Table 4.1: Overview of different confidence classifiers

investigated regarding their suitability for probability estimation with logistic regression.

All confidence classifiers make use of distributions of the n-best hypothesis lists. A major benefit of such features is that they can solely be computed on classifier output, without using ‘internal’ information. The same is true for the features *Agre* and *Stab* which cover sequence characteristics. As the experimental results show, the confidence-based fusion approach significantly improves the overall recognition rate.

All confidence approaches generally perform better for face ID than for voice ID on the given data. On the observed task many short utterances existed in the system which does not seem to contain enough discriminative information for robust voice identification. When longer sentences were recorded, voice identification provided better classification results and showed significant improvement in the multimodal fusion.

Together with multimodal hypotheses, on the highest layer, confidences are calculated that can be passed on to other dialog system components. The interface to the other components is an n-best list with confidences, independent of how many samples have been recorded or which modalities are available. Looking at a sequence of those confidences furthermore allows us to reliably detect unknown persons, even though a single incorrectly classified hypothesis may have a low confidence value as well. These confidence measures of the highest level are used in the following section for user ID modeling on a dialog level.

4.3 Bayes-Approach to Multimodal User ID in Dialog

Modeling the user's ID is a typical classification problem, given a set of input parameters and a classification result. In addition to work presented in the previous section on multimodal fusion of ID classifiers, additional attributes can be integrated for identification. For example while the first name of a person does not solely identify one person it contributes to the identification process, as can do many other features. Even the time of the day when a person talks to the robot can contribute to the probability of meeting a specific person. All these aspects can be interpreted in form of Bayesian probability theory, by modeling probabilities of observations given a true state of nature. To integrate these different aspects using Bayesian theory, a formalism exists which is known as Bayesian networks or belief networks. In fact the approach presented here benefits from work presented in the previous section, which did not only introduce multimodal fusion, but also confidence estimation for classification output. Such confidence measures for ID classification have proved valuable for estimating probabilities of correctness which allows a probabilistic integration in the Bayesian model, as will be shown in the following.

A typical affordance in an interactive dialog system is that hypotheses are computed during runtime of the system, i.e. while the dialog continues, instead of collecting all relevant information and only then applying classification. In fact, the dialog flow is also influenced by the decisions that are made by the ID classifier, leading to better results and shorter dialogs, if the model can generate good hypotheses early in the dialog. The approach proposed here has been examined for human-robot interaction, where the users engage in explicit identification dialogs. For example, the robot can ask a user for the name, or use explicit or implicit confirmation strategies given different kind of observations. Perceptual technology used for the experiments is based on sensors typically used on a humanoid robot, such as stereo vision and speech recognition. For simplicity we frequently use the term user model in the following, which in this paper refers to modeling the user's ID.

In this section we outline our approach for a user model that combines information collected during dialog, such as spoken names, spelled names, confirmations and multimodal ID classification from face ID and voice ID. Fusion of these modalities is done using Bayesian (belief) networks. A key aspect is estimating conditional probabilities, such as confidence measures for multimodal ID hypotheses. Generally speaking, these confidence measures are necessary to cope with recognition errors. For example, if a first name has been misrecognized and contradicts the multimodal ID hypothesis, the system computes the best hypothesis while taking into account probabilities of misrecognition of each input hypothesis. It can then ignore the incorrect speech recognition if multimodal ID confidence is high enough.

Bayesian networks are frequently used in data mining, to discover statistical dependencies on large data sets, with the goal of learning network structures. In our work, the network structure is created manually and different network structures are analyzed. The following chapter describes some basic properties, a more detailed overview can be found for example in Heckerman (1996). In Huang et al. (2000) a belief network has been used for multimodal user registration. It is similar to the 'simple' network structure presented in the following and compared to more complex structures.

In contrast to other work our approach takes into account unknown persons, unknown word detection plus name spelling, and features extracted from the dialog history. Special attention is given to confidence estimates. The presented approach can generally be extended with other features. For example on could consider day of time when a person interacts with the system and integrate this as a conditional probability. The approach also shows significant improvement over our previous identification model presented in Holzapfel and Waibel (2007).

4.4 Belief Networks for Person Identification

A Bayesian network is a directed acyclic graph with nodes and edges. Each node represents a variable which is either discrete or continuous, and edges are modeled as conditional probabilities. Depending on the type of variables the network is either a discrete, continuous, or a hybrid Bayesian network. In the presented work, we use a discrete network. When some variables are observed (they are then called evidence variables) other variables in the network can be queried using probabilistic inference.

4.4.1 Input Features and Network Structure

In our network we use three categories of observations as evidence for identification. Evidence corresponds to information slots filled by the dialog system. The first observation category is multimodal ID (MMID) classification which directly classifies the person's ID. The second type of observation only provides hints about the person's ID but does not classify one person exclusively. Such observation is recognition of the spoken first name. In our model, a person can have only one first name; however, different persons (either known or unknown) may share the same first name. The third type of observation is extracted from the dialog flow, such as disconfirmed names.

These types of observation can be modeled in the belief network as the following discussion shows. The structure of the belief network is determined by the definition of conditional probabilities. A standard ID classifier produces hypotheses with posterior probabilities, i.e. $P(ID|observation)$. Another way of modeling, which also describes a causal structure, is the inverted dependency structure $P(classification\text{-}correct|ID)$ with a directed edge from 'ID' to 'classification-correct'. Now, the conditional probability models the probability of a classification being correct given the ID. This probability is estimated by independent confidence classification which is multiplied by the n-best list hypothesis score with successive normalization. Confidence estimation for multimodal ID is described in section 4.4.3. Using Bayes theory to combine different classification results leads to some practical issues with the extreme values 0 and 1. This is the case, e.g. when IDs are not represented in the n-best list, thus additional factors and an offset are introduced with the following formula:

$$w_{id} = m + conf * a(2 * score_{id} - 1) \tag{4.7}$$

The values m (offset), the confidence of the classifier, and the scaling factor a influence the rating of the original ID-score from the hypothesis list. The desired probability is then obtained by normalizing w_{id} by the sum over all w_{id}.

In a similar way, spoken name recognition (speech recognition results) is integrated into the network as evidence. A conditional probability $P(name\text{-}correct|name)$ models first name and last name recognition. An additional edge $P(name|ID)$ connects names to IDs. It is set to 1 for persons that have been entered manually and can be set to a smaller value to model uncertainty in the knowledge base when a person has been learned interactively.

A fourth type of information is not used as evidence in the network, but influences conditional probabilities in the network. For example confirmation of a name is a feature that is observed by the dialog model. In this case the evidence, i.e. the value of the observed name, does not change, but the probability of the name being correct increases.

4.4.2 Network Structure

Figure 4.9 shows the structure of a simple belief network integrating multimodal ID, first name and last name recognition. An abstract ID node represents the ID of a person; other nodes represent evidence as pointed out above.

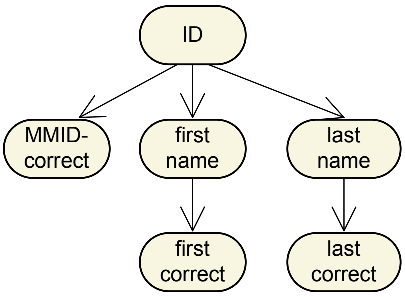


Figure 4.9: Simple user ID belief network

The simple network structure works well for many situations with only known persons. However, some important aspects are missing. For example the network does not model the reduced probability of a name after it has been rejected. For rejected (disconfirmed) names a separate blacklist is added. It accounts for the fact that also rejections are error prone. A name on the blacklist is assigned 1/10 of the probability of non-rejected names. Also the problem of unknown first name / last name combinations is addressed. An unknown detection node increases the likelihood of an unknown person by 100 * prior user probability if first name / last name combinations are observed that don't match the database of known persons. The factor 100 has been chosen experimentally to 'compete' against multimodal user ID. Figure 4.10 shows the extended network structure.

Some considerations had to be made so that the proposed user ID model can be used in an online system. The main considerations relate to dynamic updates in the network. Some parts of the network are static, which are the structure of the network and the node names. Node values, i.e. person IDs, first names, last names, etc. are generated automatically from database entries. Edges in the network, i.e. conditional properties, are also updated

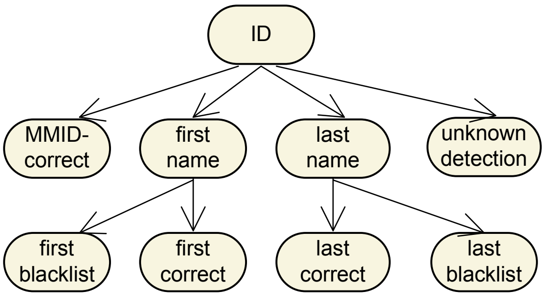


Figure 4.10: Extended user ID belief network with blacklist and unknown model

dynamically, e.g. the edge representing correct recognition of multimodal ID is updated with each new hypothesis according to the confidence value.

4.4.3 Multimodal ID and Confidence Measures

As mentioned before, confidences are very helpful in theory to estimate a 'trust-level' of a classifier, especially when hypotheses from different classifiers are combined. In Grosse et al. (2008) we have proposed an approach for confidence-based fusion of face ID and voice ID, which uses logistic regression for confidence estimation on several levels (on single hypothesis, sequence hypothesis and output confidence). We use this approach to model the multimodal ID ('MMID') node in the belief network, but restrict ourselves to using video information, leaving out voice segments. The reason for this is that even though the recognition rates are better with voice information, no sufficient voice data was available for independent training and evaluation. The belief network uses the hypothesis score (from n-best list) plus the classification confidences, as described in the beginning of this section.

In the experiments reported here, this approach has been used in two configurations. The first configuration is closed-set person identification, where the MMID classifier always decides on a label known from training data. The second configuration is open-set person identification, where there is an additional category 'unknown' to classify persons which are not in the training set. To integrate the unknown classification in the n-best list, we estimate the hypothesis score by 1.0 minus classifier confidence, which produces stable results on the given corpus.

4.5 Experiments

4.5.1 Dialog Data Collection

Data used for experiments has been collected during different robot receptionist dialogs with user ID and name learning (Holzapfel and Waibel, 2007). The dialog manager uses different strategies (a fixed strategy was employed per scenario) to identify a person's ID, first name and last name. During the dialogs, speech and image sequences have been recorded for voice ID and face ID, speech recognition results have been logged and all interactions have been transcribed. From this data we obtain a corpus of annotated sessions, with a timeline of events including all dialog system input.

With this data we have then conducted the evaluation of different user models. The advantage of the approach is that once data has been collected, different user models can be compared on the same data. While there is an effect of the applied user model on the dialog flow, recorded information can be observed by all models. Thus, a user model that has not been used for recording will be slightly underestimated. The best comparison can be drawn by the end of an evaluated dialog session.

4.5.2 Baseline Approach

The baseline or 'confirmation' approach uses a rule-based system and a confirmation strategy to determine the ID. It uses three slots: *MMID*, *firstName*, *lastName* with different slot states, and the output slot *userID*. *userID* and *MMID* have the states EMPTY, SET, CONFIRMED. *firstName* and *lastName* have the states EMPTY, UNKNOWN, SPOKEN, SPELLED, CONFIRMED. A slot is set to EMPTY when the information slot is empty or when its value has been disconfirmed. The update rule for setting the *userID* value takes into account reliability of the slot values. For example CONFIRMED has the highest reliability, and spoken name input is preferred over multimodal ID. The latter only is considered when first name and last name slots are empty, which happens typically at the beginning of a dialog or after a name has been rejected.

4.5.3 Evaluation

The evaluation compares different configurations of the Bayesian networks with each other and against the confirmation approach. The evaluation has been conducted on two conditions: person is known (i.e. has talked to the robot before and MMID training data is available) vs. person unknown (no training data available). Each condition is evaluated with open set vs. closed set MMID classification, each with a database of 25 known persons. In the unknown condition, 46 sessions are available for evaluation; in the known condition 43 sessions are available. Before evaluation, MMID models have been trained with independent training data for the two conditions: known and unknown. The set for the unknown condition includes all sessions from the known condition, however the respective person was excluded from the MMID training data and the person database.

To evaluate the approach, different metrics are used. 'UID rows' is the percentage of all correct hypotheses, i.e. all input event during the interaction. 'UID end' is the percentage of all correct final hypotheses, i.e. the last hypothesis of each dialog. 'UID norm' is a normalized correct rate, i.e. the average correct rate per dialog. It prevents that long sessions get higher weight than short sessions. For example, the shortest sessions without face ID input has only 6 input events, in contrast to the longest session with 250 input events.

set / condition	events	MMID	sessions	MMID end
closed / known	1870	84,44%	43	81,4%
closed / unk	2165	0,00%	46	0,0%
open / known	1870	80,53%	43	74,4%
open / unk	2165	89,01%	46	89,1%

Table 4.2: Task overview: number of input events, MMID per event, number of sessions, MMID at end of the session

Table 4.2 reports recognition rates of the multimodal ID classifier representing the observations of the MMID node. Table 4.3 shows the numbers from the evaluation runs with closed set and open set classification. The user models listed in the tables are the baseline *confirm* model, the simple Bayesian model *bayes-p* without black list, the *bayes-bnr* model with black list but without resetting of user names after disconfirm, the *bayes-blu* model including black list and unknown person detection, and the *bayes-bl* model with black list and resetting of names. The unknown/closed set has been excluded since models are not suitable for this category, only *bayes-blu* and *bayes-bl* achieve 100% for UID end, the others achieve 0.0%.

The overall best model is the *bayes-bl* model, which outperforms the *bayes-blu* model in the open-set condition, where the unknown detection from face ID is more reliable than unknown detection from name recognition and static properties of the *bayes-bl* network. In the closed-set condition both are almost equal. In the closed set/known condition the simple model obviously performs best, since it does not produce false alarms for unknown.

The ‘UID-norm’ value looks worse than ‘UID rows’ for most conditions. This is reasonable since some sessions don’t have any face ID at all. These sessions start without relevant information and only at the end of a session a good hypothesis can be found by the model. In general this also mirrors the fact that user ID hypothesis improves with the dialog flow.

Given the kind of evaluation with a static dialog corpus, the effect of the user model on the dialog flow cannot be measured. Despite the fact that the dialogs had been recorded with the baseline user model, the results show that the belief network operates more reliably than the baseline model. The recognition rates are better especially at the end of the dialog. This significantly improves the robot’s perception who the robot is talking to and improves memorizing persons.

condition	task	UID rows	UID end	UID norm
known/c	confirm	85.94%	83.7%	74.77%
known/c	bayes-p	83.74%	95.4%	77.90%
known/c	bayes-bnr	75.13%	93.0%	76.52%
known/c	bayes-blu	73.32%	93.0%	75.13%
known/c	bayes-bl	79.68%	93.0%	76.41%
known/o	confirm	77.38%	72.1%	67.19%
known/o	bayes-p	80.80%	95.4%	73.54%
known/o	bayes-bnr	72.19%	93.0%	72.16%
known/o	bayes-blu	69.84%	90.7%	70.59%
known/o	bayes-bl	76.58%	93.0%	71.94%
unk/o	confirm	91.45%	100.0%	88.20%
unk/o	bayes-p	86.00%	58.7%	83.53%
unk/o	bayes-bnr	86.33%	60.9%	83.59%
unk/o	bayes-blu	91.50%	100.0%	88.11%
unk/o	bayes-bl	91.50%	100.0%	88.11%

Table 4.3: User ID evaluation with closed set ‘/c’ and open set ‘/o’ multimodal ID

4.5.4 Conclusion

The presented approach to user identification in dialog considers aspects of an online system where information is delivered and updated sequentially.

The approach also considers special aspects of a dialog system where information is confirmed or rejected during dialog. In this aspect it extends a pure multimodal ID approach. It is also suitable for open set person identification.

Different belief network structures were compared with each other and against a baseline model that purely relies on dialog information with confirmation and rejection of the best hypothesis. The results show that the best configuration depends on the task to be fulfilled. Especially, the selection of the best architecture depends on the expected number of unknown persons, i.e. if the classifier works in open set vs. closed set mode. In any case, the best configurations perform better than the baseline model and are suitable for person identification in dialog. These results also confirm results from the evaluation of confidence-based multimodal ID.

Chapter 5

Reinforcement Learning for Person ID in Dialog

5.1 Introduction

5.1.1 Overview and Problem Definition

This chapter addresses optimization techniques for person identification and name learning dialogs using reinforcement learning. Writing error tolerant and robust dialog strategies generally is a tedious and costly effort. In recent years, reinforcement learning has successfully been applied for approaching this task by machine learning techniques instead of writing dialog rules manually. As this technology has produced promising results, it seems promising to apply this technique also for the task of identifying persons, where the dialog system has to cope with different recognition errors from speech recognition and multimodal perception, and where optimal strategies are to be found to achieve the dialog goal despite possible recognition errors. New challenges arise when adapting this technique to a multimodal system with the task of person identification including integration of different modalities, multimodal user simulation, and multimodal error models.

These challenges are addressed in the following sections, where two experiments are presented. The first experiment, which has already been published in Holzapfel and Waibel (2008a), introduces the reinforcement learning framework with a multimodal user simulation. It is extended in the second experiment with a multimodal user ID model introduced in chapter 4. The first experiment presents evaluations of the approach in simulation and in a real user experiment. The results show that the success of the strategy strongly depends on success of name recognition.

In the presented system, the reinforcement learning agent does not learn the strategy for the complete and complex dialog system. Rather, it is restricted to the person identification task, to learn the strategy of a single dialog module, which can be solved efficiently. As we have demonstrated in Holzapfel and Waibel (2007), such a modular architecture can combine handcrafted modules and modules trained with reinforcement learning to combine the advantages of both worlds. The modular dialog design will be introduced in this thesis in more detail in chapter 6.5. Within the restricted dialog module with mostly system initiative, user reactions are adequately modeled by the simulation, as we have already shown in Prommer et al. (2006), with an interaction task for a barkeeper robot. In Prommer et al. (2006) we have also evaluated the system with real users and obtained comparable results for evaluation in simulation. As the user simulation allows to run a large number of dialogs (e.g. 100,000 dialogs have been used in this thesis for evaluation of each strategy), it averages out speech recognition problems for spoken names which have the largest impact on dialog success when learning unknown names, and therefore represents a applicable evaluation method for this task.

5.1.2 Scenario and Dialog Setup

The scenario for the dialog manager is to control interaction in a robot receptionist scenario. The full robot receptionist is described in more detail in chapter 8. For now, we want to concentrate on the aspect of person identification.

The scenario of person identification was addressed in a series of experiments. A first experiment was conducted as a Wizard-of-Oz experiment, where the robot acted as a parcel receptionist, where one part of the dialog was to identify the person (more details on the experiment are given in section 8.3.2). The Wizard-of-Oz experiment served as a data collection and analysis of the dialog task. From this analysis, the receptionist task was decomposed into the dialog modules greeting, parcel reception, name learning, directions and goodbye. In further experiments, the Wizard was first replaced by a handcrafted dialog strategy. Further experiments including experiments with reinforcement learning were conducted with person identification only or in combination with social network modeling (chapter 9).

The modularization of the dialog, as motivated in the introduction of this chapter, allows reinforcement learning to focus only on the strategy of a single module. Combination of different modules into a complex system is presented in chapter 6.

5.2 Reinforcement Learning Setup

This section describes a reinforcement learning approach to automatically acquire a dialog strategy which is optimal with regard to a predefined metric, the reward function. The design of single modules separates concerns and allows training of the name learning module, which can be conducted in reasonable training time and in isolation of other dialog concerns.

One promising approach for optimization of dialog strategies in general is with reinforcement learning. The idea of reinforcement learning is that one cannot define correct and incorrect actions for each state as in supervised learning, but rather to expose the system (usually referred to as the agent) to an environment in which it can take a series of actions, where each action is associated with some reward. The agent is supposed to learn from these observations and optimize its expected reward. Reinforcement learning in this definition is a class of learning problems. Various systems exist that apply reinforcement learning, and several algorithms exist to solve the reinforcement learning problem, see Sutton and Barto (1998). One problem of applying this technique to dialog systems is the large number of data (i.e. dialogs) required for training of the system. This challenge is nowadays addresses by training a user simulation and to simulate a large number of dialog interactions for training.

In Prommer et al. (2006) we have argued for a standardized process model for training a dialog strategy with reinforcement learning. Figure 5.1 shows the process model. In step 1, a Wizard-of-Oz experiment is conducted to analyze the task and collect initial data for statistical model training. In step 2, a user simulation is created with statistical simulation of user actions and error models for system components. In step 3, the state model (states of the Markov decision process - MDP) is defined. In step 4, the dialog strategy is trained in the simulation. In step 5, the dialog strategy is integrated in the online system and experiments with real users are conducted. As the initial Wizard-of-Oz experiment provides only a small amount of data for statistical models, additional data is collected from the online interactions (step 6), to update the statistical models of the simulation.

5.2.1 MDP State Model

One important aspect of defining a model for reinforcement learning is the state model. In theory it has been shown that if the state model fulfills the Markovian Property, Q-Learning converges to the optimal policy. The Markovian Property requires that the state transition probability only depends on the current state, which is composed of system state s_t and the system action a_t with the discrete time index t.:

$$P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, s_{t-1}, ..., s_0, a_t, a_{t-1}, ..., a_0)$$

In practicable applications however, and especially in dialog, this property usually does not hold, but still good policies can be found. A trade-off needs

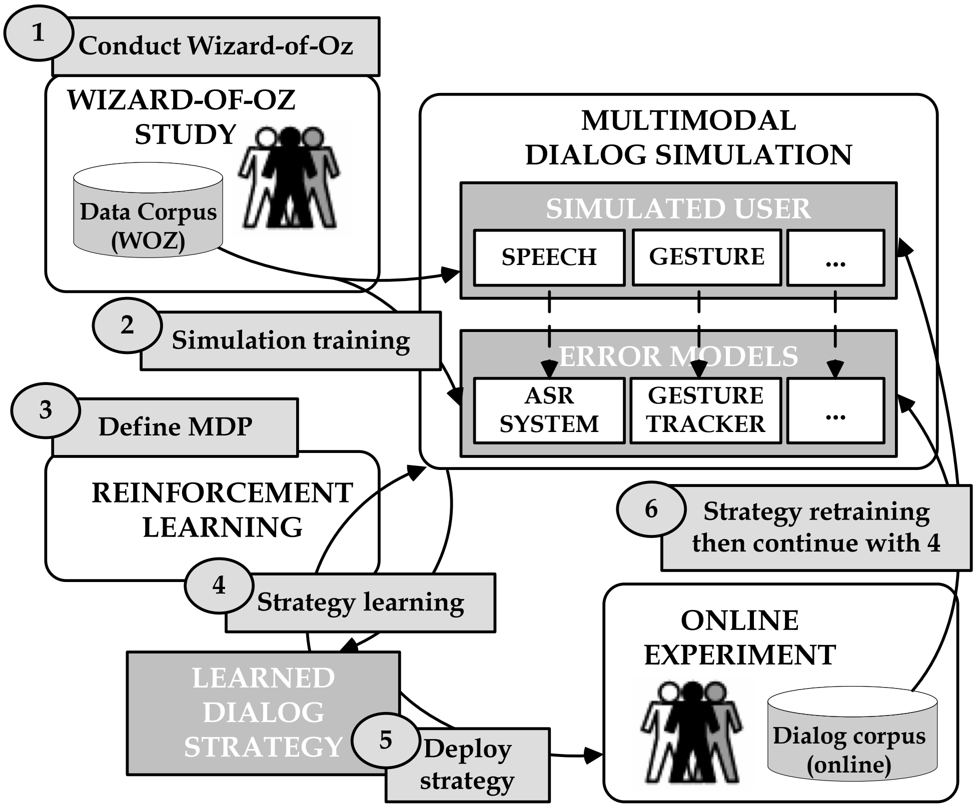


Figure 5.1: Process model for reinforcement learning in dialog systems with multimodal user simulation

to be found between encoding fine grained information and history versus simple models, to find a model which can be computed with the given data and in a reasonable amount of training runs.

The MDP state model (s_t) used in our experiments encodes information about the information state of semantic slots plus information about the progress of the dialog. The dialog manager uses the information slots with associated MDP state values as shown in table 5.1. States representing the progress of the dialog are shown in table 5.2. The actions (a_t) which are available for the dialog strategy are the same for both the handcrafted as well as the learned strategies. The available actions are listed in table 5.3.

Important for the learned strategy is the chosen reward function. It defines which dialogs are 'good'. In our scenario, learning correct names is rewarded (+10), learning wrong names is punished (-10). From experiments we found that some persons accept names which are almost correct. We try

Information Slot	MDP state	MDP state values
ASR name input	Name-ASR	empty, filled, oov
Spelling input	Name-Spelling	empty, filled
Voice ID	VoiceID	empty, filled
	VoiceIDConf	low, medium, high
Face ID	FaceID	empty, filled
	FaceIDConf	low, medium, high

Table 5.1: Dialog information slots and mapping to MDP states

MDP state	values	description
nNameFailed	0,1,2+	number of failed attempts to confirm a name
nNameConf	0,1+	number of successful attempts to confirm a name
nASRNameFailed	0,1,2+	number of failed attempts to confirm a name from speech recognition
lastAction	*action*	name of the previous action

Table 5.2: Dialog state variables in the MDP state

to quantify this effect with the Levenshtein distance between learned and correct name (distance = 1 is rewarded +3; distance = 2 is rewarded 0). Each additional turn is punished with -1, so dialog length is kept moderate; repeating the same system action is punished with -0.5. Other functions can be chosen to increase the importance of different factors.

category	actions
get information	ask_name, ask_name-spelling
confirm	conf_name-asr, conf_name-spelling
	conf_faceID, conf_voiceID
finish dialog	accept-name, abort

Table 5.3: actions available for the reinforcement learning strategy

5.3 Multimodal User Simulation

To build a user simulation a common approach is to model user actions with statistics estimated on collected data. In addition to that, (error-) models are created that describe the behavior of the system's recognition components, i.e. a statistical model of errors. The idea behind this approach is that statistically describing user actions and error models is simpler than directly learning the system's strategy.

5.3.1 Multimodal User Models

Existing approaches for training a user model range from simple models, such as the bi-gram model, to more complex models (Eckert et al., 1997; Levin et al., 2000; Pietquin and Renals, 2002). In previous work (Prommer et al., 2006) we have achieved good results using a simple bi-gram model for statistics on a semantic level. In the addressed restricted task bi-gram statistics provided good estimations already on a small amount of training data, which would not suffice to train more complex models.

The quality of the bi-gram model $p = P(\mathsf{act}_{\mathsf{user}}|\mathsf{act}_{\mathsf{system}})$ highly depends on the defined abstraction granularity of simulated user and system actions and the task restriction. In our work we have adopted the general bi-gram model to a more fine-grained model of bi-grams over semantics of user actions (input speech act + semantic attributes) given the system's speech act. Statistics for user actions have been trained on a single dialog goal, i.e. name learning, in isolation of other dialog goals, from transcribed dialog interactions.

In addition to speech-only interaction our multimodal system models non-verbal information from voice ID and face ID. Voice ID, like speech input, is computed turn-wise for each spoken utterance. Inspired by recent work by Krsmanovic et al. (2006) who concatenate data from speech snippets to simulate data that is provided for voice ID during runtime, we adopt the approach and simulate recognition input by taking samples of real recorded data adapted to multimodal identification and real dialogs.

Face ID at first glance is not turn based. However, since face ID only updates the dialog state during a new turn, this is imitated in the simulation by grouping ten to twenty images for face recognition per turn. To produce a variety of hypothesis values, we use real images from one person taken during data collection at 2 fps, which are cut into sub-sequences of ten to twenty images. From these, the simulation environment randomly picks single sequences.

Problematic with this setting are the high computing requirements. Just considering face ID, given a database of roughly twenty persons, the face ID recognizer can process two to four images per second on a standard 3GHz Pentium processor. A minimum training requirement of 1 million dialogs then poses an impracticable computational burden. The biggest part in time consumption is to detect a face within an image and to produce a per-image ID classification using the nearest-neighbor classifier. Both problems can be pre-computed given a fixed database of known persons, when the state space of the classifier remains constant. The combination of pre-computed single-image hypotheses to a sequence hypothesis is much faster and can adequately be conducted during simulation. A similar approach using audio snippets has been applied to voice ID recognition. With these settings, the system runs a full dialog in simulation (including dialog state update, policy update and action selection) in 1.2 ms at 3.5 turns per dialog on average, which is roughly 0.3 ms per turn. Note that the chosen setup does not allow direct simulation of the effects of storing more and more persons in the database, but rather allows training of a strategy for a fixed database setting of known persons.

5.3.2 Error Models

Simulation of user actions is not sufficient to model the input for the dialog system. The missing link between user actions and dialog input is described by error models, which statistically simulate typical errors made by the recognition components. For example, the difference between experiments using close speech and distant speech is simply a different error model. Face ID as well as voice ID do not require additional error models since their errors are implicitly modeled by applying real classifiers (partly pre-computed) to simulated data. Speech, in contrast, is modeled statistically and requires additional error models for speech, phoneme, and spelling recognition. Spoken name input is modeled as a speech act with semantic parameters. For spoken name input, the speech act is *informName* with a semantic parameter *NAME.* The error model first statistically models concept confusion and deletion, i.e. probability for recognizing a wrong concept (confusion) and the probability for not understanding any concept at all (deletion). Secondly, statistics are applied to model confusion and deletion of the semantic parameter(s).

5.4 Simple Integration for First Name Identification in Dialog

5.4.1 Training and Evaluation in Simulation

Training of the strategy was conducted with the MDP state and action models described in section 5.2. Training was conducted with the Watkins-Q-lambda algorithm with exponential cooling of epsilon, and the learning rate alpha. All models were computed with all combinations of a list of 11 equally distributed lambda values from 0.0 to 1.0 and a list of 11 equally distributed discounting factors from 0.0 to 1.0. To test the effect of the number of training runs we experimented with different dialog numbers per training, using 1 million to 100 million dialogs per model. Reasonable training runs are 10 million (10M) dialogs and more, since training with 1 million (1M) runs still contains a couple of state-action pairs that have never been visited, especially in states that occur only seldom. There is still a significant difference between training sizes of 10M and 100M dialogs, so high numbers of training dialogs still means improvement of the dialog strategy. On the other hand, first training runs with 100M dialogs took 32 hours on a Pentium4 3 GHz processor. After a few code optimizations we could lower the training time to roughly 5 hours. Considering training time, all models which have been trained with different configurations, i.e. 121 configurations for all combinations of discount and lambda values per MDP state space, have been trained with 10M dialogs, single configurations have been trained with 100M dialogs for comparison of the best models. All evaluation numbers presented here have been obtained from running 100k dialogs in simulation, which have shown stable results, from which the average reward is computed.

Training and evaluation requires splitting the data into three parts. The first part is used to train user ID models for voice ID and face ID, and bi-gram statistics. A second data set is used as simulation data, for training of the dialog strategy, and a third set is used for evaluation of the dialog strategy. Depending on how the set is split, the dialog strategy training and evaluation sessions include more or less unknown persons. On a set with a large number of unknown persons, the resulting reward is lower than with a set restricted to known persons, since unknown persons are harder to recognize and to register.

5.4.2 Baseline Strategy

The implementation of the handcrafted strategy follows a simple pattern: Alternatively ask the visitor for his name or for the spelling of his name. If either of both is given, try to confirm the name. When speech recognition reports an unknown name ask for spelling. In the beginning, if face ID produces a hypothesis, try to confirm the associated name. If neither is set but voice ID is given try to confirm the associated name. Do not ask for the same name twice. As soon as the name is confirmed quit the dialog and store the name. If a predefined threshold of turns is reached e.g. 15 (in the simulation), or after 3 unsuccessful confirmation questions (in the online experiment), the dialog is aborted without storing the name.

5.4.3 Evaluation in Simulation

Figure 5.2 shows evaluation results for the close-speech condition of the baseline strategy in comparison to strategies trained with reinforcement learning (RL). The categories shown are ‘sim’ for the simulation set which was used for reinforcement learning, and ‘eval’ for the third held out data set. The model abbreviations are ‘H’ for the handcrafted model, ‘F’ for RL with face ID only, ‘V’ for RL with voice ID only and ‘M’ for RL with multimodal input (face ID + voice ID). Figure 5.3 shows the name-assignment errors made by the different strategies under close speech condition. An error is an incorrect assignment of a name at the end of a dialog. Almost correct names were counted separately, where the learned name differs by only one letter from the correct name. Both simulation sets included 25% unknown persons. For roughly six percent of all turns no voice ID information was available. On the remaining set, the recognition rates for voice ID are 59% and recognition rates for faceID are 68%.

5.4.4 Experiments with Users and Discussion

The results of the simulation show better performance of the reinforcement strategy than the handcrafted strategy. An interpretation is that the reinforcement learning approach learns more complex rules, when to confirm multimodal input, in combination with recognition confidence, dialog length, and failed name recognition. The charts show slight differences between the sets. The ‘M’ model (multimodal input) performs generally best, which matches our expectations, because it can choose among different modalities. All learned models have a higher number of correct dialogs, at a minimal cost of 0.1 turns more per dialog on average.

To evaluate the system with real users we have deployed the strategy to a small user study experiment. Since there are high variations in results depending on the user, the experiment can only be interpreted as a tendency

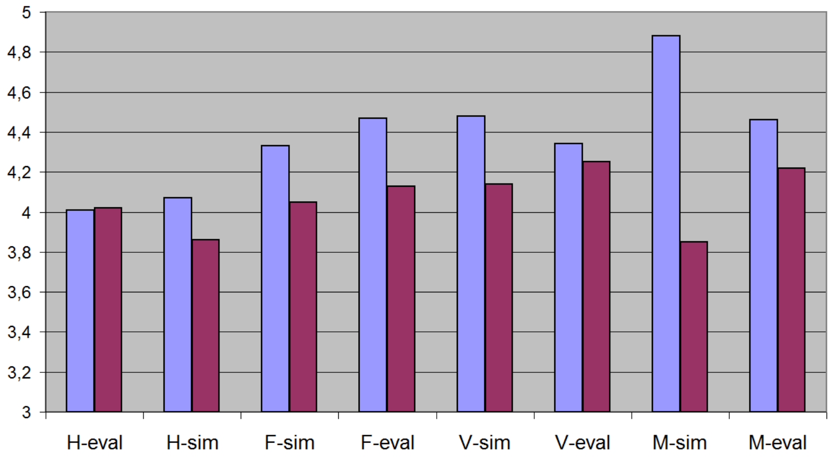


Figure 5.2: Evaluation scores for different strategies in the user simulation, showing reward (first column) plus turn numbers (second column).

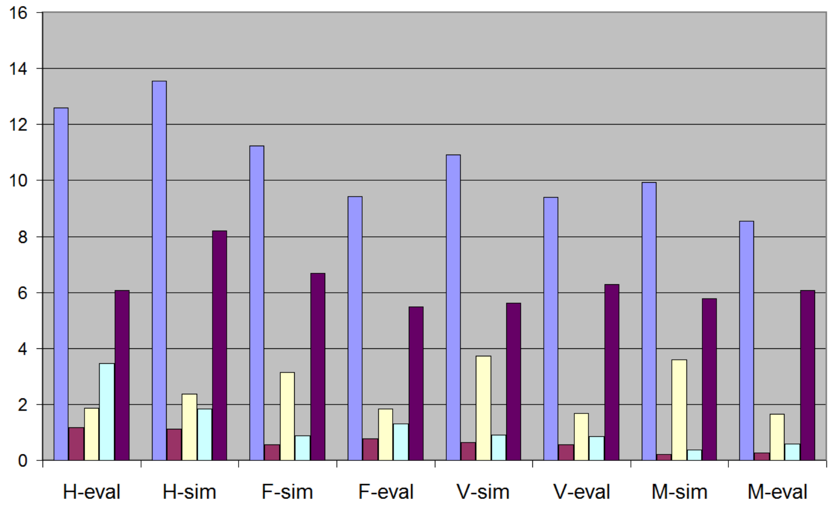


Figure 5.3: Incorrect assignments made by the different strategies in the simulation in % (out of 100k dialogs). The bars indicate: total errors, almost correct names for known persons, almost correct names for unknown persons, incorrect names for known persons, incorrect names for unknown persons.

but not be treated as a reliable comparison of strategies. Further and larger evaluations are necessary for this.

The experimental setup was the same as for the Wizard-of-Oz experiment with the same microphones and cameras. This time the users were told that the only purpose of the interaction was the name learning dialog. We con-

User: Hello *[act_hello] ; FaceID: Peter*
Sys : Am I right that your name is Peter?
User: No, my name is John. *[inform_name [NAME ”OOV”]]*
Sys : Please spell your name!
User: J O H N. *[inform_spelling [SPELLING ”JOHN”]]*
Sys : Have I understood correctly that your name is John?
User: Yes. *[yes]*

User: Hello *[act_hello] ; FaceID: John*
Sys : Am I right that your name is John?
User: Yes that’s right. *[yes]*

Table 5.4: Two example dialog interactions in the reinforcement learning scenario

ducted dialogs with 11 persons, 7 persons using distant-speech, some of them haven’t used speech recognition before, and 4 persons using close-speech. The experiments were conducted with two conditions: known and unknown. A person at first was unknown, and after completing the learning dialog, name and ID information was stored in the database to conduct further dialogs with the known condition. To obtain more dialogs with the unknown condition the ID and name information was removed from the database, to iteratively conduct dialogs in the unknown and known conditions.

Table 5.4 shows two sample dialogs (original names are replaced by ‘Peter’ and ‘John’). The first dialog shows a sample dialog trained with reinforcement learning. The second dialog shows a simple confirmation dialog with only one user turn to confirm the correct recognition.

The close-speech experiment produced similar results as evaluation in simulation. Average dialog length was 4.7 in the unknown condition and 2.6 in the known condition, excluding greeting and goodbye. The distant speech experiment produced worse results than predicted by the simulation. While first interactions with speech experts went very well, naive users had more problems to complete the dialogs. For example, users spelled their names too slowly, which was not handled correctly by automatic speech segmentation. After an introduction users could complete the task more easily. Additional errors were caused by spelling recognition performance which mostly was not 100% correct. All numbers from the experiment are shown in table 5.5. The problems are rather to be assigned to system conditions than to the dialog strategy. The numbers also show that the task to register an unknown person is much harder than identifying a known person. Unknown persons can neither be recognized by face ID or voice ID, or, if they could be recognized, but during previous interactions no name was stored, this cannot be communicated by the system. So currently the only way to get known by the system was by spelling one's name, which was easier to complete when used to the system.

	1	2	3	3	4	5	6	7
unknown	5,3,4	3,7,5	15,15,15	14,4,5	15,15,12	5,4	15,11	15,15,6
known	3,8,1	1,4,2	1,3		1,6		1,3	3

Table 5.5: Number of turns during distant speech dialogs. The columns mark subjects 1 to 7. 15 turns marks unsuccessful dialogs.

Recognizing an arbitrary name is a challenge for speech recognition, since there are too many names than can be kept in the recognition vocabulary at the same time. Experiments in the presented setup have been conducted with a name vocabulary which was restricted to known persons. Unknown names could be detected as OOV and can be spelled for learning. In the following setup we extend this model by modeling recognition of names with a dynamic vocabulary recognition approach. With this model, recognition is first conducted on a small set of names from all known persons. In case an OOV is detected, the speech recognizer automatically switches to a larger vocabulary including the top 1000 names plus names from the social network. As names from the larger vocabulary are associated to higher speech recognition error rates, the error model of the speech recognizer is adapted accordingly.

5.5 Reinforcement Learning with Multimodal User Model

The setting presented in the previous section already integrates different types of spoken input plus face and voice identification. However, most aspects of multimodal integration are left for the dialog strategy. This section introduces a further experiment, which integrates the multimodal user ID model, which has already been introduced in chapter 4. Experiments presented in chapter 4 show improvements for person identification. Experiments presented in this section also show a better performance of the dialog strategy, by integrating the multimodal user ID model. In this experiment, the user ID model integrates observations made during the dialog in a Bayesian network, such as face identification, speech input and dialog-level observations, e.g. confirmations. It outputs a posterior probability for open set person identification, i.e. for all persons from the knowledge base plus an 'unknown' person model.

5.5.1 Data Set and Multimodal User Simulation

The experiment presented here extends the experiment from the previous section by integrating training data for the user simulation from a larger corpus, collected over a longer time period. The data set for face identification corresponds to the set used in chapter 4 for multimodal user identification. As we want to obtain results with external validity, the multimodal identification part is restricted to face identification, for which a large number of realistic data could be collected.

The user simulation models correspond to the models used in the previous experiment with modifications to face ID simulation and speech error models as follows. The motivation for computing new sequence hypotheses during simulation from single hypotheses was to create higher variability of the data. In the present set we already have a large number of interactions, and per interaction, a large number of available hypotheses. Therefore, we decided to use the original recognition results and vary the delay of arrival. The alignment of face ID to turns, i.e. to spoken user input, is shown in figure 5.4. The different delay of arrival is obtained by using a random variable with equal distribution in [0;5] that selects the number of images of the original recording to skip (values i,j,k). This has the effect that each dialog can produce different combinations of face ID hypotheses and speech input. In addition it has become necessary to introduce an additional variable to model the probability that face ID produces a hypothesis at all. Note that in the above example (equal distribution in [0;5] per turn), it is very unlikely that a complete session has no face ID input at all.

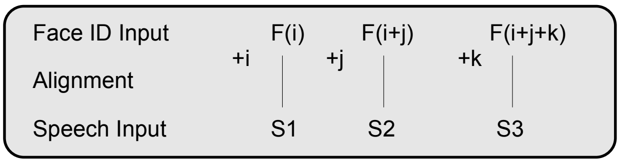


Figure 5.4: Alignment of face ID hypotheses to spoken input (user turns)

The speech simulation uses the same statistical models as in the previous experiment. The error model has been modified with an additional level of variability, to address the aspect that incorrect recognitions of names are not independent of each other. For example, the name of a specific person is not

Information Slot	MDP state	MDP state values
first name	FirstName	empty, filled, confirmed, oov
	FirstNameModality	asr, spelling
last name	LastName	empty, filled, confirmed, oov
	LastNameModality	asr, spelling
userID	userIDconf	low,medium,high max.90,max.95,max.99

Table 5.6: Dialog information slots and mapping to MDP states

covered even by the extended recognition vocabulary and can thus only be learned by spelling. The probability for recognizing this name correctly from ASR thus should be 0. We have addressed this aspect in the experiment by defining the probability of understanding the name correctly as a random variable which is selected once, at the beginning of each session. In the experiment we have used the following values as correct recognition of first name: 0.95;0.9;0.85;0.85;0.75;0.65;0.4;0.0.

5.5.2 MDP State Space

Also the state space of the MDP is built in a similar way as in the previous experiment, see tables 5.6 and 5.7. The main differences are that now the model integrates first and last names and user ID model. Both are modeled by the user model, which also decides when to set an information slot, e.g. each new input overwrites the existing value. The user ID state describes the probability of the hypothesis with the highest score. The values max.90,max.95,max.99 correspond to the absolute probability values (e.g. max.99 corresponds to 0.99 <= value <= 1.0, etc.) the values high and medium are set when the ratio of the best two hypotheses is higher than 100.0 (high) and 5.0 (medium).

The actions that can be executed by the strategy, are very similar to the action set described in the previous experiment. Table 5.8 lists the adapted set of actions.

5.5.3 Experiments and Results

Experiments were conducted with three different state configurations. Model 1 ('FaceID1') uses face ID input only instead of the user ID model, with the same granularity of probability scores (6 equally distributed values). Model

MDP state	values	description
nIDFailed	0,1+	number of failed attempts to confirm a user ID
nASRFirstNameFailed	0,1,2+	number of failed attempts to confirm a name from speech recognition
nASRLastNameFailed	0,1,2+	number of failed attempts to confirm a name from speech recognition
lastAction	*action*	name of the previous action

Table 5.7: Dialog state variables in the MDP state

category	actions
get information	ask_name, ask_first-name, ask_last-name ask_first-name-spelling, ask_first-name-spelling
confirm	conf_first-name, conf_last-name, conf_ID
finish dialog	accept-name, abort

Table 5.8: Set of dialog actions

2 ('BayesID2') restricts the granularity of the userID model to the values low, medium and max.90. It acts as a baseline to understand the contribution of more fine-grained probability levels of the user ID model. Model 3 ('BayesID3') uses the full MDP state as described in section 5.5.2, i.e. as BayesID2 it also includes the Bayesian user ID model.

Runtime behavior of the training algorithm was a critical issue in the previous experiment. Also this experiment needed significant optimization of the runtime behavior. While the whole RL process has already been implemented efficiently, the user ID model uses a Bayesian network implementation which was not adequate for training millions of dialogs in an acceptable amount of time. Two code optimizations led to a reduction of 99.7% of the runtime to 0.06ms per turn on a 2.1GHz Opteron processor (single threaded). The first optimization was to reduce calculation costs of n-best lists wherever possible. For example, output of the Bayesian network only requires the best two hypotheses, without creating a full n-best list including time-expensive sorting. The second optimization was to break the Bayesian network computations down to the least necessary computations. Since only the probability of the user ID needs to be computed given the fixed set of observations, this can be

condition	reward
each additional turn	-1
action repeated	-1
ask for spelling	-0.5
dialog aborted	-5
dialog end with correct full name	15
dialog end with correct first name	3
dialog end with correct last name	2
dialog end with wrong name	-10

Table 5.9: Reward Model

model	training reward	evaluation reward	dialog success	#dialog turns
FaceID1	14.14	12.88	86.4%	2.99
BayesID2	13.84	12.41	83.8%	2.97
BayesID3	15.05	15.03	91.3%	2.65

Table 5.10: Results of the different models

done very efficiently.

The reward model used in this experiment differs slightly from the previous model. The different rewards are shown in table 5.9. All reward categories are accumulated, especially, dialog end with correct full name receives 15+3+2 reward points.

Training was conducted with 10M training runs and 100k evaluation runs on two different data sets. Table 5.10 shows the results of the three models by comparing their rewards on training and evaluation sets and dialog success rate, i.e. correct identification rate per dialog for known and unknown persons, and the average number of dialog turns on the evaluation set. It can be seen that the face ID integration is better than the Bayes model with restricted model confidences, and that the full state space with probability levels of the full Bayes model outperforms the other models. This indicates that the posterior probability scores in fact contain valuable information. The full Bayes model also generalizes better over different evaluation sets, as it is the only model that does not degrade from the training to the evaluation set.

5.5.4 Discussion of the Results

The evaluation results confirm the initial assumption that the user model provides additional information which is not captured by the state model in the first experiment which integrates only face and voice ID in the classifier. The Bayesian user model also shows higher robustness, when observing the transfer results from training to evaluation conditions. However, this is only the case, if the output probabilities of the Bayes net are represented in the MDP state space with a sufficient level of detail. As can be seen from 'BayesID2', a lower resolution of the Bayes net output ignores relevant information. Thus, a trade-off was found between detailed sampling for accurate representation of the state space and a small number of sampling points for efficient computation of reinforcement learning. Improvements in the future might be to use the presented user model in a POMDP approach and integrate the Bayesian model directly into the POMDP's belief state. However, up to date systems are still very limited in the number of dialog slots.

Another aspect which is frequently discussed in reinforcement learning, is how to obtain a good reward model, as it is responsible for the outcome of the training. So far, there is no common general approach to define the reward function with less influence by the person who designs the experiments. As the reward function defines what is desirable for the system, it merges a multidimensional state into a scalar value by weighting different aspects. Therefore, the output depends on the importance given to these aspects by the human, e.g. to give higher weight to dialog length or to dialog success rates.

In Prommer et al. (2006) we already have successfully tested reinforcement learning in a multimodal user simulation for a barkeeper robot dialog, including speech and pointing gestures input, with comparable results for real-user and simulated user experiments. Also in this experiment, we argue that the user simulation, due to the type of models, is close to a real human-computer interaction, with the exception that especially subjective criteria cannot be measured. Since the real interactions highly depend on the speech recognition and face identification error rates in each dialog, which have high variance for unknown persons, and in addition, we have observed a significant training factor by users, the variance of the results of each dialog is high compared to other scenarios. For this reason, we prefer the simulation results, where 100,000 dialogs have been conducted to obtain the final evaluation results and attribute a higher reliability to the results obtained during simulation. As subjective feedback cannot be assessed in such a simulation, a separate section is donated to assessment of subjective criteria in the interACT receptionist in chapter 8.

5.6 Conclusion

Reinforcement learning in multimodal user simulation produces results comparable to a handcrafted strategy with even better results in the conducted experiments. It furthermore has the advantage that it can be obtained automatically and be retrained for new environments.

Two experiments have been conducted with simple integration of multimodal identification and a multimodal user model based on multimodal fusion and a belief network. For both experiments, user simulation techniques were presented for multimodal simulation. Significant speed ups have been proposed for multimodal simulation to achieve realistic training times.

In the simple integration experiment, ID hypotheses from different recognition components are integrated in dialog, and depending on the trained conditions (error models, distant speech vs. close speech), the strategy selects which hypothesis to trust and thus implicitly implements a confirmation strategy over multiple modalities. Confidence measures evaluated for face ID provide additional improvements. The system combines identification and learning tasks within one dialog.

The second experiment which combines reinforcement learning with a multimodal user ID model, the user model outperforms the simple integration scheme especially on unseen data, which we attribute to the higher robustness of the user model against varying recognition conditions.

Part II

Knowledge Acquisition in Dialog and Learning over Time

Chapter 6

Dialog-Based Learning for Knowledge Acquisition

This chapter describes a dialog-based learning approach for knowledge acquisition of a humanoid robot, a generic entity model, and a modular dialog architecture which implements dialog modules for the realization of learning tasks. The approach presented here is applied to specific tasks in the following chapters.

This chapter introduces the model of dialog-based learning, the knowledge model, learning scheme, and dynamic knowledge sources of the dialog system. Afterwards, the design of dialog modules for different learning tasks is discussed and a model for knowledge mending is introduced. They are built upon by the following chapters which describe and evaluate the realization of the model for different learning tasks. The dialog architecture builds on the TAPAS dialog manager described in the part I. It is extended here with additional details of the learning strategies.

The knowledge base, i.e. descriptive knowledge representation, defines and stores information and knowledge of the system, and provides a structure that can be updated by the dialog component. After describing the knowledge model in the following, we introduce the dialog model and the connection between both components.

6.1 Introduction to Dialog-Based Learning

As motivated in the introduction, especially section 1.1, the system studied in this thesis uses learning strategies for

- extending the system's knowledge and verifying knowledge
- correcting or discarding stored information

The setting for learning in this work is a fully developed system that is extended during runtime (in contrast to work which focuses on learning from scratch, e.g. early stage language acquisition). The problem defined in the introduction of this thesis furthermore specifies that the goal is to acquire information about persons and objects in real-world scenarios, and that these are represented in a knowledge model which combines semantic information, properties and multimodal identification models. With the fully developed dialog system, it will be possible to extend an existing environment model of objects (as described in the chapter on object learning) or to build a model from scratch (as in the experiments with the interACT robot receptionist and person modeling). The knowledge model presented in the following therefore must be able to combine these representations, and must be expendable by the dialog. Further challenges must be considered for the conversational system, which must be able to deal with new words, recognize new semantic constructs and detect unknown entities. Learning of a new entity therefore must consider detecting where models must be updated and updating speech vocabulary, recognition and understanding grammars, and semantics and multimodal identification models on the fly. These requirements, and especially the aspect of long-term maintenance through dialog, exceed the current state of the art, as laid out in section 2.4, and require a new integrated approach for dialog-based learning.

In this chapter we argue how these aspects are covered by the learning model. Evaluation of the approach is conducted in the following chapters by testing single learning tasks (e.g. objects, semantic categories, person models) and finally in a long term study of the interACT robot receptionist.

In contrast to (standard) supervised learning, dialog-based learning is conducted autonomously by the robot without manual assignment of correct labels by annotation. The learning process is restricted to knowledge provided by other agents - persons - and to kind of knowledge that can be communicated by the chosen style of communication. The system, which implements the dialog-based learning approach, has been studied in the context of a humanoid robot, and is restricted to speech communication and visual perception. As a consequence, the style of communication is restricted to these modalities and their characteristics.

Still, a complex dialog system is required which can collect information required to conduct task-specific operations on the knowledge base. To be able to address separate issues separately, the dialog system is broken down into separate interaction patterns implemented by dialog modules. Each dialog module implements a separate learning task, e.g. identification of a person, or learning a new word as a property of an object, which has a well-defined dialog goal, has a reduced implementation complexity, and can be combined with other modules for a complex dialog system. The design of the knowledge model and dialog strategies considers an error-sensitive interaction style and possible restrictions of the communication channels.

6.2 Knowledge Model and Learning Scheme

6.2.1 General Object Entity Model

The general object entity model is a generic model for representing any object or person in the knowledge base. Figure 6.1 displays the generic entity model of the knowledge base. The model is tailored to capture the specific requirements of multimodal representation, such as integration of semantic models, multimodal classification data and reference to speech recognition grammars, but it is also general enough to cover the learning tasks addressed by this thesis. In the following sections, specialized models are used, as different entity classes have different sets of attributes, e.g. a person entity has an attribute representing the first name, while a DVD has an attribute representing the title of a DVD.

Every entity represents an object, which is a complex data structure. Each object is associated with four main types of information.

- **ID**: Each object is associated with a unique identifier. The *Objects-Database* contains all known objects with their identifier as primary key for reference. Learning of a new ID corresponds to learning a new object entity and storing the object in the database with a generated ID.

- **MMID**: MMID represents multimodal classification data. For real world objects, visual features are extracted from an image of the object. In case of persons, MMID classifiers exist for face identification, and voice identification. Learning of MMID information corresponds to collecting classification data and either creating a new entry or extending existing MMID information.

- **CLASS**: The class of an object defines its semantic representation as a concept in the ontology. Semantic grounding defines different types of real world objects and distinguishes, e.g. a cup from a DVD. Learning of CLASS corresponds to semantic grounding, to associate the object with an ontological concept.

- **ATTRIB**: The ATTRIB represents a list of properties. The object's class can define which properties are available. Properties are important also for spoken reference, e.g. a person's name, or an object's color. Each attribute thus is described by a label, which is a textual representation that can be uttered either by the user or by the system. Learning of an attribute can include detection and learning of new words in speech recognition.

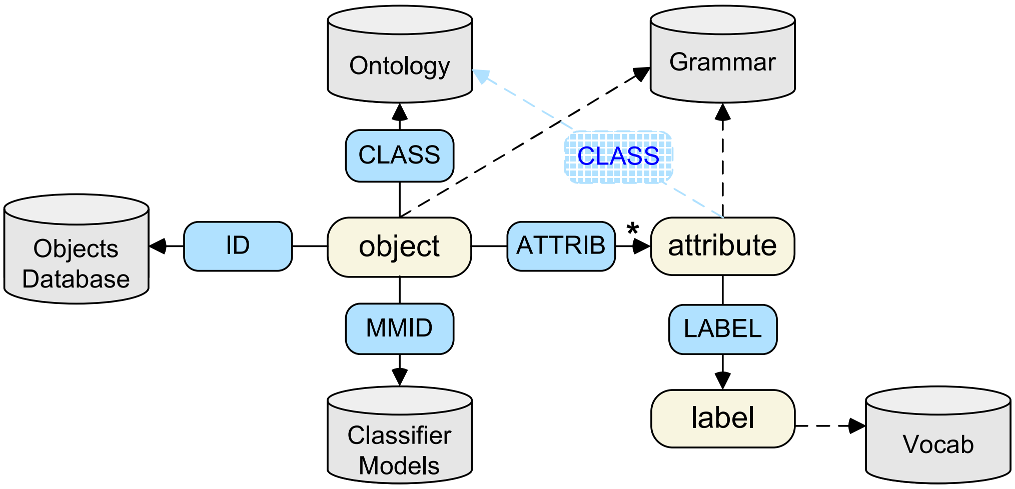


Figure 6.1: Knowledge entity model

Arrows which are painted as dashed lines in figure 6.1 represent connection to the dialog system's knowledge sources for speech recognition and understanding. While the connection between the label and the vocabulary can easily be interpreted in terms of the label belonging to the speech recognizer's vocabulary, the other connections are more complex. Connections with the grammar indicate that the grammar is automatically generated in part from database information, utilizing the object entries' property structure, and generating grammar productions using general syntactic rules, in a way that allows a user to refer to the object. This automatic generation of grammar allows dynamic updates of recognition and understanding components of dialog manager and speech recognizer during runtime, as explained in section 6.3.2.

6.2.2 Learning with the Entity Model

Dialog-based learning is conducted incrementally, and updates of the knowledge base are performed stepwise. A learning step is defined by a sequence of dialog turns during a dialog session with a user, after which one or more knowledge base update operations are performed. A dialog session is understood as a conversation between the system and a user and consists of a sequence of consecutive dialog turns. The dialog system used here identifies sessions by segmenting observed events by time constraints and explicit session ending, such as saying “goodbye”. When creating a new entity in the knowledge base, a complex structure is created, with information from at least one of the three categories MMID, CLASS and ATTRIB. Upon the creation of a new entity, a new ID is immediately associated with the entity. A question, which depends on the specific scenario, is what kind of information is necessary to associate with a new entity. The lower bound of necessary information is defined by the requirement of recognizing and identifying the object in further interactions. If possible, the system should also have lexical information to refer to the object during a communication with the user. However, this depends on the type of the object. For example, a person can be stored with first name and last name only. MMID and other information can then be added later, incrementally.

Besides the complex object structure, learning can be applied to single categories. For example, a strategy to learn a spoken word can be used to learn a new attribute of an object, and similarly to learn the name of a person. A modular definition of dialog strategies is designed for this purpose and is described with a detailed analysis in the following sections.

6.2.3 Identification with the Entity Model

It is clear that identification of an object is important to update the model incrementally with new information. This entity model allows different ways to identify an object. Pure classification is possible using MMID only. But additional steps are necessary to communicate the result. The speech modality provides information about ATTRIB and CLASS, and allows flexible formulations to describe an object. For example, to distinguish between a coke bottle and a book, the user could say “the book”. In this example, the user refers to the class of the object, according to some ontological model. To distinguish two books, more details about the desired book are used. Such information can also be provided by other modalities, for example by 3D pointing gestures as deictic reference (Holzapfel et al. (2004)). Structured semantic information (CLASS), and location in the environment (ATTRIB), are provided by the model for deictic resolution. To distinguish persons, the system requires first and last name, which in most cases are unique and sufficient for identification, also for unknown persons. Similarly, different pieces of information can be combined to identify an object.

6.2.4 Entity Model for Persons

For representation of persons the generic object entity model is specified as follows. Figure 6.2 shows the more specific entity model with a set of sample attributes. Differences to the generic object entity model are that the ID is now called UserID, the CLASS is now fixed as the semantic concept *person*, and ATTRIB and MMID are specified. The MMID contains two models, face identification and voice identification. The ATTRIB contains a sample set of attributes which is first name and last name, and two attributes of a social user model, the role within an organization and the research interest.

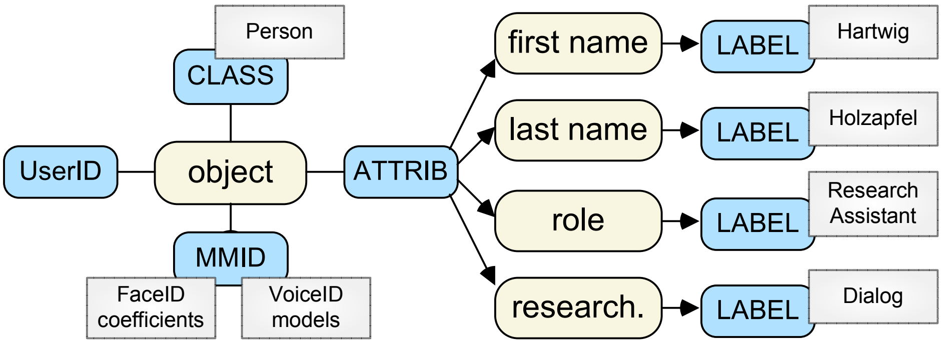


Figure 6.2: Adapted entity model for persons

6.3 Dialog Knowledge Resources

In addition to the knowledge base, which stores information collected by the system, the dialog manager uses further knowledge sources as communication knowledge for dialog behavior and knowledge sources for recognition and understanding.

The dialog knowledge resources comprise representation knowledge and interaction knowledge. Representation knowledge defines the aspects of the knowledge, which the system can talk about, and which it can extend by acquiring new information. Interaction knowledge tells the system how to obtain the knowledge via a communication process, i.e. the dialog strategy for acquiring new information.

6.3.1 Object Model

The (representation) knowledge sources are shown in the generic object entity model, figure 6.1, as an *ObjectsDatabase*, an *Ontology* of semantic concepts, a *Grammar*, *Classifier Models* and *Vocab*.

The *ObjectsDatabase* contains information about the environment including real world objects, persons and social relations. Each real world object is represented by a database entry which includes an ID (unique label), and contains information about its properties, semantic category (type), values for object properties, and an association to a list of observed textual descriptions. And as such, it implements the data structure of the knowledge base, presented in the previous chapter. The ID of an object is associated to the label of the object recognizer, for example *granini_juice_0001*; respectively *hartwig_holzapfel_0001* as an example of a person. The type values associates an object instances with ontological concepts, for example the object 'granini juice' is associated with the concept *obj_juice*. Values of object properties store additional information about the object instance, such as brand 'granini' or color 'yellow'. Observed textual descriptions could be 'granini juice'.

Type information and semantic categories of objects are modeled in the *Ontology*. The object ontology provides inheritance information (*isA* hierarchy of concepts with multiple inheritance) and defines properties that can be associated with objects. To be able to talk about object types, e.g. refer to the concept *obj_juice* by using the word 'juice', an additional mapping file is defined which is used for grammar creation in speech recognition and understanding and for spoken output. Besides representations of objects, the ontology also models speech act information and object properties. The ontology formalism follows an object oriented approach, introduced in Denecke (2000), which also allows multiple inheritance. Thus, a concept can inherit type classes and functional classes at the same time.

The *Grammar* resource represents lexical and grammatical information of the objects for spoken interaction including speech recognition and understanding. The grammar describes how objects are embedded in grammatical constructs, i.e. their lexical representation and how the objects are referenced in speech. In the following example: "please open the granini juice for me", the term 'granini juice' is a description of an object which is stored in the database. The lexical tokens — here 'granini juice' — are read from the database and dynamically update the grammar at a predefined position defined by semantic categories. The lexical tokens are modeled by the *Vocab*ulary, a resource mainly maintained by the speech recognizer with orthographic and phonetic models of words.

Finally, *Classifier Models* encode information for multimodal classification. The models used during runtime correspond to commonly used classifiers, such as visual person identification, object recognition etc. These models can be derived from data collected during interactions or be computed from manually labeled data through a training process.

6.3.2 Language Understanding and Grammars

The definition of grammars in our system follows the approach and formalism of semantic context free grammars, see Denecke (2000). This formalism defines a grammar based on semantic categories, in addition to syntactic information with the formalism of vectorized grammar nodes. A grammar node is a 3-dimensional vector of semantic category (SEM), syntactic category (SYN), and subcategory (SUB), and is represented by <*SEM,SYN,SUB*>. With this construction, the grammars inherently carry semantic information in their grammatical structure. The grammar's syntax is defined in the Java Speech Grammar Format (JSGF)[1].

The grammar is shared by the speech recognizer, which uses a context free grammar as a language model, and by the dialog manager, which uses these grammars for natural language understanding and contextual weight adaptation. In the presented approach, parts of the grammar are generated automatically from database and ontological information. Rule generation from database information makes use of semantic categories and rule inheritance, which is defined in the following way. A non-terminal symbol that is defined on the right hand side of a rule, e.g. <*obj_openable,NP,_*>, is automatically extended to its descendants, e.g. <*obj_juice,NP,_*>, if <*obj_openable,NP,_*> is not defined in the rule set. Such inheritance approaches are applied to functional object categories, e.g. *openable*, *portable*, *eatable*, etc. These functional categories are used throughout the grammar to integrate actions / speech acts with objects that are applicable to these actions. For example "please open the granini juice for me" is covered by a grammar rule that interrelates the speech act *act_open* with an object of type *openable*. The simplified rule looks as follows:

```
public <act_open,VP,_> =
          <please> <open,V,_> <obj_openable,NP,_> <recv_me>;
```

The syntactic categories used in the example are *VP* for verb-phrase, *V* for verb, and *NP* for noun-phrase. Subcategories are not used here, but are used in the grammar, e.g. for singular and plural rules or contextual utterances.

The actual generation of grammar rules from database information is realized with the following approach. So called 'import' statements which are specified as the right-hand side of a grammar rule, define grammar rule generation with database content. The presented grammar generation approach from database information extends previous work on a multimedia access dialog system in Gieselmann and Holzapfel (2005), by the definition of more complex import statements to match object descriptions, and sup-

[1] http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/

porting interactive extension of the models. The left hand side of the rule is a standard non-terminal symbol, e.g. <*obj_juice_db*>, the right hand side of the rule is started with a VOID element, which conforms to the JSGF syntax specification. The import definition includes DB connection, imported fields and semantic conversion rules with the syntax *import DB-ref* $entry_1$ $entry_2$ .. $entry_n$. Each entry consists of a table-field pair with an optional list of semantic values in the form of *table field* { sem_type_1 sem_value_1 ... sem_type_k sem_value_k }. For example, the following rule with import statement

```
public <obj_juice,N,_> =
  <VOID> { import exampleDB
      objects_juice brand \{ BRAND objects_juice:brand \}
      objects_juice type \{ TYPE import \} };
```

updates the right-hand side of the rule and generates the following productions

```
public <obj_juice,N,_> =
      granini { BRAND granini } juice { TYPE juice }
  | valensina { BRAND valensina } juice { TYPE juice } ;
```

from the database entries

type	flavor	brand	onto type
juice	apple	granini	*obj_juice*
juice	orange	granini	*obj_juice*
juice	orange	valensina	*obj_juice*

To allow understanding of any known property in combination with any known object type, for example 'red cup', 'blue DVD', or unobserved combinations such as 'green juice', can be parsed by combining two independent rules for properties and object types. The latter example is necessary to understand assignment of yet unobserved properties. If one wants to restrict grammar coverage to only known property-object combinations, the import statement is specified accordingly with more than one imported field, as done in the example above.

As mentioned above, the grammars are shared by the dialog manager and the speech recognizer. For the purpose of using the grammars as language models, the grammars are converted to a self-contained standard context free grammar. This is done by the Tapas dialog tools in a compilation step at system startup. During system runtime the speech recognizer's grammar and the dialog system's grammar share the same structure, but are different instances. Automatic updates to the grammar, which result from the learning method, are always modifications to the database rather than to the grammar structure. The learning step updates the database and modifies the runtime objects of speech recognizer and dialog manager accordingly during runtime, by adding new entries to the corresponding grammar rules. With this approach, the grammars of dialog manager and speech recognizer are always kept synchronized, a prerequisite for tight coupling of both components. The advantage of tight coupling is that speech recognition output already represents a parse-tree, and no additional parsing is necessary, to initiate language understanding, which maps grammar nodes to TFS nodes. Another advantage of tight coupling is that the dialog manager can maintain a generic expectations model. For example when the system asks the user to name the color of an object, the expectations model contains ontological concepts that can describe a color, and subsequently the speech recognizer's grammar rules are adapted to better fit the expected input. Since the expectations model contains (among others) speech acts such as *inform_color* and property descriptions such *prp_color* which are mapped to grammar rules *<inform_color,VP,_>*, *<prp_color,A,_>*, *<prp_color,AP,_>*, the presented learning approach does not interfere with this model and works in combination with this approach as well. As it has been shown previously, contextual weighting improves speech recognition accuracy significantly Holzapfel and Waibel (2006), especially for short responses, such as 'yes', 'red', or 'yellow'. It offers a benefit especially for large grammars, e.g. to prevent incorrect recognition of an object type, when a color has been said.

6.3.3 Specified Object Ontology

The robot's knowledge about objects is represented in a specific object model. The model specifies object classes, properties and views (visual classification models) of the object. Classes and properties are modeled in an ontology, where a real object can be associated with multiple properties and classes. This allows different attributes to be associated with one object. Examples for properties are color, name, and title of an object.

Figure 6.3 shows an excerpt from the system's ontology, which is used for object learning experiments presented in chapter 7. Properties are listed at the lower part of the figure. The middle section shows object classes (also referred to as types). The upper section shows functional concepts that model how an object can be used.

The ontology defines object classes hierarchically. General objects are displayed at the top; more specific (inheriting) objects are displayed further down in the ontology. Each object can inherit from one or more functional concepts. Each child of an object inherits the parent's functional classes.

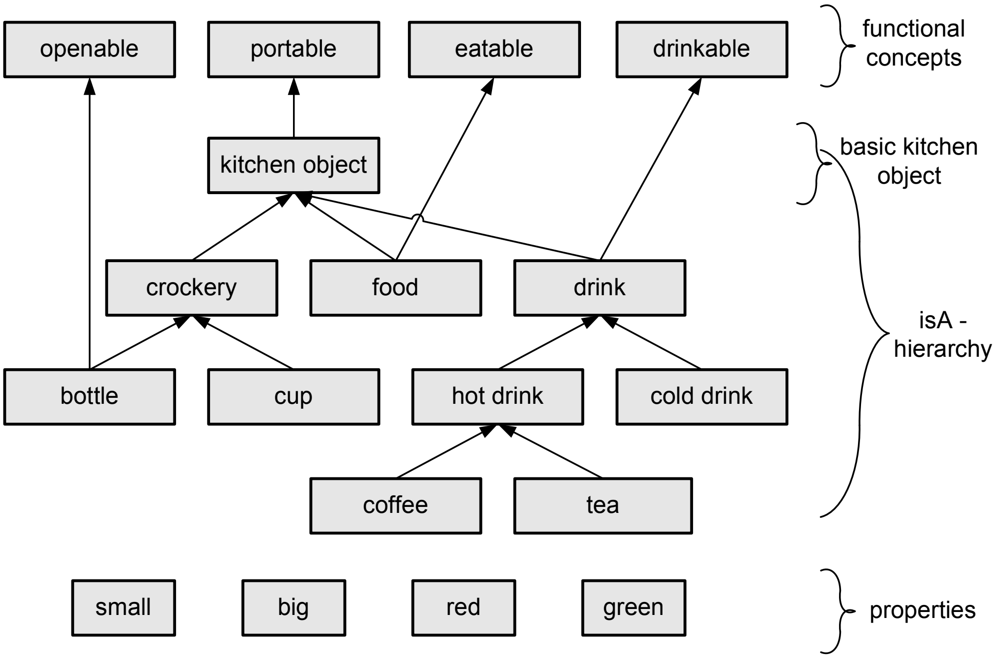


Figure 6.3: Ontology organization with functional concepts, type hierarchy and properties; example extracted from ontology for object learning experiments

This inheritance relation is used in the definition of semantics. As mentioned in the system overview, typed feature structures (TFS) Carpenter (1992) are used to represent semantics in the dialog system. The definition of TFS allows types from a hierarchy, including multiple inheritance.

For example, an object instance of a kitchen object has the semantic concept *kitchen object*, and all inheriting concepts, such as *crockery*, *drink* and *food*, are kitchen objects as well. In further inheritance, the concept *drink* is split into the concepts *hot drink* and *cold drink*.

The ontology's functional concepts describe what can be done with an object. For example, all objects which are described in the presented ontology are *portable*. However, only a bottle is *openable* whereas coffee is *drinkable*. These functional classes are used to refer to objects in the semantic grammar for speech recognition and understanding. For example, if the user tells the robot to open something, the concept which is used in the grammar is of type *openable*. All objects that inherit from *openable* are automatically inserted into the grammar and can be referenced by the user. The complete list of functional classes used in the experiments covers nine categories: *cook*, *drink*, *eat*, *fill*, *open*, *play*, *carry*, *switch on*, and *watch*.

6.4 Functions for Knowledge Base Update

Updating the knowledge base is conducted by the dialog manager with a small number of update functions. In this section we introduce the update functions and their implementation via background and interactive tasks. We refer to **update** as any modification of the knowledge base. The different update functions are **insert**, **modify**, and **delete**. With respect to the special requirements of the multimodal knowledge base, it is necessary to describe the update functions and describe which information must be obtained to execute each function.

Following the knowledge entity model, an update function is defined for each of ID, CLASS, MMID and ATTRIB. Each function thus is implemented with a different granularity level. To insert a new ID, a complex entity is inserted. For example, a person as a complex entity is inserted with an ID, a class, a set of attributes, and optionally faceID information. A lower information bound is used to insert the entity, to ensure that the entity can be identified afterwards during dialog interactions. For example, we can defined a lower information bound for a person-ID by the first and last name of a person. A new person-ID can only be inserted with the first name and the last name; other information, such as faceID and other personal information, is optional. Inserting a new CLASS corresponds to creating a new class in the ontology. The CLASS category is especially useful during interactive learning of new objects, but plays a minor role in learning person IDs. Inserting a new MMID mainly updates the classifier (i.e. faceID, voiceID, object recognition) for which a one-to-one association exists of classifier output and ID. Inserting a new ATTRIB is implemented on two levels. On the vocabulary level, new words are learned and added to the vocabulary of knowledge base and speech recognizer, and are associated to existing semantic attributes (e.g. first name). On the semantic level, new attributes are inserted, e.g. semantic properties of an object.

The modify function has the same characteristics as the insert function, with the difference that additional information is necessary what to modify. The delete function can potentially be implemented for all categories. It is usually executed by a confirmation dialog or by offline processing, and requires only the information which entry to remove from the knowledge base. In case of complex entries, relations are deleted as well. Examples for using modify are to change the first name of a person (ATTRIBUTE) or to merge to person entities (ID). Examples for using delete are to delete labels (ATTRIBUTE), delete complete entities (ID), or to delete MMID data. Among these examples, deleting MMID data is only conducted during offline processing.

knowledge acquisition type	**acquisition task**	**description**
background acquisition	init phase	initialization of knowledge base through information retrieval and information extraction
	interaction active	while the conversation is going on obtain background information
	offline	reconfiguration of the knowledge base removing deprecated entries adding new and updated information
interactive acquisition	words and names	new words are learned and update speech vocabulary
	semantics	semantic grounding and learning of new object's relations
	objects	complex real world objects
	person-ID	real persons as instances of complex knowledge entities
	social information	obtaining and affirming information about social networks of persons and social user models

Table 6.1: Terminology and Classification of Knowledge Acquisition in the Dialog System

Update functions are implemented by interactive knowledge acquisition, but also by background processing tasks. Table 6.1 lists a set of knowledge acquisition tasks.

6.5 Dialog Modules for Interactive Knowledge Acquisition

Updates during the interaction are executed by the dialog strategy. For this purpose, a modular dialog strategy approach has been chosen, which offers benefits in comparison to a unified dialog strategy. Other existing modular approach to dialog systems are agent-oriented dialog systems, e.g. Turunen and Hakulinen (2003); Nakano et al. (2006), or interaction sequences Denecke (2002), which allow to define module-like building blocks of a dialog strategy in a generic approach.

A dialog module can be a small module to learn a new word, e.g. a person's name or an object property, including word acquisition by spelling and confirmation questions to verify pronunciation. A dialog module can also simply be used to identify a person. A module, which is associated with update functionality, can execute one or more update functions after successful completion.

Following up on the discussion in the previous section (section 6.4), table 6.1 lists the main tasks that are implemented by dialog modules. These tasks represent the most challenging problems that the dialog system has to solve, to collect information necessary to execute the update functions. At the same time, they cover all aspects discussed so far, which are to insert or update a new ID, CLASS, MMID, ATTRIB and attribute labels, and are executed by dialog modules. Each of these tasks is fairly complex, so in the following, separate chapters are devoted to some of the tasks including the task's detailed implementation, experiments and evaluation.

The challenge of the top level of the modular dialog approach is to decide which module to activate for processing and reacting to the current input. In this work, the active module is determined by a score function, which is executed each time when qualifying events are observed. The module with the highest score is selected for preferred execution and conducts its strategy. The score (equation 6.1) of each module m is defined by the score that m can generally handle the input, the score that the input matches the current module's expectation, and by a penalty for switching to m from module m_{t-1}. The switching penalty also depends on the current state, because the penalty is different if the module m_{t-1} has been completed.

$$\begin{aligned} score_m(input, state) = \\ score_m(input) + expect_m(input, state) - penalty(m, m_{t-1}, state) \end{aligned} \tag{6.1}$$

Application of the modular approach allows separation of concerns. For example, each single module can be optimized individually. It is even possible to develop different implementations of a single module, for example a handcrafted strategy and a strategy trained by reinforcement learning, and switch between different implementations of the same module. By separate module implementations, this approach also offers to mix dialog approaches. For example, strategy optimization with reinforcement learning or POMDP models has been successful mostly on small dialog tasks. This way, small tasks are created by training single modules, while other modules can be created manually.

Figure 6.4 shows the architecture of the modular dialog approach. It highlights the components relevant to the selection and execution of the modules. An example of the modular dialog is shown in table 6.2. After the greeting has been conducted, the userID module is started to identify the person. In this example, this is initiated by simply asking the user for his name. By the

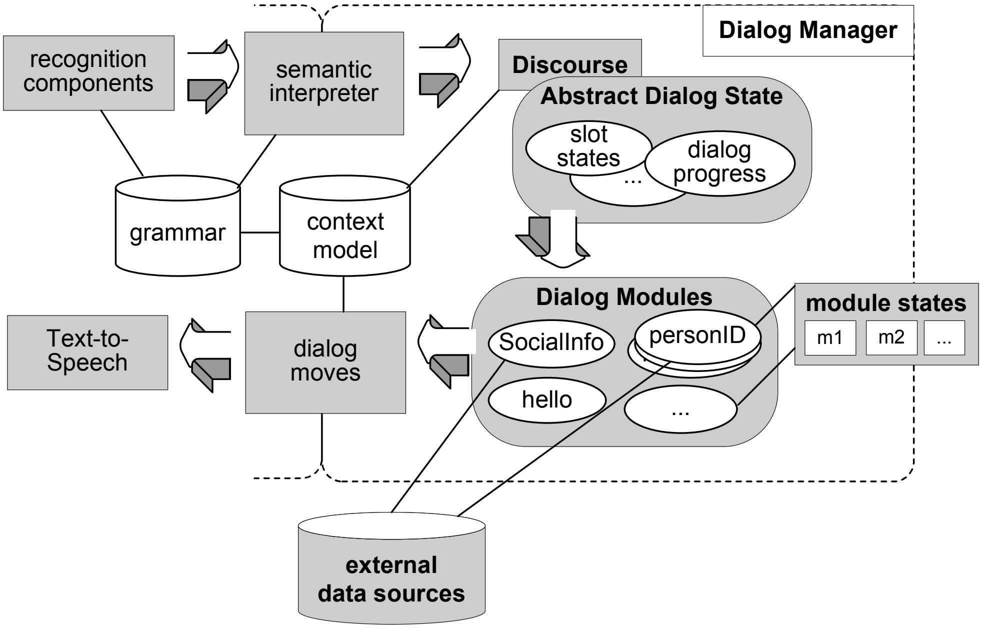


Figure 6.4: Modular dialog architecture

interaction	speech_act	system goal	module	score
U: hello	greeting	greeting	hello	0.9
S: Hi, what's your name?	ask_name	identify	userID	
U: my name is Peter	inform_name	identify	userID	0.8
S: is your first name Peter?	ask_confirm	identify	userID	
U: yes	confirm	identify	userID	0.8
...	...	social talk	social	0.5

Table 6.2: Example dialog with module selection

end of the userID module, the user has confirmed his name. The confidence now is high enough to end the module and execute an update function. Note that only for the purpose of simplicity of the example, the last name has been omitted. If the user already exists in the database, i.e. the system identifies a known person, only the MMID data is updated with newly collected data. If the user is unknown to the system, a new ID is created with newly collected MMID data, and the given name as the initial attribute. Furthermore, the system's user model now contains the ID of the user. Information that is collected in the following, e.g. by the social talk module, can then be used to update attributes of this person's model.

6.6 Dialog Module for Learning Person Entities

With the newly introduced knowledge entity model and update functions, we can now discuss the *userID* dialog module, which can insert or update a complex person entity model. This module builds on multimodal processing and user identification as introduded in chapters 3 and 4. Execution of the dialog module can either be initiated by the system (i.e. on system initiative), or by the user, e.g., the user states an explicit request such as "do you recognize me?".

The *userID* module implements the insert function (ID, MMID, LABEL) and modify function (MMID). The CLASS and ATTRIB categories are fixed, and are not modified. Execution of the insert function(s) is conducted when the goal of the dialog modules has been reached, which is defined in different ways for handcrafted and reinforcement learning modules implementations. In any case, acquisition of first and last name is the lower bound of information, before an ID is updated. In case of a handcrafted dialog strategy, the condition is achieved, when the user confirms his first and last name. In case of a dialog strategy trained by reinforcement learning, the goal condition is implicitly defined in the state-action space of the strategy. When the reinforcement learning strategy is designed appropriately, the effect is very similar, and usually finalizes the strategy after confirmation or high classification confidence.

One question is how to built a dialog that can handle known and unknown persons. In the present system both cases are addressed by a joint module, as the system does not know in advance whether the person exists in the knowledge base or not, and unknown person classifiers, such as face identification, have not provided reliable estimates by the beginning of the dialog. The system can therefore only tell if the person is known or unknown when the person has been identified by the module. The insert(ID) function is applicable, when the person is identified as unknown. In this case, the system obtains the user's name, and after finalizing the module, a new ID is created in the database, with CLASS *person*, collected MMID data, and the attributes first name and last name. However, initialization of the MMID data is not as straight forward as it seems at first glance. To complicate things, the system runs in a real-life environment and different persons can pass the robot while someone is talking to the system. Recording of MMID data is thus performed continuously by a person tracker to obtain a continuous track of the user (robustness issues have been discussed in detail in chapter 4). When the dialog session has ended, the system needs a a little while to reconfigure the vision and voice ID models with newly collected data and store the new models with modified parameters on the hard drive. In the tested system this took roughly 10 seconds, depending on the size of the user database.

Updating MMID data for a known person is conducted by the same dialog module. Usually, the module finalizes much faster than for an unknown person, for example, because the person's name is already known, or when the person is recognized immediately by faceID. The purpose of updating the MMID data for a known person, is to adapt the models and gain more robustness, e.g., against different light conditions, and different characteristics of the person. Note that additional robustness can be gained by offline processes, which can optimize ID models with (data-)age-based weighting of collected MMID data, when the system is in 'sleep' mode.

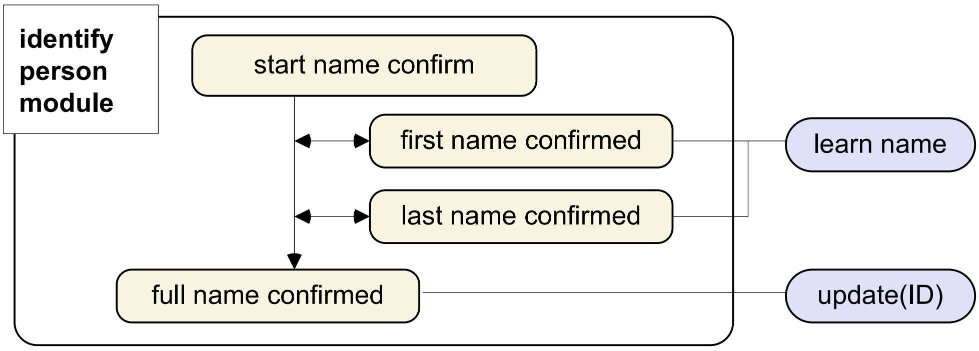


Figure 6.5: Confirmation states of the person identification module and associated learning functions

The dialog module can also learn new words during the interaction. A state diagram, which shows the module's confirmation states is shown in figure 6.5. Once the user's first name is confirmed during the dialog, the system enters the state *first name confirmed*. If the name is new to the system, the associated name learning function is called (as described in section 7.2), which updates the speech recognizer's vocabulary, dictionary, language model classes and grammar. A sample dialog with name learning is shown in table 6.3. The last name of the person does not exist in the vocabulary of the speech recognizer, and thus it is recognized as an OOV. To learn the name, the dialog strategy asks for spelling to obtain the grapheme representation. After confirmation of the name, the new name is learned. After confirmation of the full name, a new ID is created in the database, and MMID data, CLASS, and ATTRIB are created as described above.

speech	speech act	state space	update
S: I am Robbi Robinson, what is your name?	ask_name	FIRST: empty LAST : empty	
U: My name is Hartwig Holzapfel	inform_name [FIRST ”Hartwig” LAST ”OOV”]	FIRST: spoken LAST : OOV	
S: Did you say, your first name is Hartwig?	ask_confirm	FIRST: spoken LAST : OOV	
U: yes	confirm_name	FIRST: confirmed LAST : OOV	
S: However, I didn't get your last name, can you please repeat it?	ask_confirm	FIRST: confirmed LAST : OOV	
U: my last name is Holzapfel	inform_name [LAST ”OOV”]	FIRST: confirmed LAST : OOV	
S: could you please spell that?	ask_spelling	FIRST: confirmed LAST : OOV	
U: H O L Z A P F E L	inform_spelling	FIRST: confirmed LAST : spelling	
S: Did you say, your last name is Holzapfel?	ask_confirm	FIRST: confirmed LAST : spelling	
U: yes	confirm_name	FIRST: confirmed LAST : confirmed	insert(LABEL)
S: nice to meet you, Hartwig			
module finalized			insert(ID)

Table 6.3: Sample dialog between system and ‘unknown’ user. Last name is not covered by speech vocabulary and is learned during the dialog

6.7 Dialog Modules for Knowledge Mending

As the learning processes deal with uncertain information, another important aspect for maintaining a knowledge base is to deal with incorrect entries in the knowledge base by deleting information. This task is accomplished by the knowledge mending dialog module. The knowledge mending module can resolve errors as incorrectly learned names or multiple entities for a single person. Also dynamics of the real world lead to errors in the knowledge base, e.g. the status of an employee changes, or one-time visitors are registered by the system.

This section and the respective evaluation in chapter 10 include results from the Diplomarbeit of Philipp Große, which he has written at the Universität Karlsruhe (TH), supervised by the author of this thesis. In particular, design on the clustering approach and experiments with the dialog-based interaction have been joint work, details on confidence estimation are described in greater detail in Grosse (2009).

6.7.1 Error Types

Figure 6.6 depicts the four errors that occur in the knowledge base, after incorrect updates by the learning dialogs. They categorize errors of person ID entries, which we understand as the most relevant errors, as these have the highest influence on the quality of the knowledge base and multimodal classification models. Other errors can also occur, but are not addressed here. Error (1) occurs if one person is modeled by two different IDs, but the names of the person are phonetically equivalent, e.g. ‘Stephan’ and ‘Stefan’. This error can happen if the person speaks his/her name, and the system chooses the correct pronunciation, but incorrect word model. Error (2) occurs if one person is modeled by two different IDs, but the names of the person are not phonetically equivalent, e.g. ‘Stephan’ and ‘Skrefan’. Such errors mostly happen when the user confirms an incorrect name, e.g. after incorrect spelling recognition and the user is not completely sure if the system got the name right, or the users accept a slight mispronunciation. This error is enforced by the text-to-speech pronunciation, as some names sound a bit ‘weird’ (quoted user feedback), and users cannot tell if the name is pronounced correctly. Error (3) occurs if the ground truth changes over time, e.g. a person was relevant only in the past. Error (4) occurs if two different persons are represented by the same ID. This error mostly happens due to communication problems and sometimes insufficient understanding. The following short dialog excerpt demonstrates the problem. `System: Hello, you’re Philipp, right?` (meaning: system has a hypothesis from face identification and wants to confirm it) `User: Yes?` (meaning: ‘hello’? and user did not understand what the system said)

6.7.2 Dialog Tasks

As all other modules, this module is restricted to information that can be communicated verbally. This is a strong restriction, because the system cannot select any image and obtain a label for this image from the user, as is the case in standard active learning, e.g. Guo and Schuurmans (2007).

In contrast, dialog-based learning must rely on means of communication. Table 6.4 lists tasks that are implemented by dialog modules to conduct

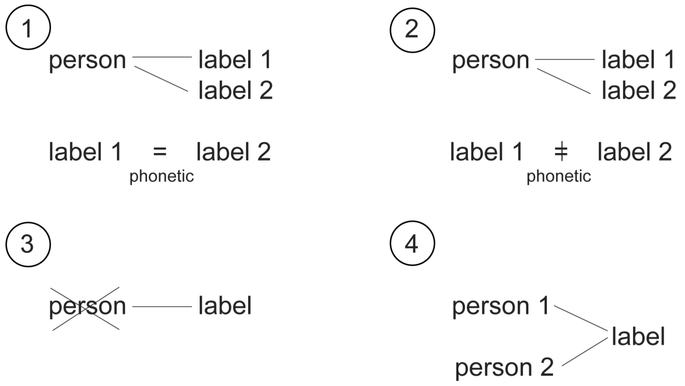


Figure 6.6: Incorrect database entry types of person identification

task	treated error
is-valid	incorrect name has been learned
	i.e. no person exists with learned name
up-to-date	deprecated entry or incorrect name has been learned
merge	one person has several entries e.g. Stephan and Stefan

Table 6.4: Tasks for error resolution with interactive dialogs

knowledge mending with interactive dialogs. Obviously, an entity is referred to indirectly, for which the system can use the entity's attributes. Thus, the first task from table 6.4 – *is-valid* – checks, if a given name can be considered as a valid entry. Note that this does not directly imply that the corresponding entity is valid. To improve the likelihood of speaking about the right entity and to interpret the dialog result correctly, different ways of pronouncing a name are considered (especially foreign names), the risk of not understanding a name correctly (for example bad pronunciation of the text-to-speech system), and differences in grapheme representation with the same pronunciation (e.g. 'Stephan' is pronounced the same way as 'Stefan').

A similar question targets at finding out, whether an entry is up to date, for example by asking something like "does Stephan still work here?" or "have you seen Stephan lately?".

A more complex operation targets at merging two entities with different names. For this purpose unsupervised clustering is conducted with probability estimation that two clusters including different sessions represent the same person. The dialog task is to obtain the right label (usually one of the two entities). This operation makes sense if the session with the incorrect label contains other valuable information. Otherwise, the knowledge base error can also be resolved with the *is-valid* task.

The list could be extended by more tasks, for similar and extended functionality. Thus we do not claim that the list is extensive. For example, a *correct* task would be a promising extension to solve mispronunciation, as a combination of identifying the incorrect entity and then replacing the data.

It can easily be imagined that the number of possible questions that can be asked to improve the knowledge base is larger than the number of questions that can be posed in a realistic setting without overburdening the user. Thus the user is a valuable but 'limited resource'. Similar to work done in active learning, it is important to provide a ranking of critical entries. These critical entries are the first in line to be checked by the mending dialogs. We present a ranking method, which produces a highly informative ranking. The rank computation is the result of a clustering and label mapping process, which implicitly already generates an assumption about incorrectly learned sessions and suggests corrections.

6.7.3 Mending Dialog Process

The objective of the presented knowledge mending approach is to model proactive error correction by the robot, with support from humans. At this point, it should be reminded that the learning method differs from state-of-the art active learning methods, where samples can be labeled with direct associations. That means that in our system, the user does not interact with a graphical user interface, where the user can label images of persons or respectively reject or accept names learned by the robot.

Figure 6.7 shows the integration of the mending dialogs and offline processes with clustering and mapping steps. The knowledge base is the database of persons collected during interactive learning dialogs, as indicated by the *Learning Dialogs* node. In regular time intervals, the system conducts offline processing, during which a cluster analysis is performed on the current state of the database to identify sessions that potentially have incorrect labels. During this process, a mapping table of person IDs (label mapping) is created and consecutive ranking is conducted, to rank problematic labels. The purpose of the clustering is to find sessions that represent the same person, but might be labeled incorrectly. Initially each cluster is associated with a single session. Clusters are merged by agglomerative clustering.

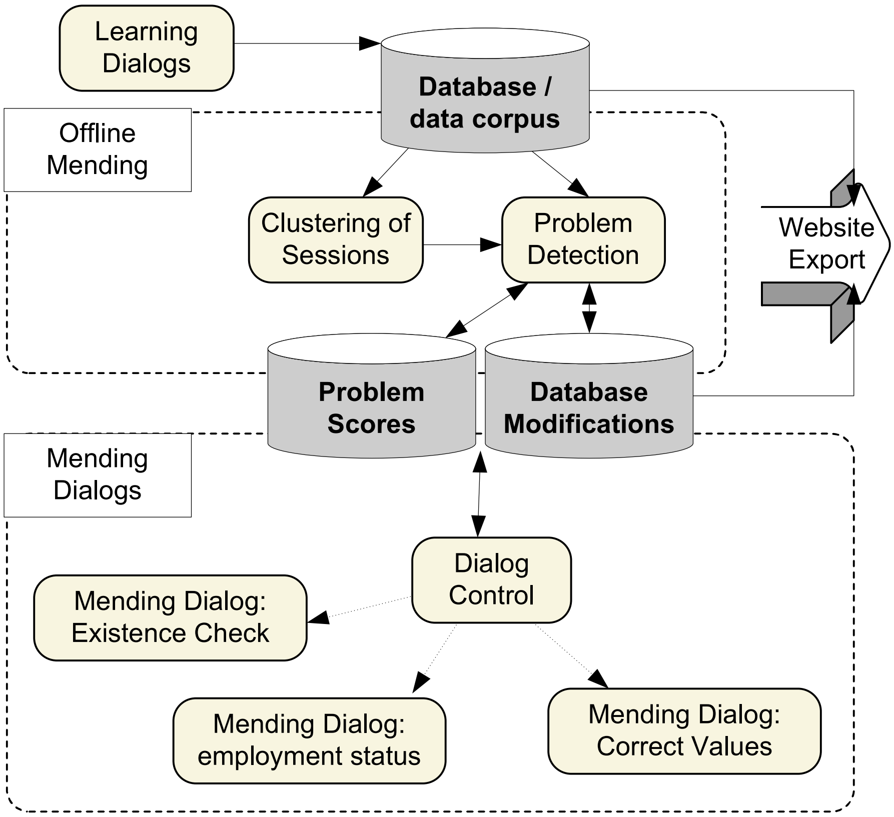


Figure 6.7: Knowledge Mending Architecture

As the original recordings in the corpus is already pre-segmented as sessions, the clustering algorithm treats a single sessions as the smallest unit, and clustering is conducted on sessions. However, the association of a session to a person ID label cannot directly communicated to the user, but the user can be asked for example about the name of a person, etc. Steps which are conducted after the clustering of sessions generally deal with labels, and also problem detection includes e.g. confusion of labels. The evaluation in chapter 10 shows significant reduction of the knowledge base errors by both, automatic clustering and dialog approaches. Chapter 10 also compares clustering parameters and corresponding quality improvements.

6.7.4 Cluster Analysis

The features used by the clustering process are extracted from recorded face identification data. As an agglomerative clustering approach, a distance measure is used for merging the two closest clusters in each step, until a stopping criterion is met. Ideally, the distance measure expresses a probability that two clusters represent the same person, where clusters are merged until the probability exceeds a predefined error level. In statistical analysis, significance tests, e.g. a t-test, are conducted to test if two sample sets differ significantly. In many machine learning applications, however, the given data do not fulfill the requirement of normal distribution of the t-test. Additionally, variation of the features is more dominated by different recording sessions (inter session variation) than by variations during one session (intra session variation). Figure 6.8 depicts the different variation aspects. As the inter session variation cannot solely be described by intra session variation, a specific data set is necessary to estimate this variation. We achieve such a model by training a confidence measure by logistic regression, where the input features express distance (of cluster centroids) and data variation metrics, which roughly compares to the idea of mean value and variance in a t-test.

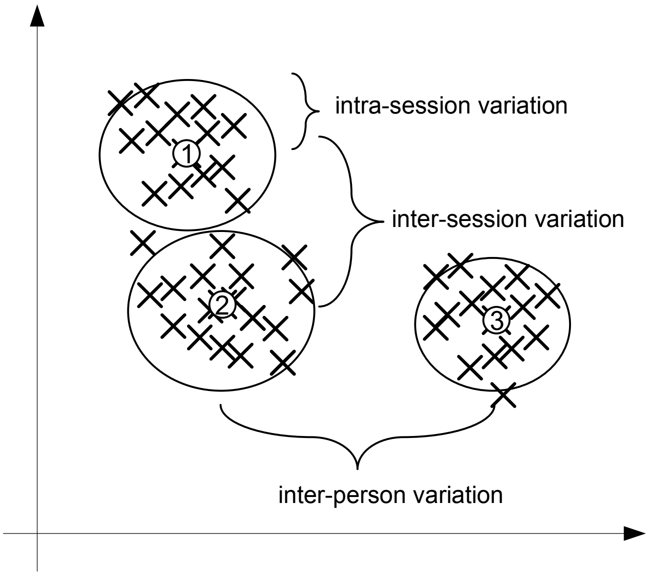


Figure 6.8: Pictorial example of intra-cluster variation, inter-cluster variation and inter-person variation

The distance measure is defined for pairs of clusters, and estimates the probability that both clusters represent the same person. We have chosen to use a confidence measure, as it represents a probability estimate. Therefore, the stopping criterion (e.g. confidence exceeds the 5% error level) has a meaningful interpretation and the threshold can be defined as a probability level (e.g. 5% error tolerance). The confidence measure is obtained by training a logistic regression model on a held-out data set.

6.7.5 Clustering Confidence

The data used for evaluation is the corpus data which was automatically recorded by the robot receptionist, which will be described in the following chapters. From the automatically recorded corpus, 59 sessions have been used for training and evaluation of the offline clustering approach, the remaining 106 sessions were used for evaluation of the clustering and dialog-based mending approaches.

As has been determined in Grosse (2009), the best features for confidence estimation on the given task are cluster mean distance (mean Euclidian distance of face identification coefficients) and cluster size measured by the number of sessions in the cluster. To evaluate the benefit of the clustering confidence measure, we take a look at the clustering plot. An error plot of the clustering approach is shown in figure 6.9. It shows four lines representing the fusion error rate, session rewrite rate, error rate of session ID labels, error level of the confidence estimation (i.e. 1 minus confidence) and the correct fusion detection rate. The error rates are calculated as follows. The *fusion error rate* is the percentage of sessions that are clustered incorrectly (i.e. if the cluster contains a label that does not match). The *session rewrite rate* is calculated the same way as the fusion error rate, with the exception that the ground truth is not the human annotation but the labels acquired by the system. The *session ID label error rate* is the number of sessions with incorrect labels that are not clustered together with the correct label, divided by the number of all sessions. The *correct fusion rate* is the number of correctly fused sessions divided by the number of all sessions. A session is counted as fused correctly if the sessions cluster contains more than one session, and all sessions are from the same person.

The figure also shows that the confidence, trained by logistic regression with cluster mean distance as input features, provides a reliable measure of when to stop clustering. The critical area, when the confidence increases is marked red. Most knowledge base errors with a low reverse confidence happen because of incorrect assignment of labels during learning.

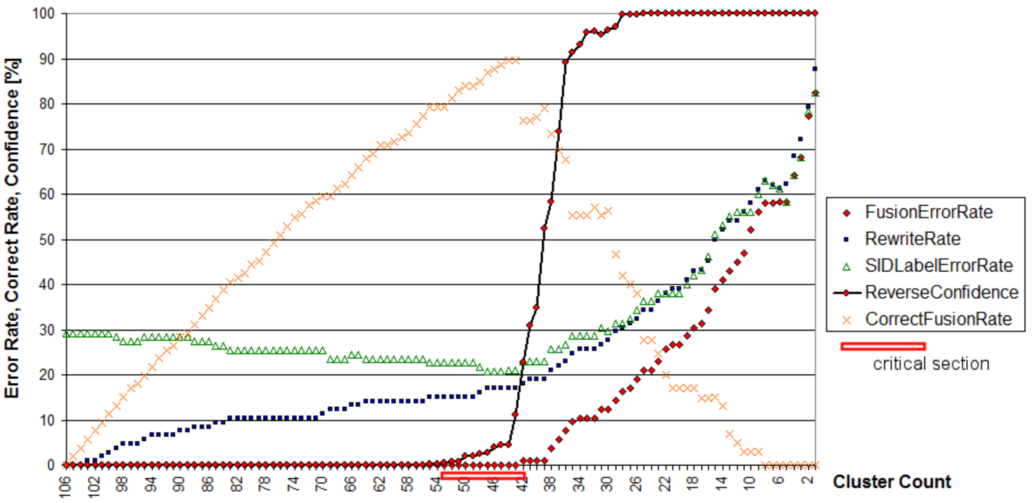


Figure 6.9: Analysis of automatic clustering of sessions and evaluation plot of clustering confidence

In most cases, these errors can be resolved completely by the system, as can be seen in the following evaluation. For example, when a person has spoken multiple times with the system, labels which have falsely been learned but fall into the same cluster with correctly learned labels might simply be overruled. Just to remind: In some situations it is safer to discard data from the problematic sessions (instead of re-labeling problematic sessions with the overruling label of a cluster), as to prevent vision models from getting corrupted by false samples. Many of the remaining errors match the type of errors that can be resolved by the dialog interactions. For example, some errors are due to incorrect name spelling, which is predominant in the visitor category, but infrequent in the employee category due to good prior vocabulary models.

6.7.6 Offline Problem Detection

The second step during offline processing is a per-cluster 'label mapping'. It represents a weighted list of labels per cluster, where weights are computed as defined in equation 6.2, where a cluster is a set of sessions and the label of a session is the personID label, which has been assigned to the session by the system.

$$weight_j(A) = \frac{freq_j(A)}{|cluster_j|_s} \tag{6.2}$$

where

$$cluster_j = clusterwithindexj \tag{6.3}$$
$$|cluster_j|_s = numberofsessionsincluster_j \tag{6.4}$$
$$freq_j(A) = |\{sid|label(sid) = A, sid \in cluster_j\}| \tag{6.5}$$

The final step of creating a list of problematic labels is to calculate scores, which are now independent of the clusters and sessions. To rank the problematic labels, different problem detection scores are used as defined in the following, which are then combined by standard rank-list fusion (low ranks indicate problems). Rank list fusion was chosen as the scores address separate problems, and the scales of these scores are not directly comparable. The rank-session score (equation 6.6) gives a low rank to labels which are recognized in a small number of sessions, as a large number of sessions indicates that the name has often been confirmed. The rank-entropy score (equation 6.7) gives a low rank to labels which occur in clusters with high variation of labels without dominance of a single label. The rank-entropy for a label is only calculated on the cluster which contains the largest number of sessions of that label, i.e. for the label's main cluster, as otherwise the metric is diluted by influence of other clusters. The rank-confusion score (equation 6.9) gives a low rank to labels which can be confused with a large number of other labels. The rank-time score (equation 6.10) gives a low rank to labels which have recently been added.

$$rank_{sessions}(A) = |Sessions(A)| \tag{6.6}$$
$$rank_{entropy}(A) = entropy(argmax_j freq_j(A))^{-1} \tag{6.7}$$
$$= -(\sum_{L \in labels_j} weight_j(L) * log_2 weight_j(L))^{-1} \tag{6.8}$$
$$rank_{confusion}(A) = |LabelMap(A)|_l^{-1} \tag{6.9}$$
$$rank_{time}(A) = lastseen(A)^{-1} \tag{6.10}$$

The function *LabelMap* refers to a global label mapping, computed over those clusters, which contain a session that is annotated with the desired label. The remaining functions are defined as follows:

$$Sessions(A) = \{sid|label(sid) = A\} \tag{6.11}$$
$$labels_j = \{L \in labels|freq_j(L) > 0\} \tag{6.12}$$
$$LabelMap(A) = \{L \in labels|\exists j : freq_j(A) * freq_j(L) > 0\} \tag{6.13}$$

task	**treated error**
merge two person entries	one person has several entries: e.g. Stephan and Stefan
discard old entry	deprecated entry: e.g. visitor who does not return or employee who has left the lab

Table 6.5: Tasks for error resolution with offline processing

The resulting per-cluster label mapping can be interpreted as automatic error correction, if the dominant label in a cluster is signed over to the other sessions in the cluster or if conflicting sessions are removed. In this case the offline processing is a fully autonomous error correction approach. Experiments, presented in the evaluation chapter 10, show that with pure offline processing, indeed a significant number of errors can be fixed, but even better and more reliable results are obtained by combining the offline processing with interactive mending dialogs. Examples of problems that can be solved without interaction are shown in table 6.5.

6.8 Background Knowledge Acquisition

Background acquisition of new information fulfills three purposes (see also table 6.1). The first purpose is creating an initial knowledge base ('init' phase). It is based on manually predefined information and as such it describes a controlled search task or data mining task and implements the insert function. The second purpose is to acquire information during an active interaction with the user. Information is provided during dialog, which exceeds information from the knowledge base, and a search task or data mining task is triggered to obtain information which can be used directly in dialog. It does not necessarily implement any of the update functions of the knowledge base. More importantly, it provides background information for the interactive learning task. The third purpose, here referred to as the 'offline' part, is to conduct background updates of the knowledge base, while the dialog is inactive. It acquires new information (insertion) by similar methods as the init phase with modified query data collected during interactive learning. Different results than during init can also be due to modified searchable information, e.g. modified web sites or updated data corpora. It also modifies or deletes knowledge base entries. Information collected during interactions and knowledge base structure is analyzed to detect superfluous, incorrect or contradictory information. The practicable usage of this is that incorrectly stored persons can be deleted from the database, or information which is not required any more but harms overall interaction quality can be removed. The criterion how the knowledge base is modified depends on the specific metric. Different metrics can emphasize for example knowledge base quality, interaction quality, or user satisfaction.

6.9 Evaluation Metrics for an Autonomous Learning System

This chapter should be completed with a few notes on how to evaluate such a learning system and to give an overview of evaluation metrics that can be derived from the previous sections. Specific metrics will be described in the following sections, where applied. It is obvious that evaluation metrics are important to measure the success of an autonomous learning system. Since a system usually is developed to optimize a specific metric, such a metric obviously influences its behavior and functionality. The approach presented in this work applies a holistic approach, i.e. besides evaluation of single components, and evaluation of the system is conducted as a so-called "end-to-end" evaluation. In accordance to existing evaluation efforts in dialog systems, interaction-specific aspects need to be evaluated, i.e. quality of the dialog. To measure the success of learning, it is necessary to assess the quality of the knowledge base. Interaction-specific metrics inform about efficiency (e.g. "dialog success") and subjective perception (e.g. "user friendliness") of dialogs conducted with users. Metrics about the quality of the knowledge base inform about quality of the learning result. Both metrics cannot be optimized independently. Optimizing dialog success does not necessarily mean to also optimize knowledge base quality. Within an evaluation over time it can be shown what kind of effect a single dialog has on the development of the knowledge base. A holistic approach thus has to consider both categories.

6.10 Conclusion

This chapter has introduced a knowledge model for representing multimodal information, dialog modules, and background learning methods for updating the knowledge model. Objects are represented as complex entities, and it has been shown, how speech descriptions and semantics are defined for representation and learning. The dialog modules implement knowledge update functions for extending the knowledge model or to support knowledge mending. A knowledge mending approach has been presented as an unsupervised approach to detect problems in the knowledge base, which can be resolved automatically or through pro-active interactions with a human user.

Chapter 7

Knowledge Acquisition in Dialog: Objects and Semantics

This chapter describes learning of objects and related semantics in a human-robot interaction scenario. The system which is presented has been deployed to and tested on the humanoid robot Armar III. Also, the experiments presented in the following sections have been conducted on the same robot. The scenario for the system is a household environment for the humanoid robot Armar III, in which the robot is confronted with different everyday-life objects. Some of these objects that the robot encounters are unknown to the robot. The robot shall detect the unknown objects, acquire information about the object, store information in the knowledge base and recognize these newly learned objects in further encounters. The learning tasks are conducted during standard task execution of the robot, which are for example requests from a human to bring a specific object to someone.

The learning scenario and learning dialog resembles other learning tasks presented in this work in many ways. The robot needs to learn visual information about the object. The robot needs to acquire verbal information, including object names and properties, also in combination with new words learning. Visual information is required to recognize the object again in the environment. Verbal information is required to understand when the object is referenced by the user, or to produce spoken output by the robot. Different from the learning dialogs presented so far, a new dialog approach for learning semantics is presented here, which addresses the special requirements of acquiring object semantics.

The presented approach and experiments have been conducted with an integrated dialog system including several learning modules for acquiring semantic knowledge, learning new words in speech recognition, and integration with visual object recognition and learning.

The experiments for learning objects are organized in this chapter as follows: Section 7.1 describes system architecture, integrated components, and overview of the knowledge model for interactive learning of objects. Section 7.2 describes detection of unknown information and new words acquisition in dialog. Section 7.3 presents an algorithm to symbol grounding for assigning a semantic category to an unknown object in dialog. Section 7.4 describes experiments, evaluation and results.

The experiments presented here have already been published in Holzapfel et al. (2008b), this chapter is slightly modified from the original publication. The publication has been written together with Daniel Neubig during his Diplomarbeit at the Universität Karlsruhe (TH), which has been supervised by the author of this thesis. The parts on new words acquisition have partly been published in Holzapfel et al. (2007) in cooperation with Thomas Schaaf and the speech recognition part has further been developed in a Diplomarbeit at the Universität Karlsruhe (TH) by Stefan Ziesemer, also supervised by the author of this thesis.

7.1 System Overview

Our approach for interactive learning of objects integrates several knowledge sources with the following aspects:

- *semantic information* about the object is acquired in dialog. Semantic information covers the type of the object and several properties.
- *verbal information* and *descriptions for spoken reference* can be acquired for a new object, which includes introduction of new words.
- *visual information* is acquired during dialog for grounding of internal object models in the real world.

In contrast to existing work, as laid out in section 2.4, the presented approach addresses an integrated system and covers all of these aspects within one system, also allowing new words and deep semantics from a structured ontology. Though this chapter does not intend to provide a fully developed theory of how objects should be learned over a longer time period by a humanoid robot, it rather intends to study a first integrated system and test application of the more generic dialog-based learning approach to object learning, including an algorithm for acquiring semantic concepts for object types, usage and properties.

The learning scenario is Armar III in a household environment, with objects from the kitchen and more general household items. Before explicit learning by 'learning dialogs' can be initiated, certain triggers are used to determine when an unknown object has been found, e.g. by the object recognition component, or when unknown words occur. Such a 'learning dialog' can acquire information for known or unknown objects, and clarify information. Learning covers new semantic categories, new descriptions for existing

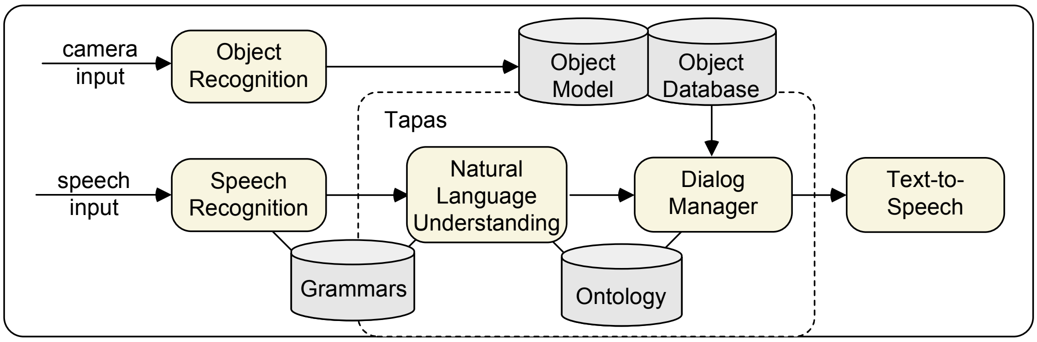


Figure 7.1: Integration of the objects recognition system in the dialog system

objects including new words, learning of object properties, and association with visual object IDs. These tasks implement the insert and modify functions, for complex knowledge entities, CLASS, MMID, new labels for existing attributes, and new attributes.

Experiments reported later in this chapter have been conducted on the humanoid robot Armar III, which is described by Asfour et al. (2006). Conducting the experiments with the humanoid robot Armar III, leads to a typical human-robot interaction scenario, which defines the type of interaction, and defines the perceptual system for our approach. While from a technical point of view, the humanoid robot is only used as a perceptual system which can go to and look at different places, users reported that interactions with the humanoid robot is fun, and the robot represents a communication partner they can talk to. Using the humanoid robot also serves as a proof of concept that the approach works on the target platform.

Figure 7.1 shows the integration of the different components within the dialog system. Dialog management is handled by the Tapas dialog tools as introduced in the first part of this thesis. In comparison to other multimodal fusion approaches presented so far, this setting uses a loose coupling scheme for object recognition, where recognition results are interpreted as high level events by the dialog manager. The central component for this approach is the dialog manager which conducts the learning strategy and interaction with the user. The dialog system setup is similar in all learning scenarios. The main difference here is the integration of an object recognizer. As in other scenarios studied in this thesis, the system setup includes speech recognition, unknown word detection and new words learning with the Janus speech recognizer. The unknown word model is integrated into the context-free grammar, which also gives information about a possible semantic meaning of the OOV, based on grammatical construction of the utterance. The models for recognizing new words for objects and new person names are identical, and both can use dynamic vocabulary approaches, as described for name recognition in section 8.1. The speech recognizer is embedded in the Tapas dialog system as described in chapter 3.

Visual processing uses stereo vision from the robotic head's cameras. For visual processing, detection and recognition of objects, we have integrated an object recognizer provided by Azad et al. (2007) and the software toolkit IVT[1]. Though visual object recognition is not the main focus of this work, we want to give a brief description of the recognizer's functionality to the extent that is necessary to follow the experiments. It can recognize textured objects using SIFT features (Lowe, 1999), and untextured objects using 3D shape models and color. Because learning of 3D shape models requires complex modeling and scanning of the object from different angles to observe its structure, this approach is currently not realistically applicable for interactive learning in real-time. Rather, the use of SIFT features allows to learn an object from features extracted from a single image taken from the scene with stereo vision, during the learning dialog and in real-time. Another advantage of this approach, is that the object's features are mostly independent of scaling, angle of view, rotation, light conditions and their position in the input image.

The object recognizer is able to recognize objects and detect unknown objects in real-time, which is triggered by the dialog system. For learning of new objects, the object recognizer can store acquired visual features, together with a given label during runtime, such that the object can be recognized immediately after learning. The label is generated by the dialog system and represents an internal 'ID' that is used to identify an object instance. The visual features are automatically segmented from a scene, using stereo vision, depth information and occurrence of visual features. The features for unknown object detection are kept in memory, until a decision is provided by the dialog manager to store the unknown object or to discard the features. More details regarding the visual object recognizer can be found in Azad et al. (2007).

7.2 Learning in Dialog and New Words Acquisition

7.2.1 Detecting Deficient Information

A dialog for learning is initiated by the system during normal interaction, when the system detects deficient information. In the scenario addressed by

[1] Integrating Vision Toolkit - IVT: http://ivt.sourceforge.net

our system, the goal of most dialogs is to instruct the robot to do a specific task. A typical task-oriented dialog is conducted when the user instructs the system to bring a specific object, serve something to drink, or put something into the dishwasher. Within such dialogs we have extracted two categories of deficient information.

- the user input cannot be understood correctly by the system given verbal information
- the specified object cannot be found, or an unknown object is detected

The first case addresses speech recognition and understanding, the second case addresses visual processing of objects in the environment. Both cases can serve as so-called “deficiency detectors”.

Deficient information in vision occurs when the object specified by the user cannot be found, or when an unknown object is detected by the system. In either case, the system first needs to detect an unknown object, i.e. obtain visual features for an object which is referred to by the user. If the system does not detect an unknown object, it cannot store any features, and therefore cannot learn information about the object. Thus the detection of features and, together with that, segmentation of the object’s shape are prerequisites for the learning process. In addition to feature detection, the object recognizer uses 3D information for object segmentation. Thus the robot can learn the object when it is held in front of the robot’s camera, as shown in figure 7.2 in the leftmost image. The object can also be learned from visual features only, when no 3D segmentation is possible and the background does not have rich texture, as is shown in figure 7.2 in the rightmost image. For the experiments described in this chapter, objects have been put at a specific location, next to the sink. This way the test subjects did not have to pay attention to where to put the object so that the robot can find it again and comparable dialogs could be produced. The objects where put on a black surface, with a standard kitchen background, e.g. parts of a cupboard and the sink can be seen in the pictures taken by the robot. In the experiments, the objects’ shapes could be segmented reliably from feature clusters only.

Deficient information in speech recognition occurs when the user produces input that cannot correctly be recognized by the system. Gieselmann and Stenneken (2006) describe different error situations that occur in human-robot interaction, for which data from text-based interactions and interactions with the real robot have been analyzed. The largest number of miscommunication errors occurs due to new syntactic and semantic concepts, i.e. new formulations, new objects, new goals, and meta-communication. In cases of unknown objects, user input typically leads to sentences that are not

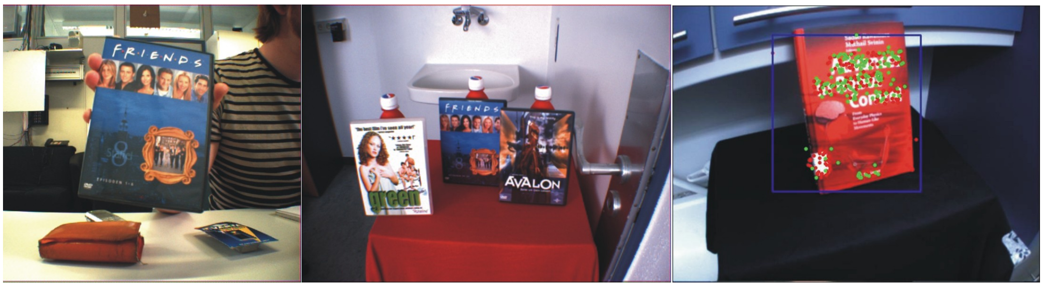


Figure 7.2: Snapshots taken from the robot camera. From left to right: object held in front of the robot's camera, multiple objects recognition, unknown object recognition during the experiment with feature extraction and shape segmentation.

covered by the grammar. As described earlier in this chapter, the grammar is created automatically from database entries, so that only attributes describing known objects are covered by the grammar. This has the advantage that speech recognition performs well for known utterances, but the disadvantage that new formulations are not covered by the grammar. To prevent this problem, the standard approach in speech recognition would be to extend the vocabulary until all words which have to be covered are contained in the vocabulary. However, in case of object names it is not clear which words need to be covered by the vocabulary in advance, since unpredictable words can occur. In speech recognition evaluation this effect is typically very small, since the standard word-error-rate (WER) is hardly affected, if once in a while, a word cannot be recognized. For the robot in turn, exactly these words can be very important. To show the effect of WER let us consider the example "please open the granini juice for me" which has been used previously. If the word 'granini' (let this be an unknown word) is misrecognized, the WER is affected in the same way, as if the word 'please' was not understood. However, in the first case, the main information for disambiguating the object in the environment is lost. Extending the vocabulary with a very large number of possible words is not a good option, since speech recognition rates for known objects would drop drastically. However, approaches are known to detect unknown words in speech. We use out-of-vocabulary words (OOVs) which are recognized by the system when an unknown word has been spoken. Our approach uses an implementation of so called Head-Tail models from Schaaf (2001) for detection of unknown words. Given an example sentence, which contains the command to switch on an unknown object, the grammar might recognize: "please open the OOV juice for me". Here, speech recognition detects an unknown word, which is encoded as OOV. For the detection, both language model scores (defined by the grammar) and acoustic scores (acoustic speech recognition models) are considered. The example sentence also gives us a first hint about the semantic category of the unknown word by observing verb-object subcategorization information, by the semantic frame given through the grammatical construct. Using OOV models has originally been studied for n-gram models. In Holzapfel et al. (2007) this approach is also described for usage with context free grammars for the recognition of unknown names. The same approach has been adopted for the present system. Following this approach, unknown words can only occur at specific positions in the grammar. The used grammar formalism defines OOV ('oov') symbols in the grammar in the following way. For example a noun phrase describing an object can be formulated as

```
public <obj_object,NP,_> =
            oov |
            <obj_juice_db>|
            <prp_juice,A,_> <obj_juice_db>;
```

Here, 'oov' replaces a full noun phrase. In analogy, the OOV can also replace a property, syntactically represented as an adjective or a noun.

7.2.2 New Words Learning

Once unknown words have been detected in the utterance, these words can be learned by the system in dialog. During the experiments, these words are either properties of objects, object types, or part of the object names. In addition to the dialogs to obtain semantic information of the object, which is described in the next section, the system needs to acquire spelling and phonetic information of the word and update the speech recognizer's vocabulary, dictionary and language model. A pronunciation for a new word is generated with a grapheme-to-phoneme converter, which is available with text-to-speech tools, such as Festival or Cepstral. Both a grapheme representation, which is obtained e.g. from spelling, and the phoneme representation are needed to update the speech recognizer's dictionary. In addition to the dictionary, the shared recognition and understanding grammars of speech recognizer and dialog manager are updated. Both can be extended on the fly, and are updated during dialog, once the new word has been confirmed by the user.

7.3 Interactive Learning of Semantic Categories for Objects

7.3.1 Learning Object Properties

The algorithm for learning properties and semantics of an unknown object includes obtaining a description from speech and clarifying properties with their values and semantic types, which is done in a dialog with the human. The dialog for learning properties allows the user to formulate any property of the object which he thinks is useful. The system already understands different values, such as color and size. Other properties, such as title or name (e.g. a DVD has usually been referenced by its title), are restricted to names stored in the database. When the user formulates a description which is not covered by existing property values, the speech recognizer can detect this as unknown words and reports an OOV detection to the dialog manager. In the case of OOV detection, the user is asked again to say only the property of the object, since additional repeats increase the chance of understanding the word correctly. If the word cannot be understood correctly, which is determined by obtaining feedback from the user, the unknown word can also be spelled by the user. The user is only asked for spelling, if the OOV-part of utterance is relatively short (which is determined by phoneme recognition on the utterance). For spoken output, standard grapheme-to-phoneme rules of the text-to-speech synthesis component are used. If the user confirms the word, it is then learned by the system, by adding the word to the speech recognizer's dictionary, and to the speech recognition and understanding grammars. The new word can then immediately be used within the same dialog. For better understanding of several words which form the title of an object e.g. 'a book on advances in robot control' an additional speech recognition module with n-gram language model and a large vocabulary can be used.

A new word learning dialog is also initiated when the user refers to an object, e.g. "bring me the red cup" and an OOV is detected for the utterance. In this case the system first needs to find out whether the unknown word is part of the object's type description or if it represents a property. The learning dialog then is conducted as described above.

7.3.2 Learning an Object's Categories

Learning of object types is conducted with an approach that combines open input by the user, who can name a category, and a prompted mode which implements browsing through the ontology. In the open input mode, the user can name a category which he would use to classify the object. The open

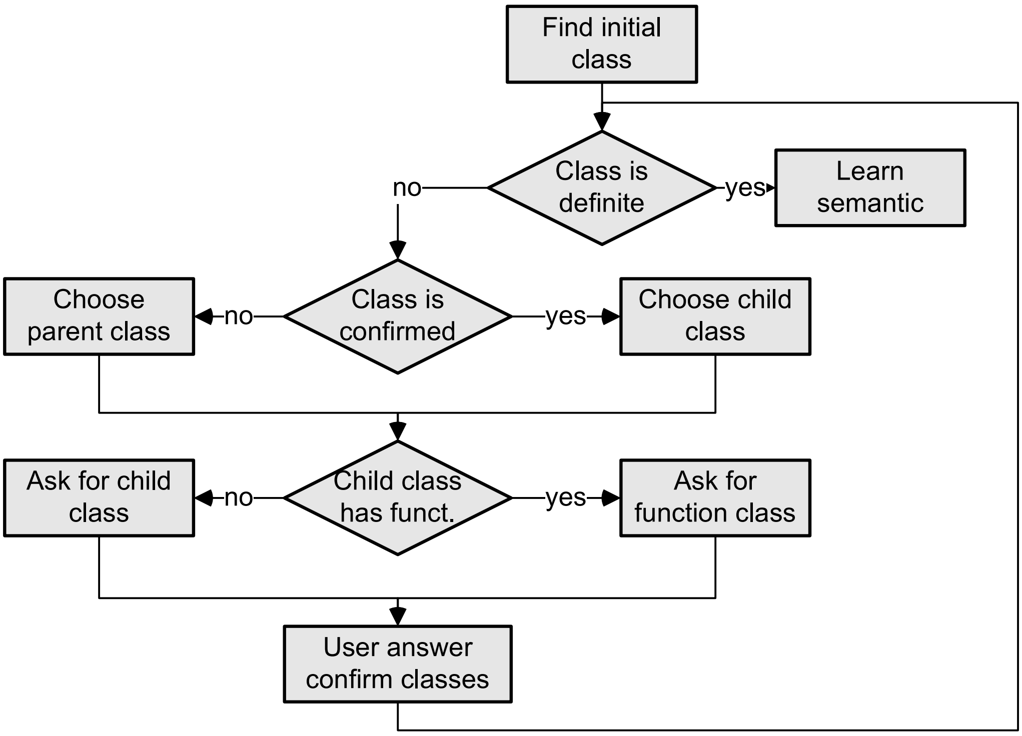


Figure 7.3: Learning scheme to acquire semantic categories for an object and dialog flow

input mode is also referred to as one shot learning, since one input by the user is enough to describe the category. A simple one shot learning dialog follows the example:

system: What type of object is this?	open question
user: It is a juice	type is set to juice
system: Did you say that the object is a juice?	confirmation
user: Yes	type confirmed

One shot learning has the advantage of quickly obtaining a hypothesis for a category. Drawbacks are that it is not necessarily obvious to the user, how the robot's internal object hierarchy is structured and the user does not know what the system can understand. For example, it was observed that functional categories pose even stronger problems to the one shot learning approach than object types. As a reply to the question "what can you do with this object" some persons replied with very complex statements, and some had to think for some time before they could come up with an answer. Thus, in the present experiments, open questions are only asked regarding the type of the object, and functional classes can be queried by system initiative only. For example the system can ask “can you eat this?” or “is this edible?” when asking for the functional concept *eatable*. Thus, the dialog is improved, when the system can choose the wording. The browsing mode addresses exactly this problem, and can choose from questions for object types and functional classes for disambiguation. It starts at a base category and iteratively tries to classify the object as one of the subclasses of the current category. This way, the structure of the ontology can be communicated and input by the user is restricted to a smaller set of possible meanings than in the open input case. Drawbacks of the browse mode are that this mode can be tiring for users, and that for large ontologies, descending the hierarchy can even take too many turns to be practically applicable. An example of the browsing mode is as follows:

system: is the item a kitchen object?	ask type
user: yes	type: kitchen_object
system: can you eat this object?	ask function
user: no	type: kitchen_object
system: can you drink this object?	ask function
user: yes	type: drink
system: is this a hot drink?	ask type
user: no	type: drink
system: is this a juice?	ask type
user: yes	type: juice

The combined approach begins with a single one shot approach and then gives the opportunity to refine the category be browsing the neighborhood. The dialog to conduct this strategy begins with a question to specify the class of the object (open input). The input is confirmed. If no children of the class are found in the hierarchy, the dialog ends here. Otherwise the robot switches to the browsing mode until a leaf node has been found in the hierarchy, or no further refinement is given by the user. The questions in browsing mode address children of the selected type or functional concepts to disambiguate subclasses and are formulated as yes/no questions. Figure 7.3 depicts this algorithm in a flow diagram. The start-node named “find initial class” represents the one-shot learning node. After the one-shot learning, the learned class can be refined by browsing the ontology’s type hierarchy or functional concepts. After posting one question to the user and a confirmation response (bottom node in the flow diagram), the loop is entered again. The combined approach makes sense because of several aspects. (i) Due to speech recognition and understanding problems the desired category cannot be understood. (ii) The user does not know the category description used by the system. (iii) The user communicates a category that is too general, e.g.

'drink'. This general category can then be refined to obtain a better model.

7.4 Experiments and Evaluation

7.4.1 Experimental Setup

For evaluation of the approach, experiments were conducted with the robot in the kitchen environment. The users could, for example, command the robot to bring a specific object from a location, or report which objects he can see at a specific location. The robot knows about several locations from its environment model, such as the sink, sideboard, stove, cupboard, fridge, etc. The robot can also understand directions such as "next to the sink", "left side of the sideboard", "in the middle of the sideboard", "in the fridge", etc. For identifying a requested object (grounding), the robot can ask for the location, which can be given by speech or using pointing gestures. If multiple objects are found at a location the robot conducts a simple dialog listing all known objects to clarify which object is unknown. If there is more than one unknown object, the user would have to move the object and present the object to the robot e.g. by holding the object in his hand as shown in figure 7.2 in the leftmost image. For the sake of obtaining comparable dialogs during the presented experiments, the setting was restricted to the sink location, with at most one unknown object and grounding restricted to speech. In case there is an unknown object at the given location, the robot ideally would ask the user to help him to learn the object and identify the object's properties. If an unknown object or unknown words occur during the interaction, learning dialogs are initiated by the system as described in the previous section.

The experiments comprise 52 dialogs which were conducted with six naive users - who have not interacted with a robot before. The goal of these dialogs was to have the robot serve a specific object or get information from the robot which objects he can see at a predefined position in the kitchen. Each of these dialogs includes detection of objects at the sink location. When an unknown object or an unknown word is detected, the learning dialogs were initiated. This way, a dialog could be very short (if only known object), in this case these dialogs are used to evaluate detection rates. Or, the dialogs took as long as required to reach the learning goal. For example, learning an object's property does not always include learning a new word. In this case, these dialogs are used to evaluate the different learning tasks of properties and concepts.

The users did not know in advance, which objects were known to the robot, and which objects were unknown. The interaction started after a

	total	category 1	category 2	comment
dialog condition		unknown	known	*dialogs with known*
#dialogs:	52	40	12	*and unknown objects*
unknown detection		correct	failed	*interaction by the user*
#detections:	40	39	1	*in 5 cases*
known detection		correct	failed	*interaction by the user*
#detections:	12	10	2	*in 2 cases*
detection summary		correct	failed	
#detections:		49	3	

Table 7.1: Overview of the experiment and recognition rates of visual object recognition

brief introduction about the scenario and the robot's task. No details were given about how the robot performs its learning strategies to prevent biasing of the users. The dialog started with a greeting or directly with a request from the user to either serve a specific object, or to report which objects the robot could see. The following evaluation section describes results, success rates and recognition rates from these dialogs.

7.4.2 Evaluation

Meaningful numbers for the experimented scenario of interactions and learning dialogs are success rates (number of successful dialogs) and dialog length (measured in number of turns). The first metric is important to measure the effectiveness of the approach. The second metric is important to measure the efficiency and burden for the user. Numbers are reported here for learning object categories and object properties for unknown objects. Also, a comparison of different learning strategies for object categories is made.

An overview of the experiment conditions and conducted dialogs is shown in table 7.1. The table shows a total number of 52 conducted dialogs, the separation into known (12) and unknown (40) objects conditions, and detection rates of known and unknown objects. A closer look at the different categories shows that out of 39 objects that could correctly be detected as unknown objects, five objects required interaction by the user. The same situation happened in the known condition, where two objects required interaction by the user. Interaction by the user means that the object could not be detected upon the first try, e.g. because the object was completely or partly out of the robot's field of view. The users then turned the objects into

task	#dialogs	success	avg turns
learn object property	40	83% (33)	1.8
- with known words	25	87% (22)	1.4
- with spelling	15	74% (11)	2.6

Table 7.2: The three learning tasks and successful completion rates in the experiment.

the robot's field of view after which in all these cases, the object was classified correctly. To further analyze the errors that were made by the system, one can look at the failed attempts, which sum up to 3 out of 52. The reasons for failure were that in one case, visual features were not sufficient for detection, and in two cases, known and unknown categories were confused.

These requests provided the basis for the evaluation of the learning algorithm in dialog. Learning of an object according to the algorithm described above includes learning of the object description for reference in speech, properties of the object, and the type of the object. The description of the object however, is a combination of object properties and the type of the object. For example, the 'red cup' is an example of combining the type of the object (the cup) with a property of the object (red) to create a description that can be used in speech (see section 6.3.2 for details). The first part was to understand properties of the object. In the second step the type of the object was narrowed down in more detail. The two parts are addressed by the different learning dialogs described earlier, and are evaluated separately. Table 7.2 shows the number of dialogs, success rates and average number of turns of dialogs conducted for learning of object properties. Learning of a property value was possible in two ways. Either the word was known (25 dialogs) or the word was recognized as unknown, in which case the word could be spelled (15 dialogs).

The more complex learning task was to learn the semantic category of an object, for which 34 dialogs were conducted. In 82% of these cases, the dialog could be completed successfully with the learning algorithm that applies the combined approach. The combined approach was applied in all 34 dialogs. From the conducted dialogs, comparison can be drawn with the one shot learning approach and the browsing strategy. The combination of different possibilities, how a class can be learned by the system, resulted in different combinations of one shot learning and browsing. In 47% of the dialogs, the class was specified directly by the user, and could be learned directly as a pure one shot learning. After the one shot attempt, the dialog was stopped by the user. The same number of dialogs (additional 16 dialogs) was conducted,

task	#dialogs	success	avg turns
One Shot Learning (47%)	16	81% (13)	2
Browse (6%)	2	100% (2)	10.5
Combined (47%)	16	81% (13)	4.2
all one shot (100%)	34	68% (23)	2
all combined (100%)	34	82% (28)	3.6

Table 7.3: Application of one shot learning, browsing and the combined approach during the experiments for acquisition of the semantic category.

where the class was refined after the one shot learning step. The remaining 6% of the dialogs was conducted as pure browsing of the ontology, after the one shot learning approach did not result in a recognized type that could be used for browsing. The browsing dialog then started with the most general class in the hierarchy. This way, the user could complete the dialog quickly with one shot learning within only two turns, if it was clear to him how to categorize the object.

Table 7.3 shows the figures and results from the learning dialogs for acquisition of the semantic category. The top three rows give the numbers for the three approaches as conducted in the experiment. The table shows the number of dialogs conducted for each strategy, the rates and numbers of successful dialogs and the average number of turns per successful dialog. Since the combined approach starts with a one shot learning hypothesis and then refines the class in further step with a browsing strategy, comparison can be drawn between one shot learning and the combined approach on all 34 samples. The number of all successful dialogs with the combined strategy is the sum of all successful dialogs. In case of the one shot learning approach, the two cases which could be learned only with the browsing strategy are classified as failures for the one shot learning approach, since no category could be identified. In addition, 3 samples of the remaining dialogs would not report an acceptable result after the one shot learning step.

7.5 Results and Discussion

The presented approach for object learning is able to detect deficient information in dialogs, and initiate a learning strategy to acquire information for learning unknown objects. The experiments show that acquisition of the semantic category is an important but non-trivial task and significant improvements can be achieved by intelligent strategy design. It can be concluded from the evaluation that the approach is adequate for learning of unknown objects, both for the learning phase and recognition accuracy of the learned objects. The dialog modules for learning integrate with the architecture presented previously and shows extensibility of the approach also to object learning and semantic category acquisition.

The approach has been tested with speech recognition experts and naive users, and has been evaluated with naive users, who are not familiar with speech recognition. Especially during the experiments with naive users it can be observed that simple dialog structures are more successful than complex dialog structures, which seem to require too much prior knowledge by the user about how the system processes information internally. The dialogs in these experiments have been purely task oriented, and the modules for object learning were not integrated with strategies for social interaction. In these experiments this was accepted by the users, mostly because they mainly were interested in achieving the learning goal. Other experiments conducted for this thesis suggest that when a system is used more often, the users wish for variability and social interaction.

In comparison to related work, the approach presented here implements a full learning system that covers the three categories new words learning, visual features learning for real objects, and learning semantics. However, it should be noted that the approach presented here does not intend to build a knowledge base completely from scratch, e.g. as the approach by Roy (2003). Rather, the approach is used to extend an existing knowledge base with new information and categorize unknown objects within a mostly predefined knowledge structure.

The system is able to learn new words, properties and types of objects. Both, properties and types of objects are important to learn since both contribute to the description of an object, which is used by users to reference an object. During reference to objects, different properties are specified by the users. The speech recognition and understanding grammar thus supports a variable combination of different properties and types for each object. Since objects are categorized with different levels of abstraction, it is necessary to model functionalities as separate concepts in the ontology. The robot can then distinguish different functionalities of an object, which can be given from context in speech or from the description of the user.

The combined approach for learning of object classes has shown better success rates than pure one shot learning. It requires only little more interaction with the user (in terms of number of turns), but it produces significantly better results in categorizing the object according to error rates and accuracy. These first results show that the algorithm provides an accurate means to categorize unknown objects in terms of semantic categories within an ontology.

In contrast to pure recognition output of the object recognizer, employing dialog capabilities significantly improves the final recognition results after confirmation. The dialog uses implicit and explicit confirmation strategies, which both give the user the opportunity to interrupt the robot and correct the recognition hypothesis in the case of errors.

The evaluated system is able to categorize and learn new objects in dialog with the user. The resulting knowledge base allows the system to recognize the detected object, talk about the object and understand when the user refers to the new object in speech. However, there are also restrictions of the approach, and there are different directions for future research in this field. Further work could be directed at combining understanding approaches, such as the one presented here, with knowledge acquisition how the robot can manipulate the object. To do so, first, additional perceptual information needs to be collected, e.g. to better segment the object's shape with 3D information acquisition. Integration with vision currently requires that segmentation of an object is possible, e.g. by 3D or feature-based segmentation, and that grounding has already been done, when the learning dialog is initiated.

Limitations of the presented approach, are that currently all object types and properties are modeled statically. To some extent, dynamic changes in the environment are reflected as properties that change over time, which is already covered by the ability to associate one object with different categories. Also different verbal representations can be associated with objects. However, the system does not cover dynamics in a way that a cup of tea only is associated with tea if it is filled with tea, and that it would be associated with coffee, if it were filled with coffee. Modeling such information requires extending the approach with a state model that keeps track of object properties, such as 'dirty', 'full', etc. Some other properties make only sense if they are interpreted as user-specific properties. For example a person's most favorite cup cannot be generalized as being the most favorite cup of everybody. But this generalization is indeed appropriate for some properties. For example, a red cup remains to be a red cup, or a book continues to have the same title, which does not change over time. For user specific properties, user ID information could be integrated as an additional variable to relate user specific properties to specific users. Another approach can be to correct wrongly stored information or discard information that leads to contradictions in the knowledge base but is not necessary for interaction with the user. To assess how the system evolves over time additional experiments are required, e.g. to quantify effects of storing objects at a wrong position in the ontology.

Another limitation is the restriction to variability of speech input using complex constructions. While the presented approach has aimed at building flexible understanding based on generic grammar constructions, the way that humans communicate with each other can include more variability. To give an example, in a pre-study to find out what kind of information can be obtained from the user, a human-human interaction experiment was conducted. One of the questions about a spoon was “what can you do with it?”. The answer to this question was “you can spoon something with it.”. Other answers to other questions ended in longer story telling that were too complex for state of the art language understanding and interpretation methods, which still require the system designer to build dialogs that restrict the user to conform to some kind of interaction style that can be understood by the system.

7.6 Conclusion

An approach for object learning by means of dialog-based learning has been presented, including detecting deficient information and conducting a dialog for learning unknown objects using a generic entity model, which incorporates multimodal knowledge sources. It can be concluded from the evaluation that the approach offers the user a method to teach the system new objects, which can then be used immediately afterwards in the communication. The usage of the new objects is more flexible than other approaches as the system acquires structured semantic information about the object during the learning phase. The presented combined approach of one-shot learning and browsing for semantic category acquisition achieves better results than applying one-shot learning or browsing, as one-shot learning leads to inaccurate results and browsing is too inefficient. Aspects that have been brought up in this work but are still open questions are how to best handle different user preferences, how to address contradictory understanding of ontological structures by humans and fuzzy categorization of objects. Furthermore, the presented approach focuses on persistent properties rather than automatic recognition of dynamic properties, e.g. the fill state of a cup.

Though object learning is still an open task, the presented work demonstrates advancements and allows to interactively acquire not only labels but structure information about objects and generic description of properties and use the newly created models for recognition. The experiments show that acquisition of the semantic category is an important but non-trivial task and significant improvements can be achieved by intelligent strategy design. The dialog modules for learning integrate with the architecture presented previously and have shown extensibility of the approach also to object learning and semantic category acquisition.

Chapter 8

The interACT Robot Receptionist

This chapter introduces a fully integrated receptionist robot that serves as the evaluation scenario of the studied dialog-based learning approach for learning over time in a social environment. The purpose of the receptionist robot is to automatically build and maintain a knowledge base of persons and model employees of the interACT lab, interACT students, and visitors. For this purpose, it is located in the corridor of the interACT lab building and engages in interactions with persons passing by and are willing to talk to the robot. The result is stored in a database and can be visualized on a Who-is-Who web page. This chapter describes the system setup and presents analyses of dialog interactions with the robot.

The general data-flow of dialog interactions, knowledge base updates and visualization as a Who-is-Who web page is sketched in figure 8.1. It facilitates proactive interactions to initiate dialogs and system initiative dialog strategies to obtain information from persons to improve the knowledge base.

Besides the technical introduction of the receptionist robot, this chapter presents user studies that have been conducted during the design time of the system to analyze social aspects and to analyze subjective user feedback. For the purpose of analysis and evaluation of the dialog interactions, single interactions are analyzed and user responses are assessed with quantifiable subjective feedback and a qualitative analysis of social aspects of the interaction. Analysis of single interactions can be used for evaluations of the system and its strategies. It can also be used to understand more closely how the interactions are perceived by the users, and to detect and solve possible problems of the system. As already introduced in the related work chapter (section 2.3.2), evaluation of a dialog system for learning involves a variety of different aspects, including objective measures, subjective measures and knowledge base quality. Objective evaluation of dialog success and dialog length has extensively been described in the previous chapters to evaluate different dialog strategies and user model approaches in the receptionist scenario, where already a few subjective measures have been included. Section 8.4 focuses in more detail on analysis of subjective user feedback. Analysis and evaluation of knowledge base quality are presented in chapter 10.

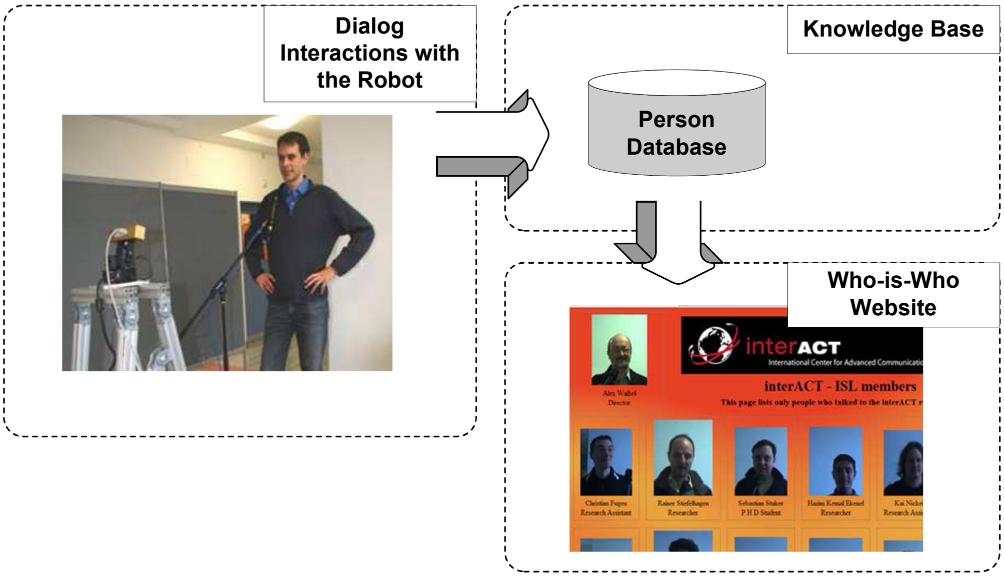


Figure 8.1: Overview of data-flow with dialog interactions, database and presentation on website

8.1 System Overview

8.1.1 System Architecture

The receptionist robot's system architecture comprises several components which contribute to its main purpose, to learn and maintain a database of persons in a social environment. The core of the system is the learning algorithm and the dialog system, which conducts the learning strategies and social interaction with users. Figure 8.2 shows a more detailed system and components diagram than shown in the introduction, and lists the main system components, knowledge models, processing levels and the general flow of data processing. As most of these components have already been introduced in previous chapters, the TAPAS dialog manager in chapter 3, face identification, voice identification, and multimodal user ID in chapter 4, unknown words detection and new words learning in chapter 7, dialog modules for dialog-based learning in chapter 6, here, we want to describe their integration and data handling. Input basically is asynchronously processed by the recognition components, and synchronized on higher levels of integration. Speech input is segmented by automatic segmentation as utterances, which are then processed by speech recognition and voice identification. Video input from a stereo camera head is processed by the Arthur multi-person tracking

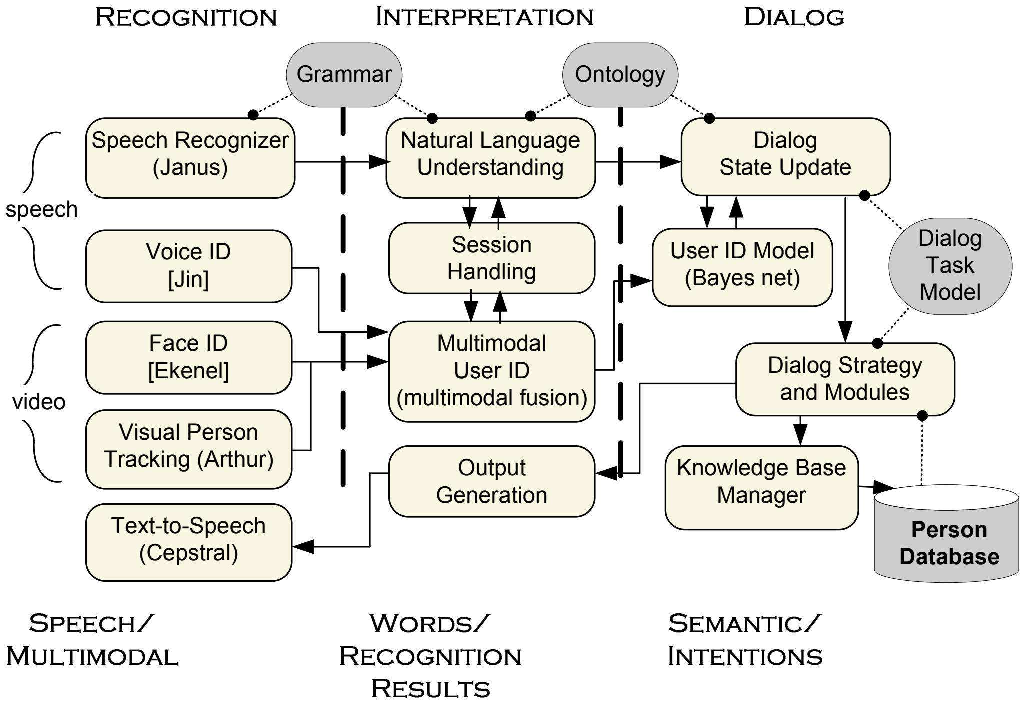


Figure 8.2: Component diagram with connection between the main components and main knowledge sources

software, as described in chapter 4, including face identification for each person track. Though the system is designed for a single person interacting with the system, additional persons appear in the field of view as well, e.g. when passing by. The multi-person tracker provides a robust method to filter out background persons by creating tracking hypotheses and face identification for each person. The continuous track of the person interacting with the system is locked by the *Multimodal User ID* component, which receives all track messages.

The interpretation layer contains natural language understanding, multimodal fusion of user identification events, where face identification and voice identification and track messages are synchronized, session handling, and response generation. The *Session Handling* component is necessary to segment input events on an interaction level, and create a notion of sessions. As already introduced in section 6.2, a dialog session is understood as a conversation between the system and a user and consists of a sequence of consecutive dialog turns. The session model is used for session-specific dialog variables, e.g. tracking of the user ID during the interaction, which affects updates of the knowledge base, and for recording and logging of session-specific data.

The highest level of integration is the dialog layer. The *User ID Model* integrates information from the *Multimodal User ID* component and information extracted from the dialog flow and provides a probabilistic model of user identification for the current session. The *Dialog Strategy* (as a general term of dialog module selection and execution of the module's strategy) decides on the next action and initiates knowledge base updates.

8.1.2 Database Setup

The *Person Database* in figure 8.2 represents the main knowledge base, which is maintained by the receptionist robot. Figure 8.3 shows the data structure of the database to represent personal information and social network structure. All entries are connected to the *Person* table, which contains the person's ID, first name and last name. The database model is shown as two main blocks labeled as *speech vocabulary* and *social network model*. The *speech vocabulary* block represents information that can be talked about using spoken interaction. The *social network model* block represents social network information, which is introduced in chapter 9. The *session corpus* block shows the connection of a dialog session with the identified person, acquired information, database updates and recorded data. It forms a history of all interactions and is used, for example, by offline processing steps such as knowledge mending, for which we present experiments in chapter 10.

8.1.3 Model Initialization and Spoken Name Recognition

The receptionist robot is started from scratch with an empty set of persons and an empty set of sessions in the database, as well as empty face identification and voice identification models. Information to initialize these models is obtained during runtime. Background models, e.g. speech recognition vocabulary and pronunciation dictionary, are initialized from the publications website of the interACT lab. Also social network analysis is conducted from publications information (discussed in chapter 9). Figure 8.4 depicts the initialization of the vocabulary models from website and telephone book for large vocabulary name recognition.

As laid out in chapter 2, recognition of names is a hard problem for speech recognition, as the number of possible names is significantly larger than the number of words that can be used in a speech recognizer for efficient and real-time recognition. If more and more names are added to the vocabulary without preprocessing, processing speed decreases and recognition rates get

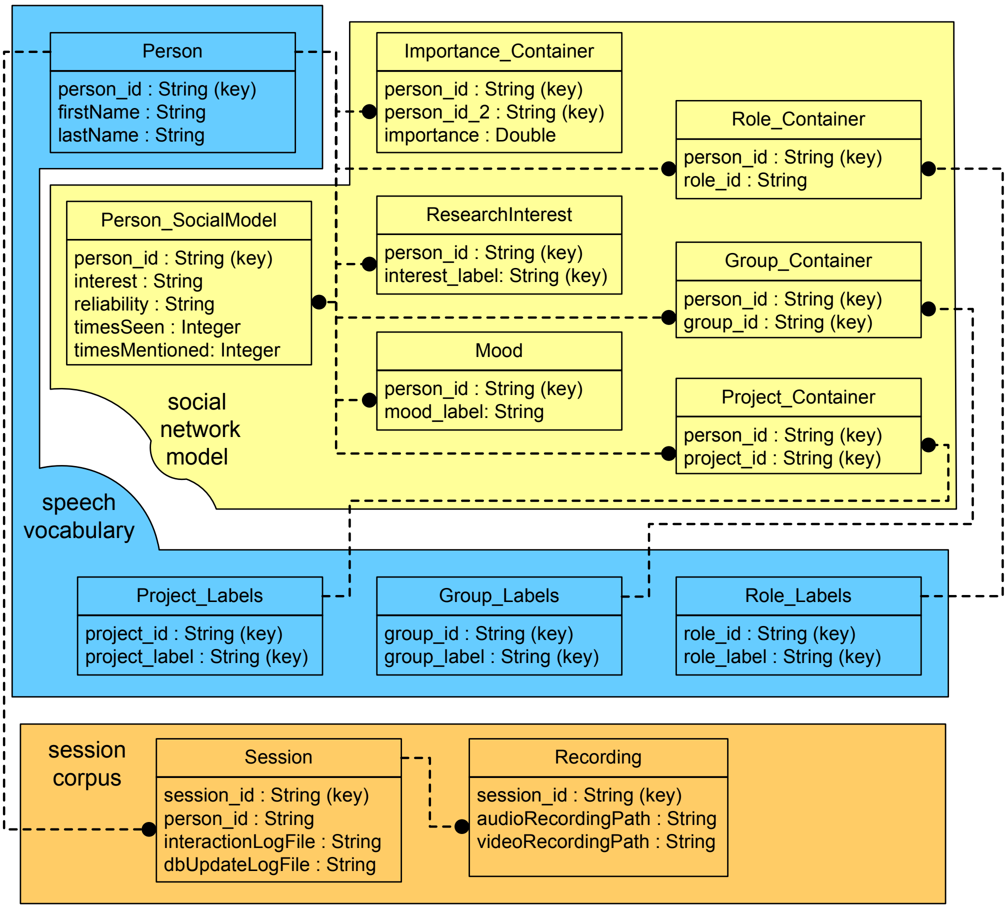


Figure 8.3: Data structure for personal information and social network structures in the database

worse, as confusion of hypotheses increases. We have addressed this problem by vocabulary selection using prior information and dynamic vocabulary recognition in a multi-stage recognition process, which improves both, processing speed for in-vocabulary names and recognition of out-of-vocabulary names with dynamic vocabulary switching. Experiments for the system have been conducted in a Diploma thesis by Ziesemer (2007). It could be shown that, depending on the probability of unknown persons, a two stage approach, which uses only names of known persons and OOV-detection in the first stage, and a large vocabulary in the second step provides better results than decoding directly on a large vocabulary.

Figure 8.5 shows a simplified example for grammar and language model

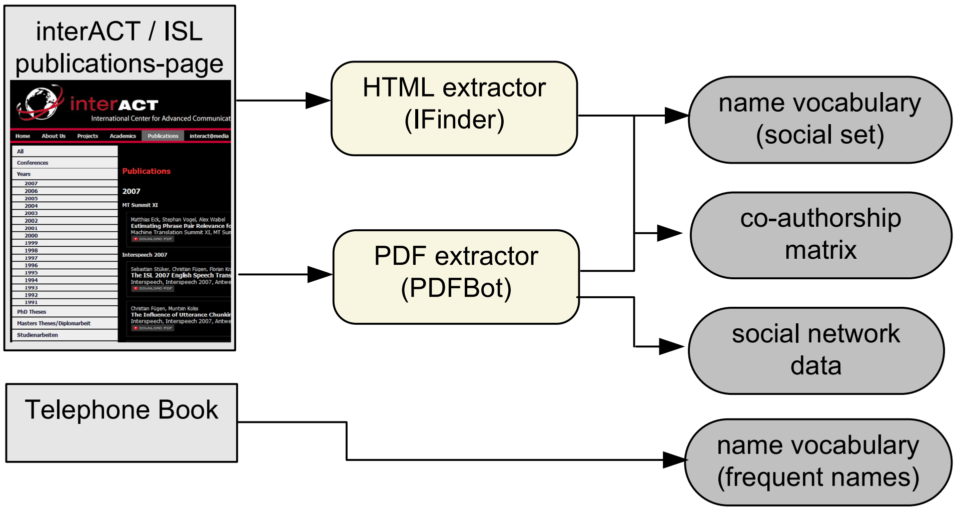


Figure 8.4: Vocabulary initialization from publications website

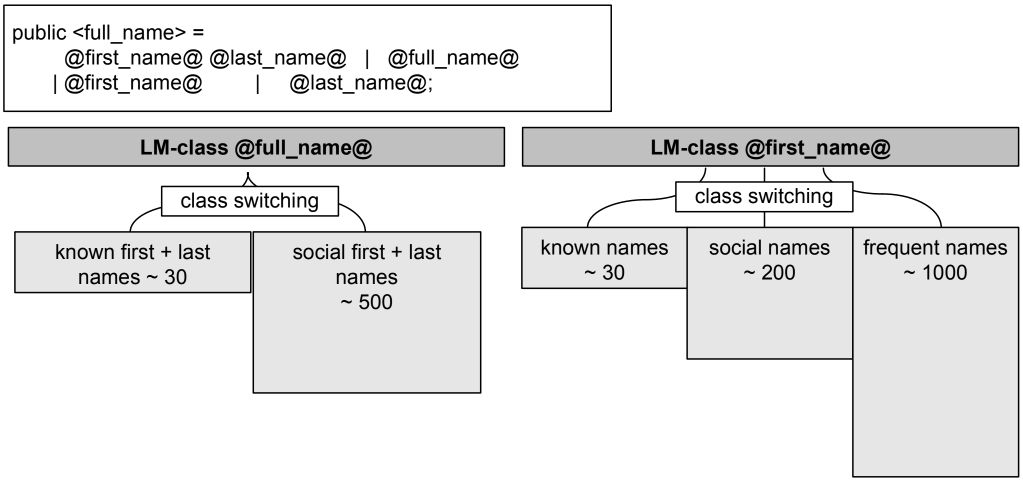


Figure 8.5: Dynamic vocabulary switching for name recognition

classes for demonstration of the vocabulary switching approach. Language model classes are often used in speech recognition, where the probability of a single word is hard to estimate, but a class of words can be estimated from data. A frequent example are navigation system and the problem of modeling probabilities of street names. Each single street name occurs too infrequently to provide good language model probabilities, and therefore the class of street names is used to estimate the probability. We apply this technique for person names, as the grammatical construct does not depend on the name, ‘John’ but only on the class, *@first_name@*. In addition, depending on the speech decoder implementation, speech recognition can be conducted more efficiently, as the language model probability is calculated only once for the name class.

The language model class *@first_name@* in figure 8.5 is reset dynamically during runtime by the respective vocabulary set. It shows one possible setup with three different sets of names from known persons (~ 30 names), names from social network analysis (~ 200 names), and the list of generally most frequent names (~ 1000 names), with OOV-detection on each layer, and re-decoding on the same utterance with the next set in case of OOV detection. A second solution is to merge the set of social names and the set of the generally most frequent with different language model weights into one set. This approach leads to the best recognition results if the system is confronted with a more or less balanced set of known persons and visitors, which are not covered by the social network results, e.g. students.

8.2 Dialog Setup

The dialog strategy is controlled by a modular dialog strategy with system initiative-based state switching. Dialog modules are understood as modular elements of the dialog strategy, as introduced in chapter 6.5. A state model is used to advance the dialog to the next state on system initiative, once a module has been completed, by executing a dialog move which leads over to the next module. For example, the *PersonID* state is entered by asking for the user’s name or by trying to confirm the hypothesis from the user model.

Figure 8.6 shows the dialog state diagram of the receptionist robot. The initial state is the *Idle* state. Once the system spots a person walking towards the robot, or speech input is observed, the system enters the greeting state. Next, the *PersonID* state is entered to identify the person. During several experiments, we have used different strategy implementations of the dialog module for identifying the user ID, including handcrafted models (chapter 9) and strategies trained by reinforcement learning (chapter 5). The *Social User Info* state is entered once the user’s ID is determined. The strategy of this module and experiments for social network analysis are presented in chapter 9. Any state can lead over to the *Goodbye* state, when the user wants to abort the dialog or wants to leave, e.g. by saying “i have to go now”, and after a short goodbye ‘handshake’, the system enter the idle state. Any other state can also lead over to the *Idle* state, when no relevant speech input events are observed, which happens, e.g., if the user simply walks away from the system.

The *PersonID* state is a crucial state for the interaction. The interaction only affects the knowledge base, if the person who interacts with the system can be identified correctly. Once the ID of a person is obtained, it is stored in the user model, and once the *Idle* state is reached again, all model updates are executed, i.e. updating the database and updating voice ID and face ID models with new data from the current interaction.

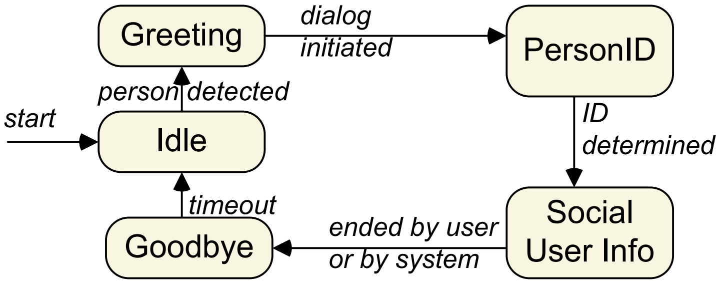


Figure 8.6: Dialog states of the interACT receptionist

8.3 Social Studies with the interACT Receptionist

8.3.1 Overview

This section presents social analyses of dialog interactions, which were conducted in a controlled experiment during the design time of the system. The purpose of the experiment was to better understand social aspects of the system and to reveal possible problems of the system design, which provided helpful insights for improvement of the dialog strategy and system components. The methods to analyze such kind of interactions have been published as the Interaction Analysis Tool (IAT).

Current ongoing discussions in the Human-Robot-Interaction (HRI) field address the issue which metrics and evaluation measures should best be applied to evaluate a system. For example, the recent workshop about Evaluation Metrics 2008 was held in conjunction with the HRI conference 2008. It addressed basic questions about which metrics should be applied to HRI and how to evaluate such systems. The current state of the art suggests a combination of different metrics, including objective metrics and subjective metrics, but also analysis of social behavior.

So far, evaluations in this thesis have mostly been conducted with objective metrics, e.g. WER. Subjective metrics are assessed with questionnaires. The social aspect addresses the *human* factor in the interaction, and analyzes the user's behavior and the user's reactions. Evaluation of these aspects includes video transcription, interviews and questionnaires. So far, such analysis cannot be automated and requires a lot of manual labor, which is very time consuming (the current tool requires manual labor for video transcription with an effort of roughly 120x real-time). It results in a qualitative analysis, which provides helpful insights, for example, what kind of problems arise in which situation, and observations how humans react to the system.

A common problem in the design of a novel interactive system is the cross-dependency of system design and user feedback. It is almost impossible to design a perfect system before assessing user feedback. At the same time it is almost impossible to assess user feedback before a system has been designed. A design cycle of several loops, often starting with a Wizard-of-Oz experiment, is the best known method to resolve these cross-dependencies.

In such a development cycle, qualitative user feedback can be used by the system developer to improve the system, and gain a deeper understanding of the flow of communication. For such analysis, a tool for close analysis of human-robot interactions (the Interaction Analysis Tool - IAT) has been developed in a multi-disciplinary team and introduced in Burghart et al. (2008). This section is based on two publications (Burghart et al., 2007; Holzapfel et al., 2008a).

Analysis with the IAT uses a multi-methodological mixture of quantitative and qualitative methods from empirical social science. Focus of these methods for close analysis of recorded video data has been activities of the users, and the interaction itself. While the quantitative video analysis reveals the "what", "when", "where", (represented as a transcription of the video sequence), the qualitative video analysis describes the "why", and "how". The quantitative analysis can be conducted as an objective analysis, and it is generally annotator-independent. The qualitative analysis however is influenced by the annotator's perception, and the resulting interpretation is annotator-dependent. Therefor, it should generally be conducted by groups of annotators, where interpretations of short sequences are commonly agreed upon with the principle of best argument.

8.3.2 Experimental Setup

The experiment was set up as Wizard-of-Oz experiment with the interACT receptionist acting as a parcel receptionist. The subjects had to interact with the robot to obtain information necessary to complete their task to deliver a parcel, the robot in turn was interested in registering the person, i.e. to understand the first name of the subject, before it could provide the information. By adopting a Wizard-of-Oz experiment, it was possible to control the interaction by a human operator who acts as a wizard and can decide which actions are taken by the system. As beforehand no detailed experiments had been conducted so far in this scenario for the dialog structure, the Wizard-of-Oz setup led to more reliable and comparable behavior. During the Experiment the wizard has some limited control of the system's behavior. While most parts of the system are implemented and run autonomously, the wizard replaces the dialog strategy, i.e. which (spoken output) actions to apply. Figure 8.7 gives an overview of the components and sketches the data-flow of the system.

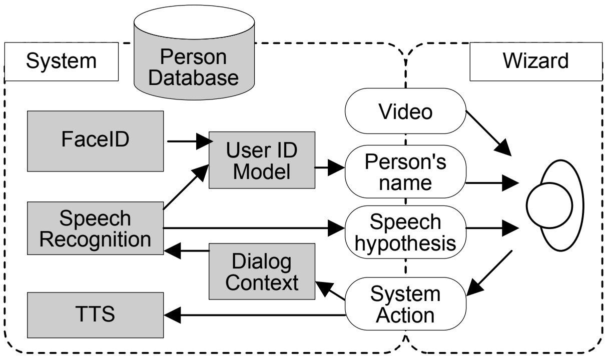


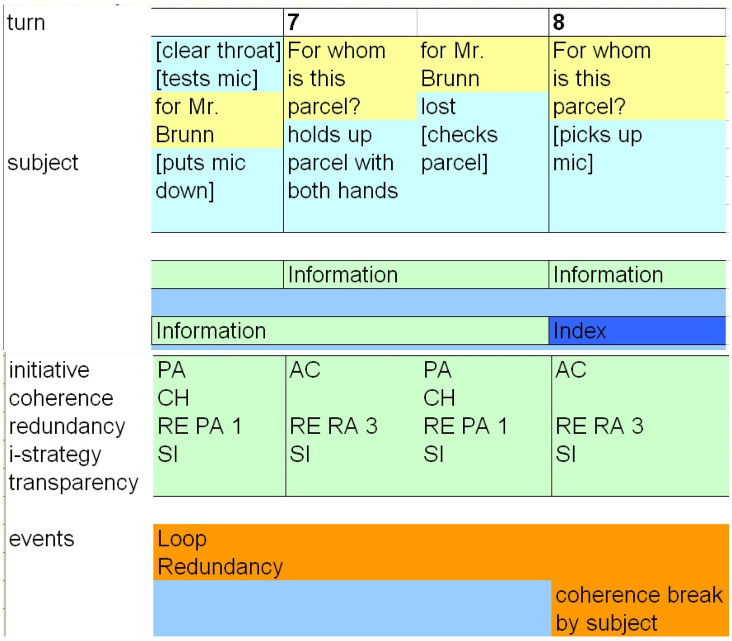
Figure 8.7: Architecture of the multimodal dialog system operated by the wizard

The conducted experiments comprised a robotic system and 16 naive subjects, eight of them social scientists, the other eight computer scientist majoring in robotics. Each subject in turn was handed a parcel, and was told the name of the recipient, but not where to find the recipient. In the aisle they could meet the robot and ask for information, which they did not know beforehand. Both groups of students were split into two: with four subjects of each group the robot acted in an empathic manner, the other times the robot adopted a rational manner. The experiments were repeated on three consecutive days with slight variations: During the second experiment, the recipient's name was different and the room was changed as well. During the third experiment, the recipient was the same as during the first day, but this time the room was locked.

Hartwig Holzapfel

8.3.3 Interaction Analysis

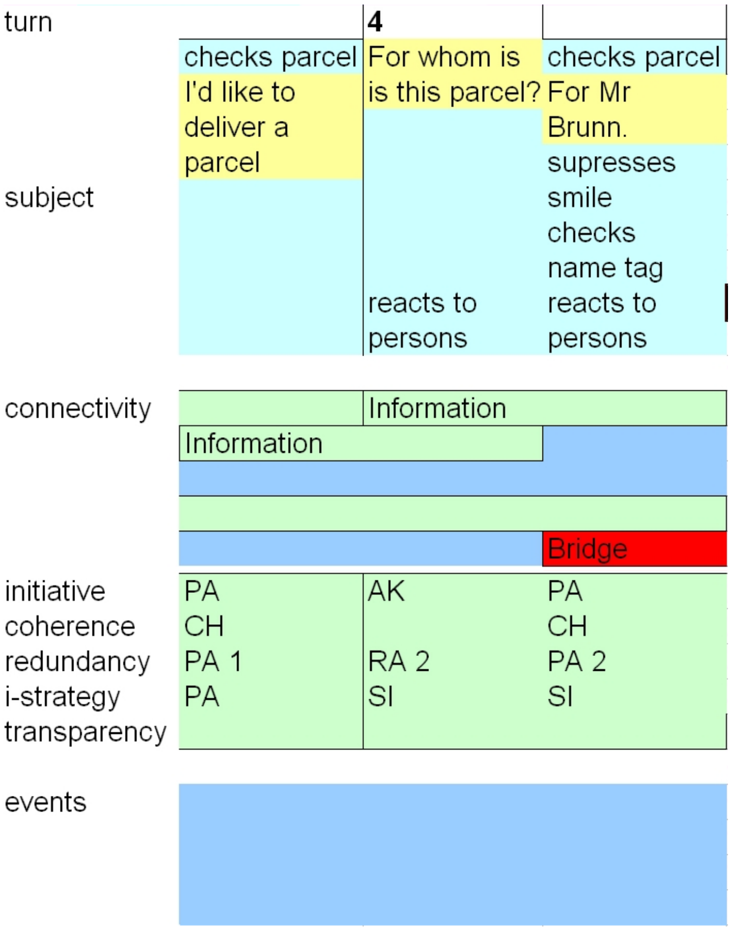
Once interactions between human subjects and robot are coded by IATs, several possibilities to analyze and compare data do exist, including close analysis of single interactions, analysis of communication problems, differences between different trials by a single person or differences of interaction styles between different persons. In contrast to earlier studies, we did not find any difference in the behavior of social science students and computer science students. During the experiments, some persons could easily and successfully interact with the robot, others encountered needed longer trials or could not achieve their goal. The analysis of these problems with the IAT provided the basis for preventing such problems in the final system of the interACT receptionist.



turn		7		8
subject	[clear throat] [tests mic] for Mr. Brunn [puts mic down]	For whom is this parcel? holds up parcel with both hands	for Mr. Brunn lost [checks parcel]	For whom is this parcel? [picks up mic]
		Information		Information
	Information			Index
initiative coherence redundancy i-strategy transparency	PA CH RE PA 1 SI	AC RE RA 3 SI	PA CH RE PA 1 SI	AC RE RA 3 SI
events	Loop Redundancy			
				coherence break by subject

Figure 8.8: Section of IAT with layers 1, 3, 4, 5 of first trial of a subject

Typical problems that arose during first interaction by naive subjects were often accompanied by the users being uncertain how to speak to the system. This is manifested in extreme cases by checking the microphone and test-speaking into the microphone. Figure 8.8 shows an IAT excerpt of such a case. The excerpt encodes that the user repeats a statement that, from a system point of view, could not be understood as the utterance was not covered by the grammar. After repeated misunderstanding, the user checks



turn		4	
subject	checks parcel I'd like to deliver a parcel	For whom is is this parcel? reacts to persons	checks parcel For Mr Brunn. supresses smile checks name tag reacts to persons
connectivity		Information	
	Information		
			Bridge
initiative	PA	AK	PA
coherence	CH		CH
redundancy	PA 1	RA 2	PA 2
i-strategy	PA	SI	SI
transparency			
events			

Figure 8.9: Section of IAT with layers 1, 3, 4, 5 of third trial of a subject

possible reasons of failure such as malfunction of the microphone, shows the parcel to the robot, checks the name on the parcel again. Though an extreme case, the selected IAT section is typical for naive subjects interacting with a robot the first time. This subject uses different strategies in order to find out why the interaction does not proceed as desired and in order to get out of the loop. Although on the third day the subject still tends to examine both, microphone and parcel, just to make sure, she achieves her goal without ado. The same IAT section as before, this time during the third day, is shown in figure 8.9.

Contemplating all trials of the presented subject, a definite adaptation of the user to the robotic system can be found. This can be seen by the following quantified data: In the first trial, 31 user utterances were needed to achieve the goal, 13 of which belong to loops, where information was simply repeated. Five turns showed a break of coherence and six times there was an omission of an answer. High redundancy existed in 18 turns, which complicated language understanding, and ten times the subject took over the dialog initiative in inadequate situations. In the third trial, only nine turns were needed to achieve the goal and no loops were detected. There were no breaks of coherence, no omissions and no changing of the initiative. The dialog recorded 11 turns, 0 turn errors. Also they way of spelling the name was changed during the third day, so that the speech recongizer could perfectly recognize the spoken input.

8.3.4 IAT Summary and Conclusion

By analyzing multiple single interactions of different subjects, as well as consecutive trials of the same subject within a specific scenario, one can reveal critical states in the interaction. By analyzing the context of these states, strategies could be developed to get the person back on track or even how to avoid such situations. At present, the tool ties a lot of labor as all recorded video data and log-files have to be transcribed and incorporated in the IAT. The second time consuming step is to fill out the different categories by hand. Prerequisite for a sound evaluation naturally is that subjects are not biased by the presence of team members or other subjects. However, as there is ongoing research in analyzing human-robot interactions, we assume that future research will lessen the high cost of manual labor required for the experiments, by automating parts of the annotation, and allow a broader application of such analysis tools. Furthermore, if larger amounts of data can be processed, it seems to be a promising approach, to introduce additional quantitative categories, and predict such categories in a similar way, as subjective evaluation is predicted from objective evaluation measures, detect problem situations, or use these categories to predict strategies of the user.

The experiments show how important it is to provide users with informative help to understand and adapt to the system's capabilities, especially when comparing successive trials of naive users. Though it is desired that the system adapts to the human, and today's systems achieve this goal more and more, there are still many aspects which require the user to understand the system's capabilities, even if they seem simple, such as how to use the microphone, or the capabilities of the robot's vision system. By closely analyzing such interactions and assessing the user's reactions, several problems could be uncovered and prevented in the final interACT receptionist system.

8.4 Subjective Evaluation of Dialog Interactions

8.4.1 Questionnaires

Subjective factors are typically assessed by questionnaires that are filled out by users after interacting with the system. The questionnaires designed for the presented system are based on 7-point Likert-Scales. We have used two types of questionnaires in subsequent experiments. Both questionnaires originate from the SASSI questionnaire described in Hone and Graham (2000, 2001) and have been adapted to the interACT receptionist scenario. The first questionnaire adopts the question style of the SASSI questionnaire with statements that are rated by users on a scale from strongly disagree to strongly agree. It was used in experiments presented in chapter 9 on acquisition of social user models. Subsequently, the questionnaire was further optimized for an experiment person identification dialogs and modified regarding question style, question order, wording, and question selection. The wording now avoids strong statements with negative emotions. The question style has been changed from statements with a disagree/agree-scale to real questions and question-specific scales, also on a 7-point rating scale. For example, this version uses scales that range from 'no fun at all' to 'very much fun' as semantic opposites. If not explicitly stated otherwise, the analyses reported here have been conducted with the optimized questionnaire. A full list of key questions is given in table 8.2, and the translation to English is given in table 8.3.

8.4.2 Factor Analysis

Analysis of a questionnaire can be conducted by evaluating every response separately or by a factor analysis. A factor analysis provides a means to discover underlying 'hidden' factors that represent the subjects' attitudes, which constitute the observed responses. Single questions represent the desired information only in parts. We have conducted a standard factor analysis with Varimax rotation, which led to a good separation of factors with clear interpretations. For the analysis and interpretation of the factors we have used methods described by Möller et al. (2007), see also chapter 2.3.2. Though the scenario is different than the scenarios studied by Moeller the analysis reveals some similarities in the interpretation and some of the factors can be related to the qualiy aspects described by Moeller.

Table 8.1 shows the results of the factor analysis and the loadings of the questions on the most important factors. During the analysis, we have iteratively removed questions with the highest loading of 0.4 or less on any factor,

Qst-key	Factor1	Factor2	Factor3	Factor4	Factor5
QST1	0.56	0.51	0.35	0.23	-0.04
QST3	0.84	0	-0.27	-0.09	0.03
QST4	0.15	0.72	0.26	0.17	0.25
QST5	-0.1	0.02	0.79	0.01	-0.1
QST7	0.4	0.81	0.17	0.17	-0.02
QST9	0.21	0.63	-0.28	-0.1	0.32
QST10	0.12	-0.51	-0.14	-0.56	-0.05
QST11	0.74	0.45	0.14	0.07	0.12
QST12	0.68	0.23	0.37	0.23	0.24
QST13_1	0.54	0.54	0.44	0.23	0.09
QST13_2	0.62	0.51	0.38	0.29	0.03
QST13_3	0.49	0.28	0.58	0.11	0.23
QST13_4	0.33	0.14	0.71	0.01	0.25
QST14	0.67	0.37	0.2	0.18	0.17
QST15_1	0.1	-0.02	0.16	0.6	0.62
QST15_2	0.68	0.05	0.16	0.51	0.13
QST15_3	0.16	0.21	0.59	0.17	0.54
QST15_4	0.27	0.15	-0.05	0.72	-0.01
QST16	0.11	0.19	0.03	-0.03	0.85

Table 8.1: Factor loadings of selected questions

which eliminated questions 8 and 13_5. We have also removed questions that had a shared loading with values between 0.4 and 0.5 on different factors, which eliminated question 2. Afterwards on only questions with loading >0.5 remain in the set. The factor analysis produced 5 factors that correspond to an eigenvalue larger than 1.0. The total variance that is explained by these factors is 73.81%. Factor 1 explains 22.73% of the total variance, factor 2 explains 17.04%, factor 3 explains 14.79%, factor 4 explains 9.70%, and factor 5 explains 9.55%.

8.4.3 Factor Interpretation

The factors can be interpreted by analyzing the loadings of the questions on the factor. The question with the highest loading is considered as the most important aspect of this factor. Usually, the interpretation of what these factors represent, can be ambiguous. Therefore, the following interpretation has been discussed by the author with peers to find commonly acceptable interpretations.

- **Factor 1: “overall impression and acceptance”**.
  The key questions of this factor are QST3 (0.84), QST11 (0.74), QST12 (0.68), QST15_2 (0.68), QST14 (0.67), QST13_2 (0.62), QST1 (0.56), QST13_1 (0.54). Comparison with the work from Moeller shows a good correlation with Moeller’s category “acceptance”.

  (shows statistical correlation with Moeller’s quality aspect acceptance)

- **Factor 2: “communication flow”**.
  Comparison with the work from Moeller shows a good correlation with Moeller’s category “interaction efficiency”.

  (shows statistical correlation with Moeller’s quality aspects interaction efficiency, transparency and cognitive demand)

- **Factor 3: “conversational control”**.
  Comparison with the work from Moeller shows a moderate correlation with Moeller’s category “symmetry”.

  (does not show statistical correlation with any of Moeller’s quality aspects)

- **Factor 4: “user friendliness and usability”**.
  Comparison with the work from Moeller shows a moderate correlation with Moeller’s category “ease of use”.

  (does not show strong statistical correlation with any of Moeller’s quality aspects, but shows highest correlation with dialog success)

- **Factor 5: “responsiveness and system response”**.
  Comparison with the work from Moeller shows a slight correlation with Moeller’s category “cooperativity”.

  (shows some statistical correlation with Moeller’s quality aspects interaction efficiency and cognitive demand)

8.4.4 Results

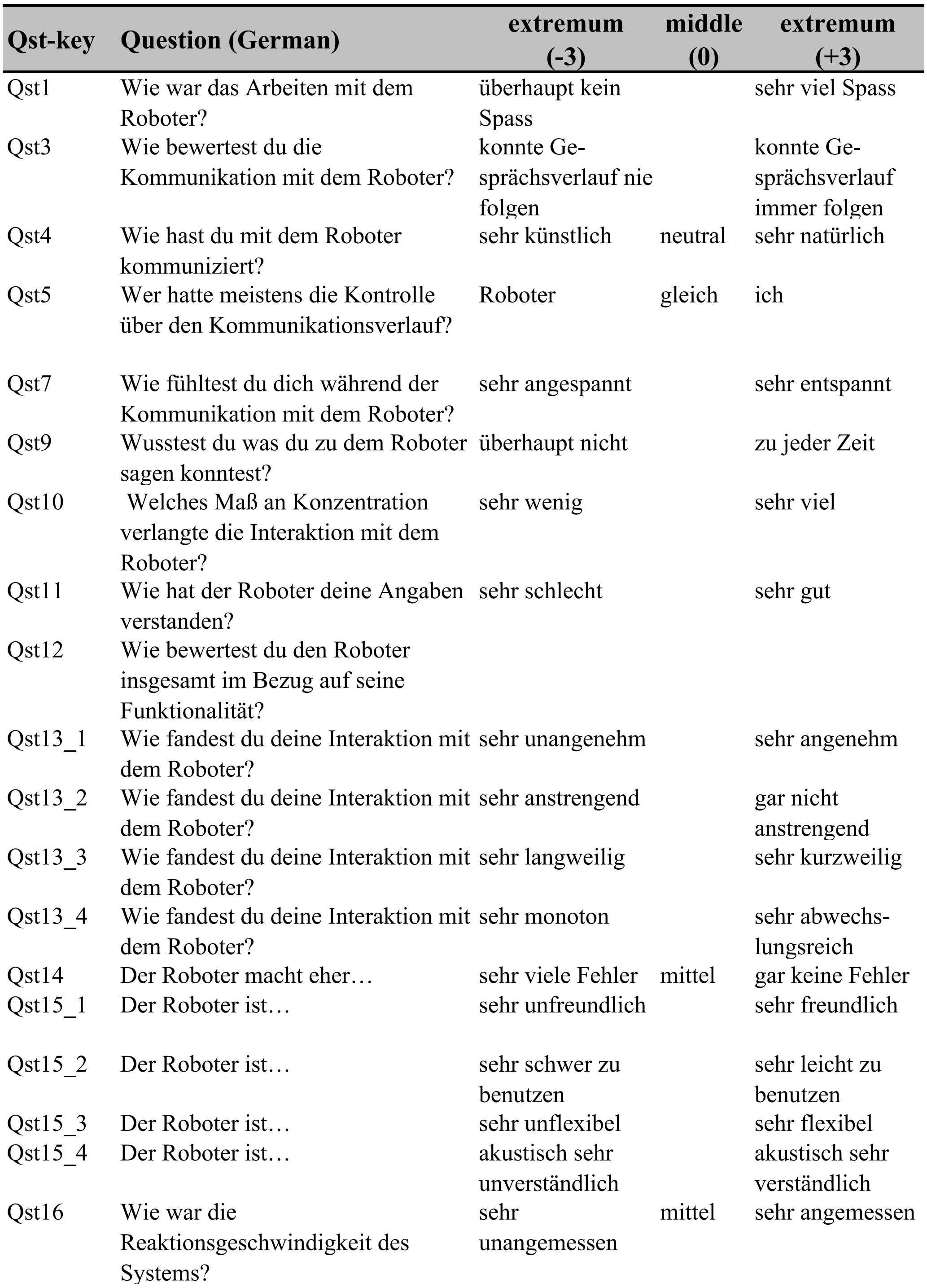
The factor analysis reveals important aspects for evaluation and the correlation of each question with each factor. The overall impression of the receptionist was rated as good (“overall impression and acceptance” = 0.39). That means, the receptionist is generally accepted by the users. Most users also rated the system to have good communication flow (“communication flow” = 0.34), be rather monotonous (“conversational control” = -0.50), very friendly and usable ("user friendliness and usability" = 0.70), and responsive ("responsiveness and system response" = 0.36).

In addition to averaged numbers, the subjective questionnaire also provides insights into specific aspects, usually problems that exist, where the feedback of a single user differs from the averaged result. They reveal that in some cases users were misled by the system, or that the system was sometimes hard to understand. Problems with the flow of communication exist for example due to problems to understand the English Text-To-Speech component (mostly students, without English as native language), or that names are mispronounced by the Text-To-Speech component. This is not surprising, as the users had different expectations of how German names would be pronounced by the English-speaking system.

Speech recognition and especially name recognition accuracy has strong effects on the outcome of the questionnaires. This is reflected by generally better scores in all categories for dialogs that complete successfully. In contrast to analyses from Moeller, not all aspects where also discovered in the present system. We interpret this fact that in the given scenario not all these aspects are important to the user, and that these aspects also depend on the scenario.

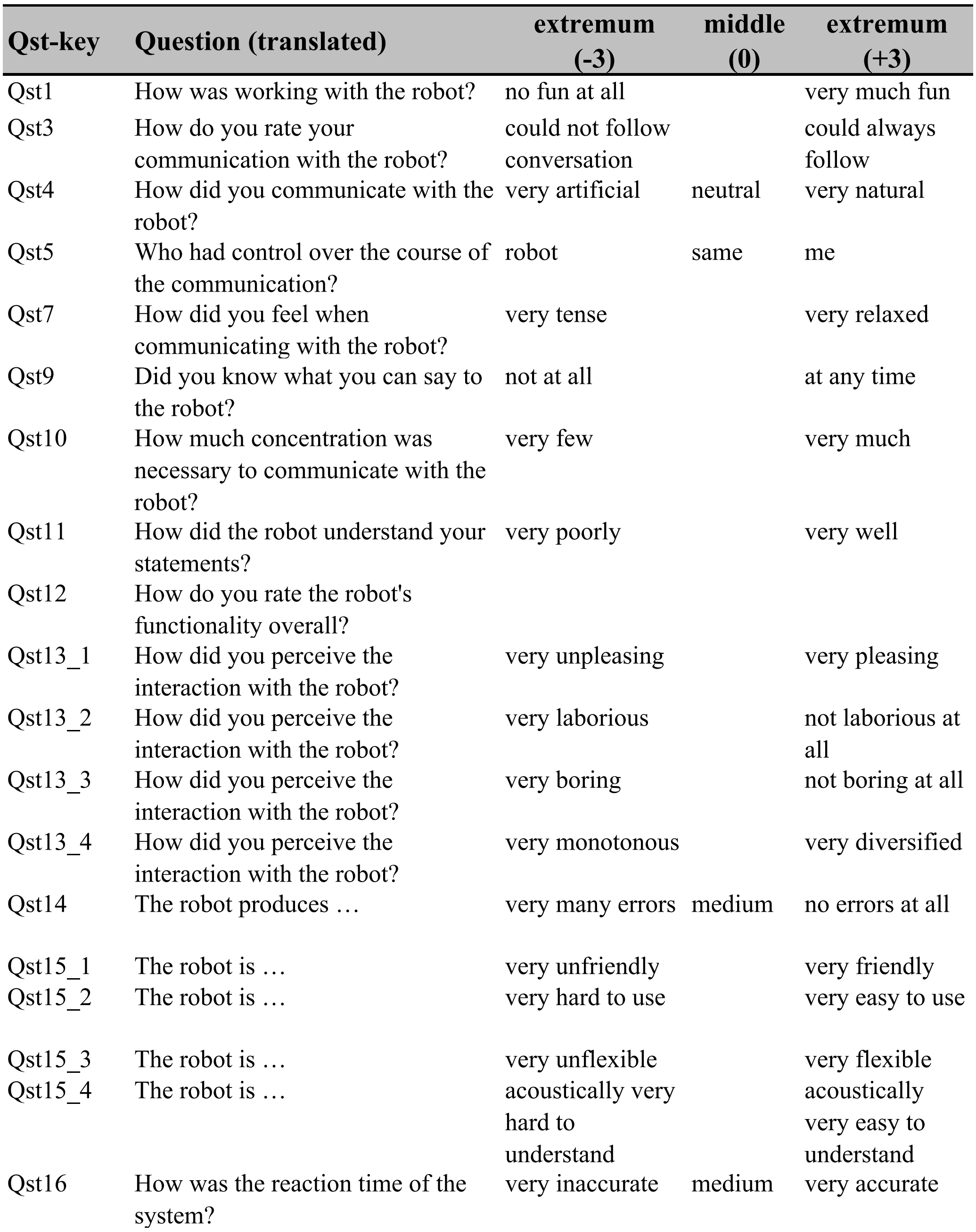
8.5 Conclusion

This chapter has introduced the interACT robot receptionist and analysis of dialog interactions with the robot. Interaction analysis using video transcription requires a large amount of manual labor but provides a detailed analysis of the interaction and reveals communication problems of prototype versions of the system that can be solved by improving the dialog system. Analysis of subjective user feedback with the final system provides insights into the user's perception of the dialog, shows different quality aspects, and shows that users generally accept the system.



Qst-key	Question (German)	extremum (-3)	middle (0)	extremum (+3)
Qst1	Wie war das Arbeiten mit dem Roboter?	überhaupt kein Spass		sehr viel Spass
Qst3	Wie bewertest du die Kommunikation mit dem Roboter?	konnte Gesprächsverlauf nie folgen		konnte Gesprächsverlauf immer folgen
Qst4	Wie hast du mit dem Roboter kommuniziert?	sehr künstlich	neutral	sehr natürlich
Qst5	Wer hatte meistens die Kontrolle über den Kommunikationsverlauf?	Roboter	gleich	ich
Qst7	Wie fühltest du dich während der Kommunikation mit dem Roboter?	sehr angespannt		sehr entspannt
Qst9	Wusstest du was du zu dem Roboter sagen konntest?	überhaupt nicht		zu jeder Zeit
Qst10	Welches Maß an Konzentration verlangte die Interaktion mit dem Roboter?	sehr wenig		sehr viel
Qst11	Wie hat der Roboter deine Angaben verstanden?	sehr schlecht		sehr gut
Qst12	Wie bewertest du den Roboter insgesamt im Bezug auf seine Funktionalität?			
Qst13_1	Wie fandest du deine Interaktion mit dem Roboter?	sehr unangenehm		sehr angenehm
Qst13_2	Wie fandest du deine Interaktion mit dem Roboter?	sehr anstrengend		gar nicht anstrengend
Qst13_3	Wie fandest du deine Interaktion mit dem Roboter?	sehr langweilig		sehr kurzweilig
Qst13_4	Wie fandest du deine Interaktion mit dem Roboter?	sehr monoton		sehr abwechslungsreich
Qst14	Der Roboter macht eher…	sehr viele Fehler	mittel	gar keine Fehler
Qst15_1	Der Roboter ist…	sehr unfreundlich		sehr freundlich
Qst15_2	Der Roboter ist…	sehr schwer zu benutzen		sehr leicht zu benutzen
Qst15_3	Der Roboter ist…	sehr unflexibel		sehr flexibel
Qst15_4	Der Roboter ist…	akustisch sehr unverständlich		akustisch sehr verständlich
Qst16	Wie war die Reaktionsgeschwindigkeit des Systems?	sehr unangemessen	mittel	sehr angemessen

Table 8.2: Questions and extreme values of the rating scale. Original version in German.



Qst-key	Question (translated)	extremum (-3)	middle (0)	extremum (+3)
Qst1	How was working with the robot?	no fun at all		very much fun
Qst3	How do you rate your communication with the robot?	could not follow conversation		could always follow
Qst4	How did you communicate with the robot?	very artificial	neutral	very natural
Qst5	Who had control over the course of the communication?	robot	same	me
Qst7	How did you feel when communicating with the robot?	very tense		very relaxed
Qst9	Did you know what you can say to the robot?	not at all		at any time
Qst10	How much concentration was necessary to communicate with the robot?	very few		very much
Qst11	How did the robot understand your statements?	very poorly		very well
Qst12	How do you rate the robot's functionality overall?			
Qst13_1	How did you perceive the interaction with the robot?	very unpleasing		very pleasing
Qst13_2	How did you perceive the interaction with the robot?	very laborious		not laborious at all
Qst13_3	How did you perceive the interaction with the robot?	very boring		not boring at all
Qst13_4	How did you perceive the interaction with the robot?	very monotonous		very diversified
Qst14	The robot produces …	very many errors	medium	no errors at all
Qst15_1	The robot is …	very unfriendly		very friendly
Qst15_2	The robot is …	very hard to use		very easy to use
Qst15_3	The robot is …	very unflexible		very flexible
Qst15_4	The robot is …	acoustically very hard to understand		acoustically very easy to understand
Qst16	How was the reaction time of the system?	very inaccurate	medium	very accurate

Table 8.3: Questions and extreme values of the rating scale. Translated English version.

Chapter 9

Social User Models and Interactive Learning of Social Networks

This chapter describes a learning scenario where the task of learning is the acquisition of a social user model. The term social user model is derived from a system's perspective on a social network structure, which is observed and reproduced by the robot. The experiments presented here have already been published in Putze and Holzapfel (2008), this chapter is slightly modified from the original publication. Research for this publication has been conducted in a Diplomarbeit at the Universität Karlsruhe (TH) by Putze (2008) and has been supervised of the author of this thesis.

9.1 Introduction

Today's humanoid robots are intended to integrate into the daily life of their owners. They still lack social awareness to gain a full understanding of human behavior and interaction. According to Drury et al. (2003), social awareness (and group-structural awareness) requires knowledge in terms of a person's role and responsibilities, its status, and group processes. This chapter introduces an approach to equip a robot or any other cognitive system with social user models. This knowledge will enable the robot to better predict and understand human behavior and to offer a more natural dialog experience for its users.

The system presented here is termed the 'IslEnquirer'. It is a predecessor of the interACT robot receptionist and acquires user models that represent social structures like roles, personal ties and cohesive groups within a computer science lab. To this end, we combine two complementing components: An offline step processes a corpus of publications using methods from social network theory and information retrieval. The result of this step is an initial social model that is then verified and extended during spoken human-robot interactions in which the robot interviews the user about social information. Figure 9.1 demonstrates this process. This extension and verification through a combination of offline and online acquisition is the main contribution of this paper.

Most existing approaches to collecting social user data are based on (automatic or manual) offline analysis, e.g. Terrill L. Frantz (2006) and Newman (2001), by studying connectivity, centrality or other properties in social networks built from existing data. In Newman (2001) for example, information from existing publication databases is used to create a social network based on co-authorship. The author then uses various measures from graph theory to investigate this network, e.g. by studying the node degrees, connected components, node distances or the number of small subgroups. This allows the author to draw conclusions about the structure of the different networks. Other works, like Terrill L. Frantz (2006) focus more on the different roles of single actors within a network.

Arnetminer (Yao et al., 2007) is a web site presenting automatically gathered information on members of the worldwide scientific community. This information includes the person's affiliation, the research interest and a list of associated researchers. The system searches the web for data and employs a combination of several classifiers and heuristics to extract the relevant information. Based on co-authorship, the system builds a social network to identify cooperation within the community. The scope of this work is much wider than the one of the IslEnquirer scenario but it is based solely on information retrieval and does not report research groups or social roles.

To our best knowledge, our work presents the first attempt to create social user models using spoken interaction, which augments the classic offline approach by gathering data directly from the subjects and thus adding complementary information. We do this by using a speech interface installed on a robot to minimize the required initiative and effort on the user's side (e.g. compared to a text based interface).

The remainder of this chapter is organized as follows: Section 9.2 presents the general structure of our social user models. Section 9.3 explains how information is extracted in the offline step and section 9.4 contains a description of the dialog component. Section 9.5 describes the experimental setup and presents the results.

9.2 Social User Models

We acquire social user models in the context of a medium sized scientific community. The attribute types which are contained in the models reflect this domain. However, they can easily be adapted to all social contexts. The modular design of our system allows a convenient replacement or addition of attribute types for other domains. We collect the following attributes:

Importance reflects the relevance of other people for the subject. **Research groups** refer to membership in a group of specialists for a shared research area or in an interdisciplinary group working on a common goal. A **Role** is the position a person occupies within the institute hierarchy. The **Research interest** describes the general subject the user is currently working on.

The social data gathered during both the offline network analysis and interactive learning is prone to noise. On the one hand, there is noise due to mistakes of the network algorithms or the automatic speech recognition (ASR) component. On the other hand, there is noise which is induced by wrong or outdated information and inherent ambiguity. The design of the IslEnquirer accounts for these observations with a user model which supports multiple hypotheses and confidence scores for single attribute values and collections of hypotheses. The updating algorithm regards the reliability of the incorporated information, for example based on the learned efficiency and effectivity of the used information channel.

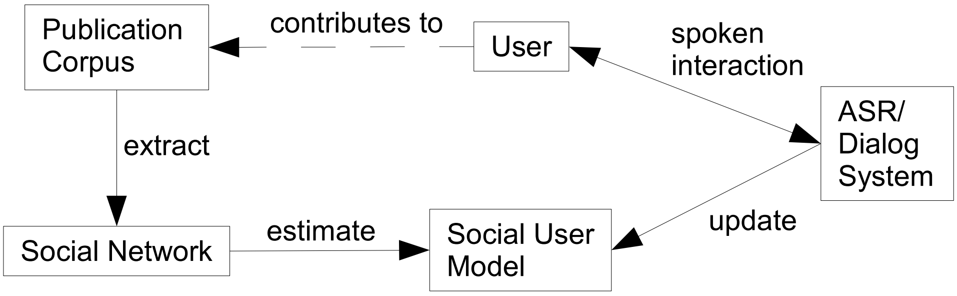


Figure 9.1: Data flow within the social user modeling system: Offline and online processing work together to build a common social user model

9.3 Offline Network Analysis

To initialize the social user models, we introduce an *offline step*. Here, we process a publication corpus gathered from the official institute web site. We use several algorithms, which are described in the following, to generate hypotheses for importance, research groups and roles from this data.

In the first step, we build a social network from publication co-authorships. A social network is a (directed) graph, where every node represents one person in the database. The nodes are connected by weighted edges. The weight of an edge depends on the frequency of joint publications of the connected people, weighted by age. After normalization, we can interpret the weight of the edge from A to B as the *importance* of B for A.

$$\text{importance}(A, B) = \frac{\sum_{p \in \text{Pub}(A,B)} \frac{1}{\text{age}_p + 1}}{\sum_{B'} \sum_{p \in \text{Pub}(A,B')} \frac{1}{\text{age}_p + 1}} \tag{9.1}$$

We can now use this graph to derive other attributes: Cohesive subgroups are a classical higher-order structure that can be identified in social networks. In our application, we interpret them as research groups as their members are working closely together. A usual way of finding groups in a social network is by searching for cliques. However, this approach alone is too restrictive for finding all relevant groups and it does not offer a possibility to compare two similar group hypotheses. In Putze (2008) we therefore propose an alternative procedure, starting with all cliques as tentative group set, iteratively merging groups similar in composition and associated research topics.

To automatically identify and assign roles, we want to form clusters of people with a similar social position. Social network theory offers multiple measures for calculating role similarity. To integrate different approaches, we calculate a Euclidian distance in a multidimensional feature space, where each dimension represents one similarity measure. This approach allows easy integration of new features and flexible combination of multiple criteria. Examples of features we used are: **Regular Equivalence**, which measures how well the neighbors of two actors correspond to each other, computed using the REGE algorithm described by Wasserman and Faust (1997). **Prestige** (Wasserman and Faust, 1997), which is a measure of relevance in social networks, i.e. an actor is assigned a high prestige value if it has strong ties to other actors with high prestige (solved as eigenvalue problem). **Entropy** of the importance relation distribution, which measures whether the person's relations focus on a small group of coworkers or are scattered around the whole institute. The other criteria are number of associated groups, in- and out-degree and an activity indicator.

Additionally, we employ TF-IDF scoring and part-of-speech tagging to create attribute labels and a speech recognition vocabulary from the available publication abstracts.

9.4 Online Robot-Initiated User Interview

The data generated during the offline step is written in a user model. The same data structures are used by the IslEnquirer to validate and extend the social user models by using a spoken dialog system. This system interviews each user on the data stored in the user model. To this end, we used the TAPAS dialog manager on a robot located in the hallway of the institute as described earlier in this thesis. This system tries to attract, identify and then interview bypassing people.

robot:	*I am Robbi, the new ISL webmaster. Do you have time to update my database?*
user:	*Sure!*
robot:	*You are Hartwig Holzapfel, right?*
user:	*That's right.*
robot:	*I just talked to Maria. She was very helpful!*
robot:	*Ok, let's start with the first question!*
robot:	*The research interest of Hazim Kemal Ekenel is face recognition. What is your general research interest?*
user:	*I am working on dialog management.*
robot:	*You told me that you are researching on dialog. Is that right?*
user:	*That's correct.*
robot:	*Now it's clear to me. Thank you!*

Table 9.1: Example dialog excerpt

The dialog consists of several modules. At the beginning of each session, the multimodal *user ID module* identifies the user (see also chapter 4, especially section 4.2). This task is simplified by the fact that most names of interest are already in the database before the person interacts with the robot the first time due to the already described offline vocabulary initialization from research papers. After identification, the dialog enters the *social user model acquisition module.*

The social user model acquisition module consists of several question subdialogs. Every subdialog concentrates on a single attribute in one social user model (although it can indirectly influence many other entries). Each question subdialog has a similar structure. The subdialog may begin with an optional example to introduce the next topic to the user. Then, the user is asked a specific question concerning the social user models. If no suitable answer is received, the system uses a repetition question, which can contain a different wording of the same question or the request to repeat or rephrase. When input is received, the system can ask for confirmation, depending on the agreement of the answer with the belief of the system. Between subdialogs, transitions are marked by an acoustic progress bar ("Just one more question.") and chit chat snippets.

The following paragraphs describe important aspects of the question subdialog in greater detail.

9.4.1 Social User Model Questions

All questions concerning social user models are generated from generic question models, which are represented in an attribute independent way to allow convenient addition of new questions. There are two main dimensions along which questions are categorized: open vs. closed questions and direct vs. third person questions.

Open questions ask the user to formulate a free answer which is then parsed by the NLU to extract a label. *Closed questions* propose a label hypothesis which the user can either confirm or reject. Open questions are intended to acquire new information in an unbiased way, while confirmation questions validate existing knowledge. At first glance, closed questions are more restrictive and less informative than open questions. They can however propose attribute values that the user would not come up with by himself. They also include the benefit of better speech recognition performance.

While *direct questions* deal with the current user himself, *third person questions* ask the user on information about another institute member. The reason for the latter is to get additional sources of information and to acquire information on people who do not regularly visit or talk to the robot.

	open	**closed**
direct	Who is your most important coworker?	Are you researching on `[label]`?
third person	What is `[subjects]`'s role at the I S L institute?	Is `[subject]` in the `[label]` research group?

Table 9.2: Examples for questions of all four basic types

9.4.2 Question Selection

Every user can only spend limited time talking to the system. Therefore, we cannot pose every question that is available but we have to choose the "best" questions, based on certain criteria. Foremost, we need to pose questions which give us as much information as possible. Usually, this means choosing attributes with a low confidence. Additionally, we want questions that are known for high success rates, which means that they are easily understood by the user and the responses are most probably covered by our grammar. Another important goal is not to bore the user by repeating the same questions and topics over and over again. Concerning third person questions, we weigh them according to the importance of the question target for the user. Finally, we have some requirements on the global dialog flow, e.g. when a question is aborted, we want the next question to stick to the same topic (see also table 9.3). To cope with these different, often adversarial goals during question selection, we use a flexible, modular scoring approach: For each question q and criterion c_i, a score $s_{i,q}$ in the interval $[0, 1]$ is calculated. These scores are weighted by w_i according to their desired influence on the selection process. For each question we calculate the product of these scores and select the one with the overall highest score.

$$\hat{q} = \mathbf{argmax}_{q \in Q} \prod_i s_{i,q}^{w_i} \tag{9.2}$$

score calculation	**explanation**
$1 - \mathbf{conf}(A)$	when Q is an open question (promises high information gain)
$0.5 - \lvert 0.5 - \mathbf{conf}(A) \rvert$	when Q is a closed question (same, avoids asking for low quality values)
$\mathbf{reliability}(Q)$	based on avg. duration and success rate of Q
$\mathbf{importance}(U, T)$	when Q is a third person question
$1 - \mathbf{frequency}(Q, U)$	for variable dialog and broad coverage
$c_1 < 1$	when A was already covered successfully during this session (constant)
$c_2 < 1$	when A is a topic change and last topic was not covered successfully (constant)

Table 9.3: Examples for criteria for scoring question Q on attribute A for user U on target T

9.4.3 Improving the Interaction

The system makes ample use of examples as they are an unobtrusive way of communicating its capabilities and limitations. This takes place on two levels: On the first level, the user is implicitly informed what degree of complexity the system is able to process. He will also adopt the proposed formulations for open questions, where many responses are valid. This helps to reduce recognition errors. On the second level, we counter varying levels of granularity which otherwise would lead to ambiguities. Giving examples at an early stage will lead to a more consistent database. As an additional benefit, a better guidance for open questions results in reduced cognitive load and a more comfortable dialog experience.

Examples are automatically generated from the set of all attribute values and their labels and represented in a form which is independent of the attribute type. Examples are selected based on the association and label scores. We prefer examples on well-known people for which we have a clear attribute assignment. Based on this scoring, we do a randomized selection to support a broad variety of examples.

The IslEnquirer dialog strategy uses confirmation questions that reduce the likelihood that incorrect information is stored in the models. For the decision whether information needs to be confirmed, we try to find a balance between dialog length and correctness: Confirmation is requested if the concept extracted from the user's reply is not in accordance with the belief of the system. For closed questions, this is the case if the user rejects the hypothesis of the system. For open questions, this decision is made by determining the maximal score of the label in question over all associated attributes of the inferred type. When this maximum exceeds a fixed threshold, no confirmation is requested.

When recognition of one label consecutively fails several times, the dialog manager switches to the *learning module*. Here, the user is asked to repeat only the used term to ensure grammar coverage for known words. If this fails, the user is asked to spell the label. As spelling of long words is tedious and error-prone, we only trigger the learning module if the number of phonemes in the OOV term does not exceed a certain threshold.

9.5 Experiments and Evaluation

For the evaluation of the offline step, we processed a total of 225 publications of which 177 came with an abstract. Of those, 142 were automatically classified as English texts and used for keyword extraction. To evaluate the dialog component of the IslEnquirer, we recorded a total of 39 sessions with a total of 19 participants.

9.5.1 Dialog Evaluation

The IslEnquirer system depends on users that are willing to share their knowledge with the robot. This makes it necessary to evaluate the subjective quality of the dialog system. We handed out an evaluation questionnaire after each interaction. It lists statements which are to be ranked on a seven point Likert scale from strong rejection (-3) to strong accordance ($+3$). 27 statements were selected and presented to the users directly after their interaction with the robot. We collected a total of 18 completed questionnaires.

The average standard deviation is comparably high ($\sigma = 1.67$ on a seven point scale). We ascribe this to the different expectations people face the robot with and the fact that people succeed differently well in adapting to the capabilities of the system. The average overall impression is $+1.0$ with $\sigma = 1.21$. Most negative feedback, e.g. that some users did not feel understood well ($-0.6, \sigma = 1.58$), can be accounted to the rudimentary learning of OOV terms which has to be extended for future implementations. On the positive side, we find that most users rate the system to be very friendly ($+1.73, \sigma = 1.0$) and fun ($+1.2, \sigma = 1.56$) and rejected the statements that the dialog is too long ($-1.71, \sigma = 1.44$), that the interaction is boring ($-1.33, \sigma = 1.81$) or that they felt tense during the interaction ($-0.88, \sigma = 2.05$). Those are very important results as they promise long term acceptance of a dialog system acquiring social user models through interactive learning. They indicate that our efforts to keep the system interesting were successful and pay off by having willing users.

9.5.2 Social User Model Evaluation

To evaluate the social user models attributes, we perform a qualitative analysis. This is supported by our observation that even official sources are not always in accordance with the judgment of the participants of a pre-study and the experiments. They therefore cannot offer a baseline for evaluation. Asking the concerned people for their judgment is another evaluation approach and can at least offer a sanity check of the collected data. Note however, that the expressed opinion might be biased and does not reflect the real situation.

For every social attribute, we study two aspects: Are all corresponding attributes from the real world represented in the created set of attribute values and do all entries in the attribute set represent reasonable concepts in the real world? We exemplarily show this analysis for research groups but the results carry over to other attribute types, e.g. roles Putze (2008).

To get an understanding of the existing groups, we rely on the following cues: We regard all groups that were mentioned during the pre-study, during recording sessions and in interviews afterwards. This is complemented by checking the group candidates for spatial coherence (i.e. all members working in a limited number of places), temporal coherence (i.e. regular meetings), thematic coherence (i.e. shared research interests) and organizational coherence (e.g. a separate web site). The analysis shows that all main groups are represented in the social user models with perfect recall (while still containing some former members). The offline step contributed a starting vocabulary and initial knowledge used in the reliable closed question (91% success rate). It also found a group which was not mentioned by the users but supported strongly by the criteria listed above. This strengthens our assumption on the absence of a baseline. The contribution of the dialog on the other hand was to find more general attribute labels not present in the offline corpus and to add new groups which were not found by the offline algorithms.

To analyze the acceptance of our models by the concerned researchers, we performed a study among all institute members to determine the average approval of the extracted research interests. In total, we collected ten data sets. Each participant was presented a list of research interest labels (from the IslEnquirer and the Arnetminer website) associated to him. He was then asked to rank each item in terms of how well it describes his general field of study on a seven point scale from -3 to $+3$. As the IslEnquirer provides a ranking for its label assignment, we calculate a weighted average. The results are given in table 9.4. It shows that the results improve when dialog information is included (relative improvement of 16% resp. 33% for the average and the best result). The Arnetminer can only compete regarding the best hypothesis only (which usually is not enough due to the complexity of research interests) and only due to the fact that people without a Arnetminer page were excluded completely: less than half of all institute members had an Arnetminer page and those were ‘easy’ because of their large number of publications.

9.6 Conclusion

The approach to social user modeling is performed with a two-step approach using both offline sources and spoken dialog, which complement each other. Using methods of social network analysis and information retrieval, we auto-

Approach	**Avg**	**Top1**
Arnetminer*	0.48	2.5
IslEnquirer [offline]	1.05	1.8
IslEnquirer [offline+online]	1.22	2.4

Table 9.4: Evaluation results for research interest attributes: Averaged over all hypotheses for all users (Avg) and averaged over the best hypothesis for all users (Top1) (* = only calculated on available entries)

matically extract an initial social user model. Secondly, this model is verified and extended by human-robot dialog. The dialog strategy is designed to choose questions that promise a high information gain and a high success rate.

It is a first attempt to provide a humanoid robot with social awareness as part of interactive learning. This is achieved by building social user models, which store information on the relations between people, the groups they belong to and their roles within the community. Social user models are modeled to be robust against noise and able to handle multiple hypotheses.

By evaluating the collected data, we validate our hypothesis that social user modeling is possible using this combined approach. We see that both components contribute to the final model. A user study indicates that the dialog system is accepted by its users.

Chapter 10

Long-Term Evaluation of the interACT Robot Receptionist

While previous chapters address evaluation of single interactions, this chapter addresses evaluation of the knowledge base quality over time. Since the dialog system learns a knowledge base and updates the knowledge base over time, evaluation metrics, which assess the quality of the knowledge base, also reflect the effectiveness of a learning strategy and show superiority of a learning strategy with knowledge mending over pure information adding. In the role of the interACT Receptionist, the system learns a pre- structured knowledge base from scratch, and development of the knowledge base can be observed over a period of time, during which different people talk to the system. The system was installed in January 2008 and is still online at the date of writing. The time period of the presented evaluation was roughly eleven months, from January 2008 to November 2008, during which the system was online on workdays, with short maintenance breaks.

Successively the metrics for evaluation of the knowledge base and knowledge base quality are introduced, and the evaluation results are presented. Section 10.1 introduces evaluation metrics for assessing the knowledge base quality. Section 10.2 analyzes the development of the knowledge base over time according to the introduced evaluation metrics. It also analyzes different errors produced by the dialog strategies and compares the strategies with respect to their influence on the knowledge base. Section 10.3 presents experiments and evaluation with the knowledge mending approach.

10.1 Evaluation Metrics

The following evaluation paradigms are used as ground truth of person ID. The ground truth set is manually defined, and represents an optimal knowledge base that the system's knowledge base is evaluated against.

- **static (closure)** - all persons of a target set are added to a static knowledge base.
- **dynamic** - includes dynamic adding and removing of persons from the

ground truth set, which mirrors the real dynamics of a society/group structure.

- **onvisit** - persons are added to the ground truth set when they first visit the robot. In contrast to closure and dynamic, this category is suited to evaluate how good the system is with respect to persons that actually talked to the system.

Depending on the evaluation paradigm, different aspects are reflected.

The *static* paradigm mirrors how well the system's knowledge represents a ground truth set of persons, and describes how well the system represents each person that could talk to the system.

The *dynamic* paradigm mirrors how well the system's knowledge represents a real population at any given point in time, and also models the effect of persons changing their status, e.g. when leaving the institute. Therefore it is restricted to persons working at the institute and excludes visitors, since visitors would be represented in the ground truth set for a short period of time only and thus obfuscate the result.

The *onvisit* paradigm mirrors knowledge base quality with respect to persons who actually spoke to the system. And therefore, it highlights the errors that have been done by the system and not what has been learned so far. This paradigm is especially useful to measure the errors made by the system. It also does not immediately benefit from unseen persons talking to the system, but therefore it is not obfuscated by the time variability when persons first talk to the system. For this reason it also uses error metrics as explained in the following.

$$precision = \frac{\#truepositives}{\#truepositives + \#falsepositives} \tag{10.1}$$

$$recall = \frac{\#truepositives}{\#representatives} \tag{10.2}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{10.3}$$

$$EER = 1 - precision \tag{10.4}$$

$$SER = \frac{\#incorrectlylabeledsessions}{\#allsessions} \tag{10.5}$$

$$F1ER = 1 - F1 \tag{10.6}$$

The evaluation measures used for the three paradigms originate from information retrieval. There, precision (10.1), recall (10.2), and their harmonic mean, the f-measure (F1) (10.3), have become popular to measure retrieval

abbr.	long name	description	values
EER	Entry Error Rate	measures the precision of entries	EER [%] ranges from 0% (best) to 100% (worst)
F1ER	F1 Error Rate	relates to the f-measure (F1) and is the harmonic mean of EER and error of recall	F1ER [%] ranges from 0% (best) to 100% (worst)
SER	Session Error Rate	incorrect session labels divided by all sessions	F1ER [%] ranges from 0% (best) to 100% (worst)

Table 10.1: Overview of the main evaluation metrics

and classification results. These measures are also appropriate for an evaluation of the knowledge base, by evaluating how many persons of a ground truth set are covered by the knowledge base and how many persons were learned incorrectly. The same measure can also be applied to correct recognition of names and other attributes. An overview of the metrics EER (10.4), SER (10.5), and F1ER (10.6) is shown in table 10.1. Since in our scenario a person is always learned with first and last names, the precision/recall measure is equivalent to computing the error rate of names. The name error rate is shown in the following equation (10.7).

$$nameER = \frac{wrongnames}{wrongnames + correctnames} \tag{10.7}$$

The error measures used for the *onvisit* paradigm, represent the error equivalent to the precision, recall and f-measure, and are computed as shown in equations 10.4 to 10.6. In contrast to the *static* and *dynamic* paradigms, the f-measure for the *onvisit* paradigm starts at 100%, which corresponds to an F1 error rate (1 minus f-measure) of 0%.

When evaluating the improvements that are made by the mending dialogs, not only the improvements of the knowledge base should be measured. As the improvements are largely obtained through interaction with a user, there is also a relevant question of how much the human ‘information source’ may be used. In general, a high number of unsuccessful dialogs can be considered as bad, and we consider disturbing the user without positive outcome an annoyance factor. However, if the dialogs result in improvements, these dialogs have a large benefit, which is good. Both aspects are combined in the benefit rate in equation 10.8. It relates the obtained corrections to the number of questions necessary to obtain these corrections. A benefit rate of 100% can be obtained by one correction and one question, but also by ten corrections with ten questions. Note that the benefit rate theoretically can be larger than 100% if more than one correction is made with a single question, and that the benefit rate can be negative if the dialogs produce more errors than corrections.

$$benefitRate = \frac{correctchanges - incorrectchanges}{\#questions} \tag{10.8}$$

In relation with the benefit rate, it is also relevant to measure how many labels are checked by the dialogs, as only a large number of dialogs can check all labels at least once and a small number of dialogs relies on a good benefit rate. This is measured by the labelClarificationRate in equation 10.9.

$$labelClarificationRate = \frac{\#questions}{\#labels} \tag{10.9}$$

10.2 Analysis of Learning Results

10.2.1 Analysis of the interACT Robot Receptionist

The result of applying the evaluation measures to the learned knowledge base is shown in figure 10.1. It shows an evaluation over time of the knowledge base of the interACT receptionist. It displays three graphs with different evaluation paradigms and metrics. The first graph considers all persons including visitors and evaluates against the closure, i.e. the static paradigm. The second graph considers only lab persons and evaluates the dynamic paradigm. It also shows the same measures applied to the website. The third graph also considers only lab persons and evalutes the *onvisit* paradigm. A person entry is counted as a correct entry, if the person confirms the learning result as being correct and the robot has created an entry, by which it can identify the person again. This does not necessarily mean that each name spelling is correct. For example if a person is called 'Stephan' and an entry is created with the name 'Stefan' (which is a typical error if the system does not have a prior model of possible names), this counts as a correct entry. However, if two entries are created, 'Stephan' and 'Stefan', then one of them is counted as an error. An evaluation including correct spelling of first and last names is presented later in this section.

The evaluation charts show stable development of the knowledge base over time, with a tendency of obtaining worse results after longer periods of runtime. However, as some of the charts show, there's room for improvement regarding incorrectly learned persons. In fact, these plots evaluate only

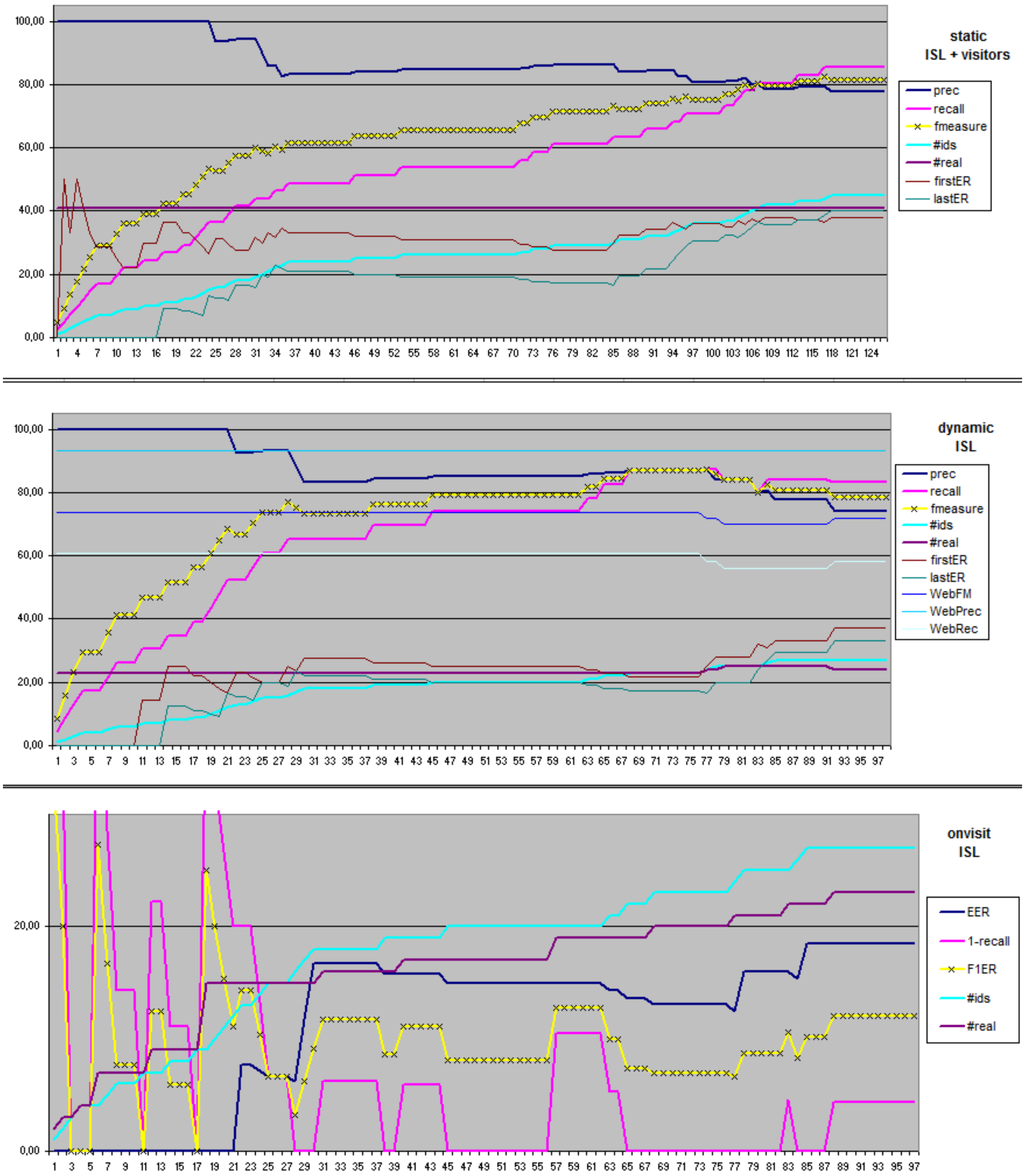


Figure 10.1: Evaluation of the knowledge base over time, displaying static, dynamic and onvisit paradigm evaluations of the knowledge base learned by the interACT receptionist

adding persons to the knowledge base. Knowledge mending, generally speaking, serves two purposes. First, inaccurate learning results can be removed from the database, which is important to remove persons that have been learned with incorrect names, or even persons that have been learned twice with different names. Second, changes of the ground truth set can be incorporated, e.g. a research assistant has left the institute. Before we evaluate the knowledge mending approach, the follwing section assesses the contribution of the dialog setup by comparing the learning curve of the dialog approach against a classification approach without dialog usage.

10.2.2 Comparison against FaceID Learner

To get a rough understanding, how the dialog-based learning approach compares against a classification approach without dialog, we have conducted an experiment with evaluation over time using face identification output for learning decisions. The model is simplified as it simulates a learning curve over the complete data corpus, for which it uses confusion rates of the classifier from the receptionist scenario. The experiment rather produces a qualitative than a quantitative comparison between both approaches, and may provide indications with some restrictions. The faceID learner reproduce the knowledge base update functions to add a new entry to the database or to update an existing entry to the database. Therefore both approaches can be compared by the recall and precision-based metrics. As the faceID learner cannot obtain the name of a person, the knowledge base update functions must be adapted to ID-classification without name information. In contrast to the dialog scenario, the faceID learner has the advantage to apply offline learning, i.e. the complete set of images can be seen before a decision is made, and that sessions without face recognitions are ignored as no faceID hypotheses are generated during these sessions. Still it is an interesting comparison, as the related work, as described previously, uses similar learning methods with successive labeling, and this analysis shows the difference of a dialog scenario with real user interactions versus a pure classification approach, where a supervisor labels each unknown class.

The basic design of the learning component is that the face identifier either classifies a person as known with a certain ID, or as unknown. This is done successively for each recorded session in the data corpus. In case the person is classified as unknown, a new entry is added to the database with the correct label. In case the person is classified with an ID, the corresponding ID model would be updated, which has no effect, except in case where the wrong ID is classified, which increments a corrupted-models counter. We imply that labeling of new entries is perfect, e.g. it is done manually after the learning has been conducted, while the dialog approach deals with recognition errors. The classification output is simulated by applying a probabilistic confusion model, which is shown in table 10.2.

real state	correct	confuse with known ID	confuse with unknown
unknown	0.8901	0.1099	0
known	0.8053	0.1155	0.0788

Table 10.2: Confusion model of face identification

An evaluation plot, which shows the learning curve on the same session data as the interACT receptionist evaluation is shown in figure 10.2. The charts show evaluation plots of the static and the onvisit paradigms. This learning mechanism has some specifics regards the errors. An error for a known person can either be that the person is identified as unknown (a person is invented) or that the person is mixed with a different person (then the face ID model is updated with mixed data and thus gets corrupted). A special "corrupted" category is also evaluated for the onvisit paradigm, which does not count corrupted IDs as errors, and thus shows better results than the standard evaluation. The "corrupted" category also illustrates the problem of the faceID learner, and why the dialog approach is superior. More persons are misrecognized in the faceID approach either as another person or as a new person, which has a negative effect for the knowledge base. While the f-measure in the static paradigm for the interACT receptionist steadily increases and achieves a maximum of over 80%, the f-measure of the faceID learner decreases interim and ends at below 70%. The effect is even more obvious in the onvisit paradigm, where the f-measure steadily decreases after an initial learning phase.

The results are even more encouraging, as the faceID learner assumes that each session contains enough face images of a person, which is not the case for the real interactions of the interACT receptionist where some of the dialogs are conducted without face recognition.

10.2.3 Discussion: Dialog Success and Knowledge Base Quality

The first to sections of this chapter have analyzed the long term development of the interACT receptionist. Analyzing the development over time shows generally stable learning results of the system. But after some time, the error rate (knowledge base) increases, even though the dialog success is unchanged and even increases as the identification models are adapted to the modeled persons. Three evaluation paradigms, *static*, *dynamic*, and *onvisit*, were analyzed and each paradigm shows different aspects. The *dynamic* paradigm motivates the necessity of knowledge corrections, as here it becomes obvi-

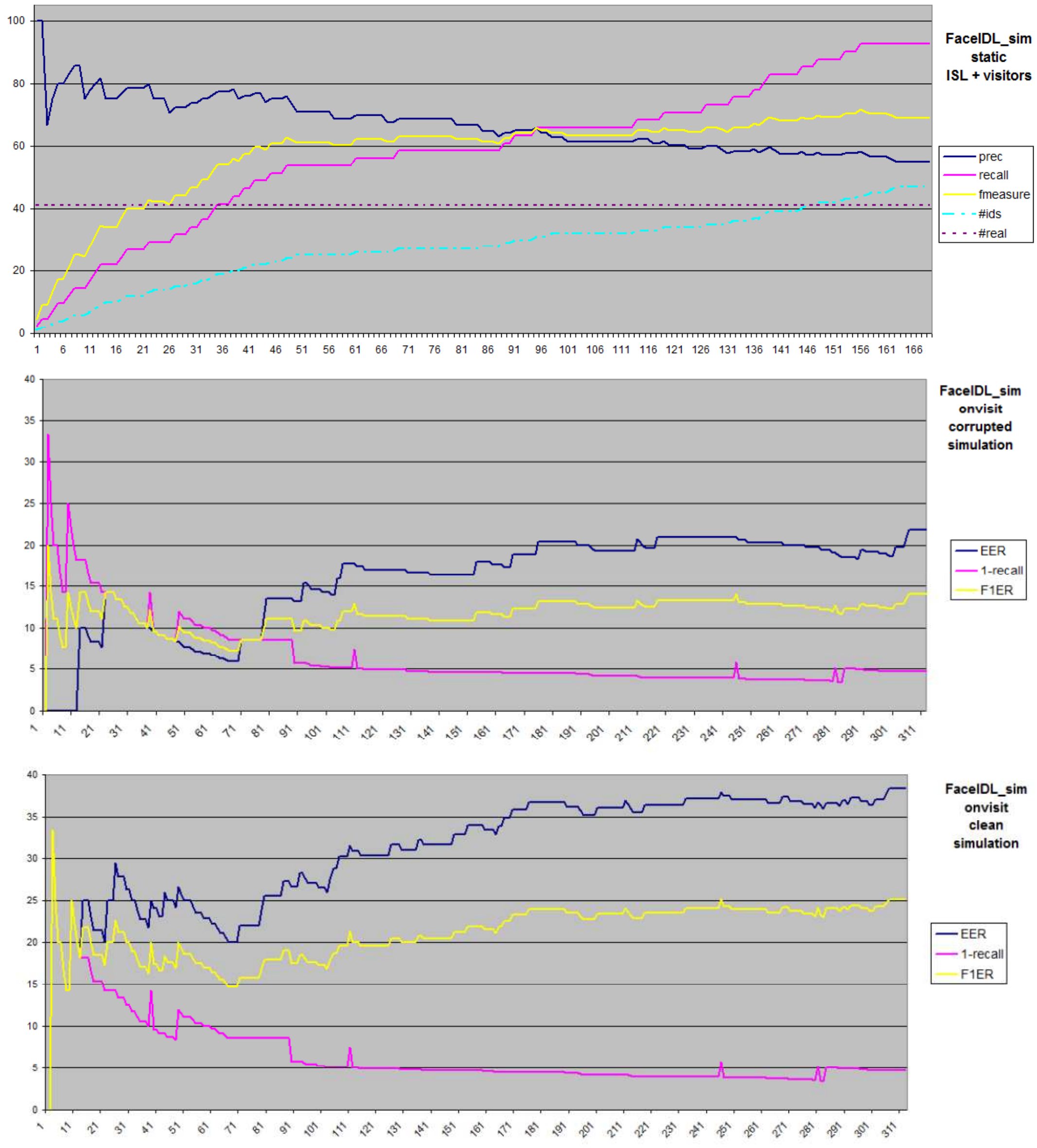


Figure 10.2: Evaluation figures for the FaceID learner simulation

ous that the knowledge base quality decreases after some time. While the *dynamic* and *static* paradigms are strongly influenced by events in the environment and when – or if at all – persons talk to the system, the *onvisit* paradigm is most helpful to analyze errors that are induced by the dialogs.

There is also a question of when to count a person as "has been learned correctly", i.e. is an entry correct if the name is pronounced correctly, if the person agrees that the name is pronounced correctly, or if the name is written correctly. In the first part of this chapter an entry was counted as being correct, when the person intentionally confirmed the learned name as being correct. As described before there are slight differences to correct spelling, especially visitors, whose names have not been in the vocabulary before the interaction and who had to teach the robot their name by spelling, sometimes confirmed a name which sounds correct but does not have the right spelling. Especially the spelling part was hard to complete and sometimes tedious for some persons, as almost all visitors are non-native English speakers and not comfortable with spelling words in English. In addition, the spelling recognizer is a part of the system which needs improvements, as we did not have an up-to-date spelling recognizer at hand and therefore used a prototype system with acoustic models from standard speech recognition from Ziesemer (2007). The system could be improved by acoustic models trained specifically on spelling data, as has been done in system which achieve higher recognition accuracy.

A further point of improvement is pronunciation by the text-to-speech system. As many names have been German names, and the language during the interaction was English, the German names were pronounced with English letter-to-sound rules. In addition, the synthesis was not always clear to understand, as some people stated in their questionnaires. Therefore, a few initially correct names were rejected at first, and some names were accepted by the persons as they could not tell that the name was not pronounced fully correctly. Interestingly, no one of the users asked the robot to spell the learned name, which would have been the only way to tell if the name was learned correctly.

Another interesting analysis shows the effect of a dialog success rate on the development of knowledge base quality. The outcome of the dialog-based identification module is one of four cases: Correct identification (corr), no identification (noid), wrong identification and confusion with other person (corrupt), false learning and creating a nonexistent person (invent). Thus, there are three different error categories. Risk calculation of which errors have worse effects on the knowledge base, requires to look at how often persons talk to the system. A histogram of visiting persons is shown in figure 10.3. The y-axis denotes the frequency of a person visiting the system and the x-axis denotes the rank of the frequency. The diagram shows three other functions to approximate the histogram by common statistics. The function 70/rank is a simple estimate of 1 divided by the rank of the person multiplied by a constant factor. The functions Zipf,s=1 and Zipf,s=1.1 follow Zipf's law, which was originally designed to model the statistics of common words in texts Zipf (1965), and is described by equation 10.10, where r is the rank, N is the total number of entries, s is the value of the exponent characterizing

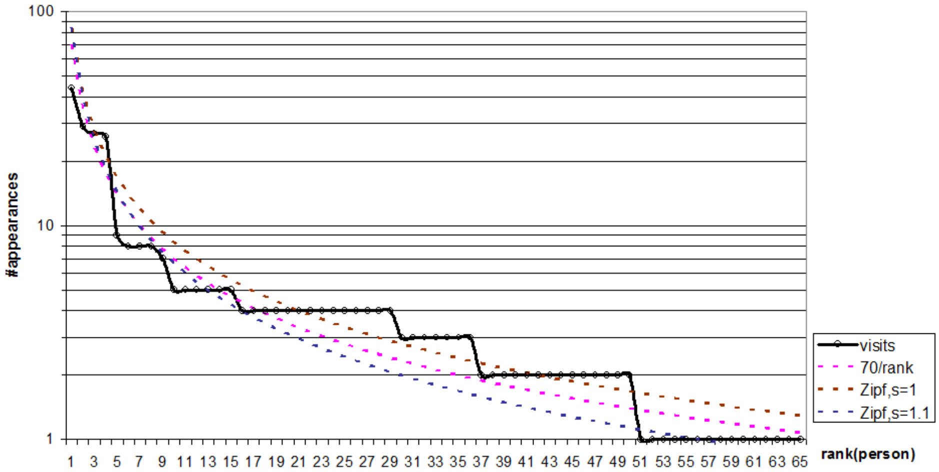


Figure 10.3: Person histogram of visits, and Zipf's law

the distribution[1]. In the classic version of Zipf's law, the exponent s is 1. Figure 10.3 shows Zipf's law for s=1 and s=1.1.

$$f(r,s,N) = \frac{1}{r^s} * \frac{1}{\sum_{n=1}^{N} 1/n^s} \tag{10.10}$$

To give an example how Zipf's law can influence decisions of the dialog strategy, we looked at the frequency statistics. Persons who talked to the system very infrequently, e.g. only once or twice, are mostly visitors. As the task of the interACT robot receptionist is to model only a network of persons working at the interACT lab, it can qualitatively be inferred that the 'noid'-error is not as severe as the corrupt and invent errors, which both decrease the quality of the knowledge base. And therefore during dialog strategy training, it should be given more weight to getting the correct name of a person and rather abort a dialog than to conducting short dialogs with a higher risk of obtaining incorrect names. However, as subjective user feedback shows that users prefer short and successful dialogs, both aspects should be balanced, and we can imagine working with different confidence levels of user ID for different knowledge base operations in the future. A tradeoff between correct identification and short dialogs had also been chosen in the reward function for the training of dialog strategies in chapter 5. However, this tradeoff between different error categories is shifted again, when we consider that some error categories can be resolved better by the system, which is discussed in the next section.

[1] http://en.wikipedia.org/wiki/Zipf's_law

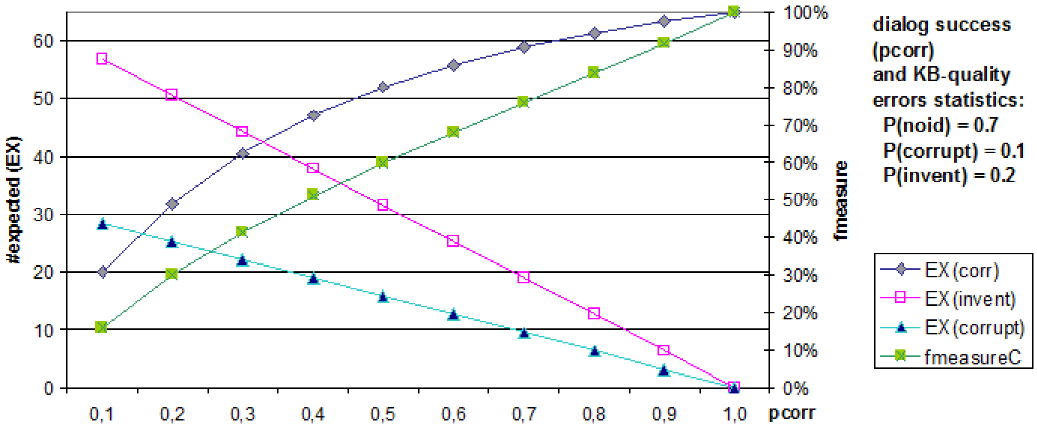


Figure 10.4: Plot of dialog success, resulting knowledge base modifications and f-measure representing knowledge base.

With the data studied in this chapter, we can further analyze the effects of dialog errors on the long term results of the knowledge base quality. From the results of the dialog strategy in the interACT receptionist in identifying persons we can estimate the expected error categories as follows. P(noid|error)=0.70, P(corrupt|error)=0.1 and P(invent|error)=0.2. The partition of an error into these three error categories remained fixed, and the probability of error P(error) was used as a variable from 0.0 to 0.9. Figure 10.4 shows the plots of knowledge base errors obtained from a dialog simulating which replays the recorded corpus with the given the dialog error configuration. It can be seen that there is more or less linear correlation between the dialog success rate and the f-measure. The analysis however is kept quite simple and represents a lower bound of expected errors, as the error simulation does not include that the person identification rate might get worse with more knowledge base errors. These results are not surprising, as they show that better dialog design leads to better knowledge base quality. But they provide an estimate to understand what kind of improvements of the knowledge base can be expected when improving the dialog strategy by a certain degree.

10.3 Knowledge Mending Results

Evaluation of the knowledge mending approach has been conducted by applying the already defined knowledge-base quality metrics. Three approaches are compared with each other, the 'noClustering', the 'offline' and the 'online' approach. The baseline approach, i.e. the standard learning behavior without knowledge mending, is labeled as 'noClustering'. Following the description of the mending approach in section 6.7, two knowledge mending approaches are evaluated. The 'offline' clustering approach conducts a pure non-interactive mending strategy by clustering. Each cluster is assigned the label which occurs most frequently in the cluster, on a balanced set of samples. The clustering step is executed once per day. Two different error levels for the stopping criterion of the agglomerative clustering algorithms have been chosen for evaluation. As we have seen during clustering confidence training, an error level of 5% induces only little errors, while at a level of 10% more severe errors can occur, which corresponds to our intuition. At a low error level, a small amount of correctly labeled sessions are expected to be merged incorrectly. At an error level of 5% we expect this error category to be rather insignificant, especially since most persons have interacted with the system more than once. In this case, a single discarded session does not influence the evaluation result. This kind of fusion error can be resolved by the online approach in some cases. However, when the error level is too high, i.e. larger than 10%, we expect that such errors occur more frequently, which would then form clusters of existing persons and thus should degrade the clustering result. Therefore, we generally favor an error level of 5%.

The online clustering approach is a combination of the offline clustering with dialog interaction for existence checks with a trusted person, as described in section 6.7. Evaluation was conducted by automatic day-wise clustering and conducting mending dialogs with the author of this thesis as the trusted person.

The evaluation set includes 106 sessions from the automatically recorded corpus, which have been left out from the confidence training.

10.3.1 Mending Results for All Interactions

Due to the obviously higher error rates in the visitor category, i.e. mostly incorrect spelling of names, evaluations are shown separately for the full interaction set including visitors and the set restricted to persons from the interACT lab (i.e. employees and students), who have significantly lower error rates. Each figure shows a graph of

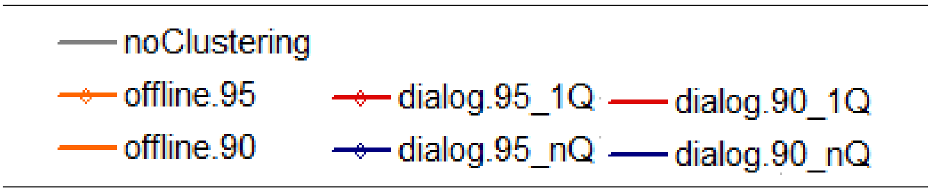


Table 10.3: Legend for evaluation plots: noClustering baseline, offline and dialog with single-question dialog (1Q) and single-cluster dialog (nQ), and confidence levels of 90% and 95%

- ‘noClustering’ — learning over time without mending
- ‘offline’ — learning over time with offline clustering for 90% and 95% confidence levels and automatic label correction
- ‘dialog’ — learning over time with offline clustering for 90% and 95% confidence levels and interactive mending dialogs

and the two dialog conditions

- ‘1Q’ — dialog is restricted to only 1 clarification question, i.e. problematic label
- ‘nQ’ — dialog is restricted to only 1 cluster, i.e. ‘n’ questions to clarify contradictory labels in problematic cluster

Table 10.3 shows the legend for the coding of these categories in the following figures. The knowledge base quality plots for the full evaluation data set is shown in figures 10.5 (Entry Error Rate), 10.6 (F1 Error Rate), and 10.7 (Session Error Rate).

The index of the graphs is incremented by 1 for each interaction, mending dialogs have odd numbers, e.g. 7.5. Thus, the total number is a counter for interactions. The set includes 23 mending intervals and the plots show characteristic spikes (mostly downward), were the mending takes place. It can be seen that the number of errors significantly decreases (shown by the entry error rate). The improvement of the F1 measure can mostly be explained by the improved precision of the entries. The recall changes insignificantly, as the mending approach does not improve the recall. Note: a recall below 100% means that some persons are not modeled by the database. These would be added by the learning dialogs, not by the mending approach.

The figures also show comparison of the dialogs restricted to one question (1Q) and the dialogs restricted to one cluster (nQ), to evaluate if even the strong restriction provides improvements to sorting of the problematic labels.

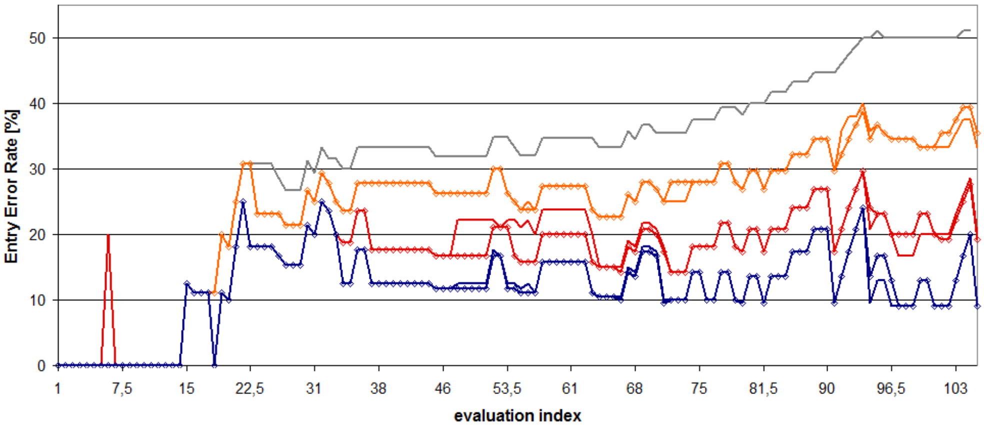


Figure 10.5: EER – mending results full corpus – baseline, offline and dialog approaches with confidence levels of 90% and 95% – legend in table 10.3

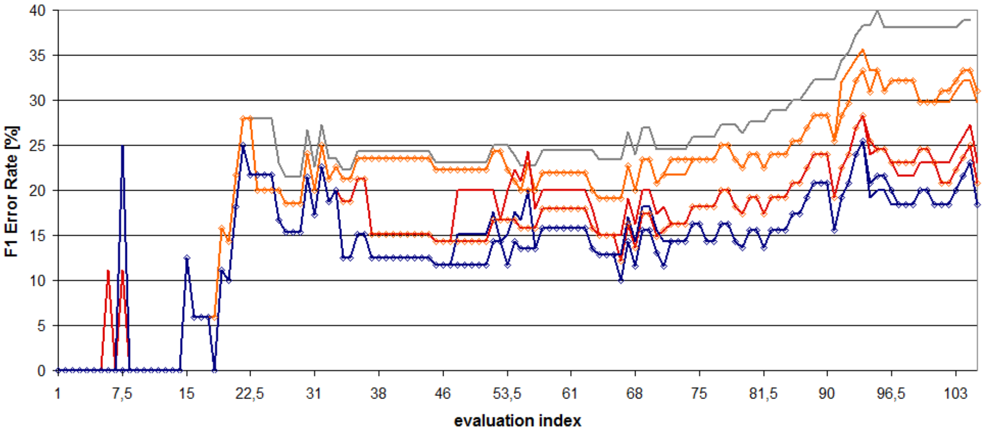


Figure 10.6: F1ER – mending results full corpus – legend in table 10.3

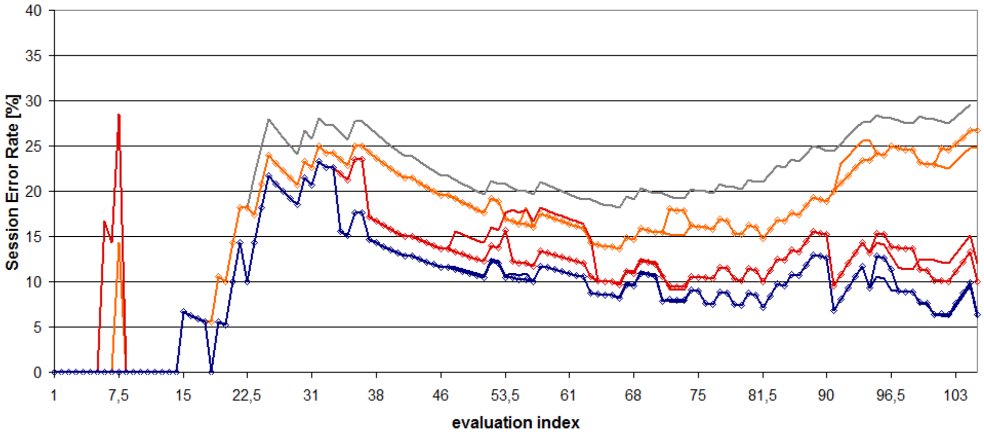


Figure 10.7: SER – mending results full corpus – legend in table 10.3

dialog.95_1Q	**dialog.95_nQ**	**dialog.90_1Q**	**dialog.90_nQ**
92,0%	74,3%	84,0%	70,3%

Table 10.4: Benefit rate for the full corpus

At the final stage, the baseline has 51.1% EER, the ‘nQ’ and single-question dialog tasks (1Q) have 9.1% EER and 19.2% EER for a 95% confidence level and 9.1% and 20.0% EER for a 90% confidence level. Correspondingly, we can look at the benefit rate and at the label clarification rate (equation 10.9). The label clarification rate is 78% for nQ with 95% confidence level, 82% for nQ with 90% confidence level and 56% for 1Q dialogs, meaning that for the 1Q dialogs, 54% of the labels have not been checked at all during the dialogs. From the results we conclude that problem selection provides a good selection of problematic labels. The corresponding benefit rate is shown in table 10.4.

10.3.2 Mending Results for In-Domain Set

Even better results, especially in terms of the f-measure, are obtained when looking only at the set of persons that were entered into the database correctly at least once, termed ‘in-domain’ persons. The set includes all persons from the interACT lab plus some of the visitors. The numbers in this analysis are significantly better than the results obtained for all persons including visitors. This is mostly due to better prior name models, i.e. better speech recognition vocabulary, which are obtained from social network analysis for members of the interACT.

The results for the 90% and 95% confidence intervals are shown in figure 10.8, figure 10.9, and in figure 10.10. Table 10.5 and table 10.6 give an overview of the final results. The corresponding benefit rates are shown in table 10.7. The significant improvement of the entry error rate shows that the entries are very precise after the mending, along with more than 50% reduction of the F1 error rate, as the recall remains constant. Also the offline approach alone achieves significant improvements.

10.3.3 Summary and Discussion

The presented approach enables a robot to proactively detect and correct information stored in its database with significant improvement over the baseline system. The highest improvement is achieved by the dialog approach which solves all labels in a cluster per dialog (tagged ‘nQ’) with relative improvements of 92.7% SER from 17.8% to 1.3% and 88.3% EER from 40.5%

conf level	metric	results			improvements	
		learner	offline	dialog	offline	dialog
95% /1Q	Session Error Rate	17.8	12.2	3.9	31.3%	78.1%
	F1 Error Rate	25.4	16.7	11.1	34.4%	56.3%
	Entry Error Rate	40.5	23.1	13.0	43.1%	67.8%
90% /1Q	Session Error Rate	17.8	12.2	2.6	31.3%	85.2%
	F1 Error Rate	25.4	16.7	9.1	34.4%	64.2%
	Entry Error Rate	40.5	23.1	9.1	43.1%	77.6%

Table 10.5: Mending results for in-domain persons and relative improvements for the offline and dialog_1Q approaches

conf level	metric	results			improvements	
		learner	offline	dialog	offline	dialog
95% /nQ	Session Error Rate	17.8	12.2	1.3	31.3%	92.7%
	F1 Error Rate	25.4	16.7	7.0	34.4%	72.5%
	Entry Error Rate	40.5	23.1	4.8	43.1%	88.3%
90% /nQ	Session Error Rate	17.8	12.2	0.0	31.3%	100.0%
	F1 Error Rate	25.4	16.7	4.8	34.4%	81.3%
	Entry Error Rate	40.5	23.1	0.0	43.1%	100.0%

Table 10.6: Mending results for in-domain persons and relative improvements for the offline and dialog_nQ approaches

to 4.8% with the 95% confidence level, and relative improvements of 100% SER from 17.8% to 0% and 100% EER from 40.5% to 0% with the 90% confidence level. The configuration of the clustering approach with 90% and 95% confidence levels produce comparable results, and when looking at the full set of all interactions, the results are even identical. However, the 90% confidence level has higher variation, and due to taking higher risks, some errors can be solved earlier in the dialog, but also irrecoverable errors are more likely to occur than with a confidence level of 95%.

To measure how many dialogs are necessary to obtain a certain level of improvement of the knowledge base, the benefit-annoyance ratio, or short

dialog.95_1Q	dialog.95_nQ	dialog.90_1Q	dialog.90_nQ
54.2%	46.9%	58.3%	48.5%

Table 10.7: Benefit rate for the in-domain corpus

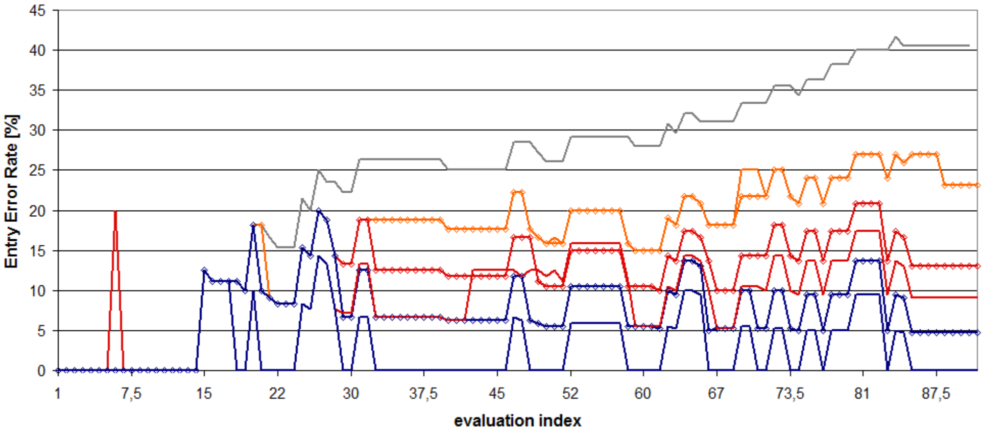


Figure 10.8: EER – mending results in-domain – baseline, offline and dialog approaches with confidence levels of 90% and 95% – legend in table 10.3

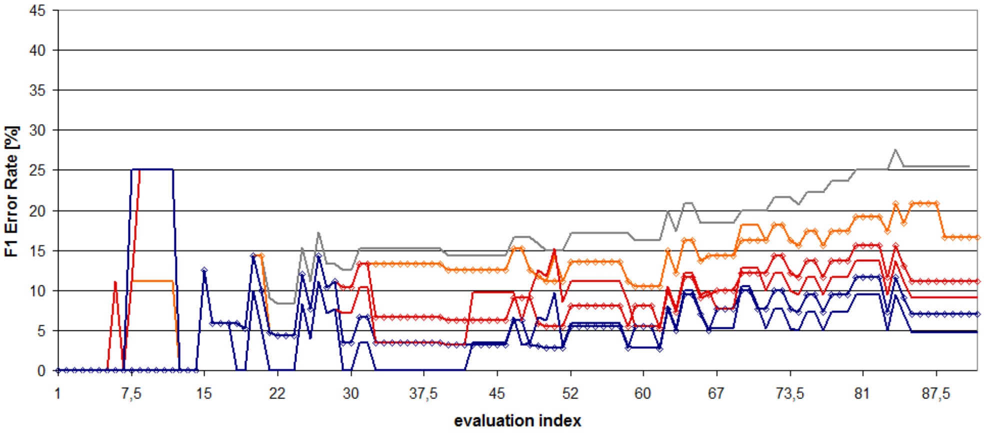


Figure 10.9: F1ER – mending results in-domain – legend in table 10.3

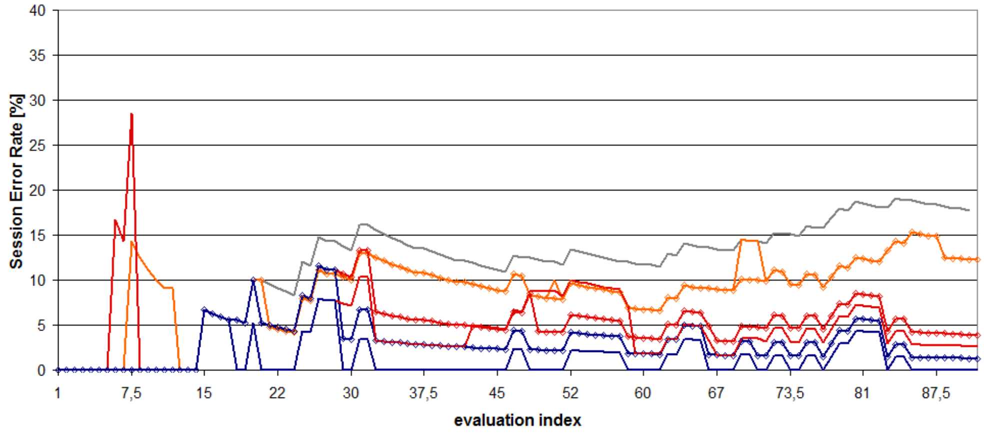


Figure 10.10: SER – mending results in-domain – legend in table 10.3

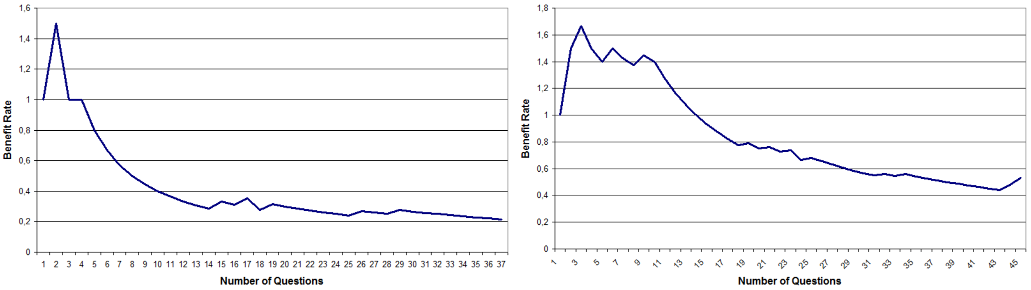


Figure 10.11: benefit rate with mending at once; left: in-domain, right: full corpus

the benefit rate, was introduced in equation 10.8. It relates a measure of improvement (the benefit) to the number of dialogs conducted (the annoyance). As the benefit rate is high for a small number of dialogs, i.e. 92% for dialog.95_1Q, and gets smaller for a larger number of dialogs, i.e. 74% for dialog.95_nQ, we conclude that the problem detection mechanism does a good job in finding problems in the data set, and that this helps to restrict the interactions to a small number without annoying the user, i.e. 25 questions were posed for the full data set in dialog.95_1Q.

To analyze the benefit rate in further details, it would be desirable to analyze the plot not only for two values, but for a varying number. A plot over the 'online' system, i.e. the same evaluation method as just reported with mending dialogs during the learning phase, has only limited explanatory power, as with new interactions, new errors are induced. But, a meaningful plot can be created with 'mending at once' where all mending dialogs are conducted in a single session after the learning dialogs. Figure 10.11 shows these plots for the in-domain set and for the full corpus with a confidence level of 95%. Both plots, and especially the in-domain set which contains corresponding correct entries for almost all incorrect entries, show a peak in the beginning and then a decrease in benefit rate, which indicates that most errors are resolved early and therefore the problem detection algorithm offers adequate pre-selection. The plot over the full corpus does not decrease as strongly as the plot over the in-domain entries. The interpretation for this is that the full corpus does not contain correct entries for each person. Therefore, the clustering step does not group together sessions from the same person with different information. For example, the slight increase of the benefit rate at the end happens because one person interacted with the system frequently but each time confirmed the same incorrect name. The clustering step now could group together different sessions, but all containing consistent information. Therefore, from the system's point of view, these sessions cannot

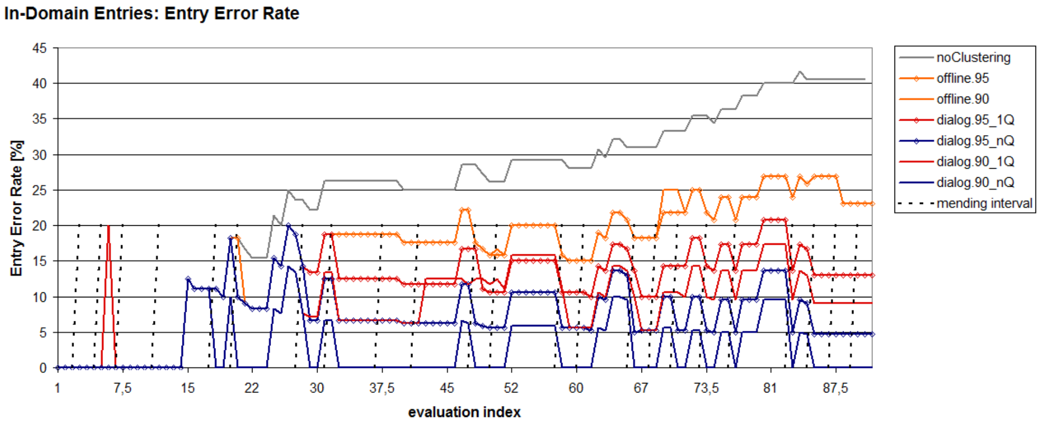


Figure 10.12: EER plot with annotations where mending dialogs have been conducted, based on figure 10.8

be distinguished from a person who interacts several times with the system and is identified by the right name. A possible solution could be to use additional features extracted from the dialogs, e.g. measuring uncertainty of having learned the name correctly, instead of pure database information. In case of the in-domain set, the ranking of errors are better, as for the in-domain set, incorrectly labeled sessions are grouped together with correctly labeled sessions leading to contradictory information which can be solved in dialog. The results are a strong peak of the benefit rate followed by rapid decrease.

Another way to interpret the success of the mending approaches qualitatively in the online system is by looking at the improvements achieved by each mending interval. Figure 10.12 shows the annotation of the mending intervals in the Entry Error Rate plot. It can be seen that drops of the error rates co-occur with mending dialogs, and that quite often errors that are made by the system can mostly be corrected by the next mending dialog. This is most obvious between interactions (i.e. evaluation index) 67 and 75.

Another conclusion is that the mending dialogs can either be conducted during the runtime of the system or as a final run 'at once'. The advantage of conducting mending dialogs during the lifetime of the system is that the quality of the knowledge base is better throughout the system's lifetime, and it can be argued that the recognition components can be trained with cleansed data.

Already by using automatic agglomerative clustering, 30% - 50% of the errors can be either solved or at least detected, before the clustering produces the first fusion errors. As clustering with modifications of the database can lead to irrecoverable errors, the database states before and after clustering are maintained, and stored in a background model, so that dialog interactions for mending are conducted with the unmodified database. Also as more persons are added to the database the clustering approach gets more reliable and stable than during initialization phase, which is another reason to maintain the original model without early hard error correction.

When looking at the errors that exist after application of the offline clustering approach, it can be seen that most of these originate from the system incorrectly creating an arbitrary person entry (which has been termed "inventing a person") – which is addressed by interactive error resolution in dialog. Some errors are due to incorrect name spelling, which is predominant in the visitor category, but infrequent in the employee category due to good prior vocabulary models. In case of the invented entries, almost 80% of the errors have been resolved by mending dialogs, which conduct a clarification strategy, e.g. by asking a trusted person if the person with the label in question exists, and if so, if the person is employed at the institute, is a student, or a frequent visitor.

The mending dialogs are conducted with a trusted person who has to have some knowledge about the population, as otherwise, information cannot be confirmed. The presented approach enables a robot to proactively detect and correct information stored in its database. It is not the intent of this approach to provide a speech interface for a human operator who modifies the database via speech instead of keyboard input. Therefore, the interaction is initiated by the robot only. By restricting the type of interaction to this style, the recall is not improved significantly, as it is left to the learning dialog to acquire this information from interaction with the user. When we take a look at how humans solve such a task, there is still room for improvement by other types of mending dialogs, e.g. to correct the pronunciation of a name if the user can assume who the robot refers to but recognizes that the name is not fully correct.

We also assume that further potential exists in improving the offline clustering approach. Though the benefit rates show that problematic labels are detected early, a few potentially problematic clusters remain undetected, which likely originate from different light conditions due to the long term of recording and moving of the platform. As this is in general a challenge of current face identification works, and the applied feature extraction methods are already robust against partial occlusion and different light conditions up to a certain degree, we can expect additional improvements by new algorithms and robust feature extraction, e.g. for images with and without shadow, light from the front or from the side, etc. In the course of this thesis, only a limited set of clustering methods could be tested. Additional distance features or clustering methods could improve the offline clustering and the benefit rate.

10.4 Conclusion

This chapter has presented an evaluation of the interACT receptionist in a long term study in a real-life environment. It could be shown that the system can acquire information about a population and maintain this model over a longer period of time including error correction. While single dialog interactions have shown high success rates and initially the learning curve of the system increases, the system reaches a point of saturation at some point with optimal knowledge base quality. After this point, the knowledge base quality decreases due to the accumulation of errors that happen during the interaction and due to real-life conditions such as severe background noise or unpredicted kinds of interaction. By knowledge mending, a successful approach has been presented which significantly improves the knowledge base quality by offline clustering and interactive dialogs.

In contrast to approaches such as active learning or data cleansing in databases, the presented approach is not intended to allow a human to 'hand-label' some kind of data set. A learning system such as studied here has internal knowledge about which it can communicate, but which does not allow direct access by a human annotator. Such a system could in the future be applied to a humanoid robot to proactively acquire information about its environment, and as it could be shown here, also has the ability to extend this learning process over a longer period of time and correct induced errors.

Chapter 11

Conclusion

11.1 Summary

This thesis has introduced a dialog-based learning approach with unsupervised learning mechanisms for acquiring information and maintaining a knowledge base. In contrast to supervised learning, such a learning mechanism enables robots to learn and extend their knowledge autonomously without manual intervention by a human supervisor, which is an important ability when used in an open set real-world environment. It could be shown that the approach can be realized for effective knowledge acquisition with robust dialog strategies and for different learning tasks, and that the approach can be applied in a realistic task over a longer period of time through ongoing automatic knowledge maintenance with offline processes and dialog interaction.

In part I, a fully integrated framework for human-robot interaction has been presented, including techniques for robust multimodal dialog processing and a novel framework for user identification in dialog including dialog strategy optimization. It could be shown that the presented techniques improve recognition accuracy by tight coupling of recognition components and produce reliable user identification results in a natural interaction. It could also be shown that the presented multi-layer user identification approach improves identification accuracy with confidence estimation, sequence hypotheses and multimodal fusion, and that additionally, the dialog achieves significant improvements over non-interactive identification. Furthermore it was shown that a multimodal user simulation can be used to train dialog strategies by reinforcement learning and that dialog strategies trained this way lead to better dialog results on independent data sets and in real user experiments.

In part II, a dialog-based learning approach has been presented with a modular dialog concept and dialog strategies for acquiring information and maintaining a knowledge base. The dialog-based learning approach comprises a complex system which builds on the framework presented in part I. It could be shown that knowledge acquisition and maintenance is possible in a real-world setting in a long term study, as information can be added reliably and that errors, which result from long term usage of the system, can be detected with high precision. Erroneous entries can autonomously be detected and be removed by the system, which leads to precision of knowledge base entries of over 95%, and in advantageous situations of 100%. It could also be shown that knowledge acquisition is possible with a generic knowledge entity model in a complex system, which comprises learning of multimodal knowledge sources of recognition components, i.e. speech recognition vocabulary, grammar, face identification, and voice identification, semantic models, i.e. natural language understanding grammar, ontology, and environment model, i.e. objects, persons, and associated attributes and relations, e.g. social networks.

11.2 Discussion

11.2.1 Evaluation

In part I and part II different evaluation methods have been applied to measure the success of the systems. Evaluations in part I are based on well defined metrics used in the dialog systems community, which allows to assess subjective and objective measures in a quasi standard way. Both, subjective and objective metrics have been used to assess the system performance, and especially objective measures can demonstrate the improvements by approaches such as tight coupling, multimodal user identification and dialog strategy learning.

Part II focuses on analyzing different aspects of dialog-based learning, for which less standardized metrics or evaluation standards exist. In case of social network modeling it was possible to compare the dialog-based learning approach against other publicly accessible services and demonstrate better results by the dialog-based learning approach. To be able to assess the behavior of the proposed methods, gold standards are introduced to measure the success of different techniques, measure relative improvements when combining different techniques, and to evaluate an overall system. Such an overall system evaluation was conducted for the interACT receptionist in a long-term study in a realistic environment. With the goal to set up evaluation metrics which are easy to understand by humans, i.e. one can personally estimate how well the system performs, an evaluation scenario was created for the interACT receptionist. Its performance is measured by how well it can model a group of persons working in the interACT lab and present the result on a “Who-is-Who” page. For such kind of evaluation we could apply known metrics from other areas, namely adopt the precision/recall measure, to define a metric for knowledge base quality. With the assessment of knowledge base quality, a learning curve of the system can be plotted to study its learning behavior over time. By using the same evaluation metrics, the dialog-based approach for knowledge mending has significantly reduced errors in the knowledge base to a minimum of less than 5%, in optimal cases to 0%.

11.2.2 Outlook

This thesis has the intent to study a complete multimodal dialog-based learning system and to realize learning functionality on all relevant system levels. Therefore, not all learning aspects are fully exploited and new challenges have been determined during the studies. It was shown that good knowledge mending ability is necessary to avoid polluting the database and knowledge models when applying dialog-based learning in a realistic system over longer period of time. In this thesis it could be shown that already by simple questions about which person is known, the knowledge base quality can be improved significantly. Further ideas have come up to extend the presented task for example by allowing a user to correct the name of a person if he thinks that there might be for example simply a mispronunciation. To further improve information that has been collected person-specific, one might want to re-evaluate information that has been collected if the session's label is corrected and assigned to a different person ID. Another idea is to use probabilistic labeling of the models, as one can estimate the likelihood that the dialog updates data correctly, and use the associated probabilities during the clustering approach, but also to influence the way that recognition models, e.g. face identification are trained.

In contrast to approaches such as active learning or data cleansing in databases, the presented approach is not intended to allow a human to 'hand-label' some kind of data set. A learning system such as studied here has internal knowledge about which it can communicate, but which does not allow direct access by a human annotator. Such a system could in the future be applied to a humanoid robot to proactively acquire information about its environment, and as it could be shown here, also has the ability to extend this learning process over a longer period of time and correct induced errors.

Bibliography

Allen, J., Ferguson, G., Miller, B. W., Ringger, E. K., and Zollo, T. S. (2000). Dialogue systems: From theory to practice in TRAINS-96. *R. Dale, H. Moisl, and H. Somers, eds.: Handbook of Natural Language Processing*, pages 347–376. *(12)*

Aoyama, K. and Shimomura, H. (2005). Real world speech interaction with a humanoid robot on a layered robot behavior control architecture. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 3814–3819. *(13)*

Asfour, T., Regenstein, K., Azad, P., Schröder, J., Bierbaum, A., Vahrenkamp, N., and Dillmann, R. (2006). Armar-III: An integrated humanoid platform for sensory-motor control. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, Genova, Italy. *(121)*

Azad, P., Asfour, T., and Dillmann, R. (2007). Stereo-based 6d object localization for grasping with humanoid robot systems. In *Proceedings International Conference on Intelligent Robots and Systems (IROS)*, San Diego, USA. *(20, 122)*

Becher, R., Steinhaus, P., Zöllner, R., and Dillmann, R. (2006). Design and implementation of an interactive object modelling system. In *Proceedings of ISR 2006 and Robotik 2006*, number 1956 in VDI-Berichte, pages 22–27, Düsseldorf. VDI-Verlag. *(20, 24, 29)*

Beringer, N., Kartal, U., Louka, K., Schiel, F., and Türk, U. (2002). Promise - a procedure for multimodal interactive system evaluation. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation'. Las Palmas, Gran Canaria, Spain.* *(17)*

Bernsen, N. O. and Dybkjaer, L. (1997). *Designing Interactive Speech Systems: From First Ideas to User Testing.* Springer-Verlag New York, Inc., Secaucus, NJ, USA. *(17)*

Bischoff, R. and Graefe, V. (2002). Dependable multimodal communication and interaction with robotic assistants. In *Proceedings of the International Workshop on Robot-Human Interactive Communication (ROMAN).* *(12)*

Bohus, D. (2007). *Error Awareness and Recovery in Task-Oriented Spoken Dialog Systems.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA. *(14)*

Bohus, D. and Rudnicky, A. I. (2003). Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Proceedings of Eurospeech*, pages 597–600. *(11, 12)*

Bohus, D. and Rudnicky, A. I. (2008). The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, In Press:–. *(11)*

Bos, J., Klein, E., and Oka, T. (2003). Meaningful conversation with a mobile robot. In *Proceedings of the EACL*. *(13)*

Bredenfeld, A., Jacoff, A., Noda, I., and Takahashi, Y., editors (2006). *RoboCup 2005: Robot Soccer World Cup IX*. Springer, Lecture Notes in Computer Science. *(14)*

Burghart, C., Holzapfel, H., Häußling, R., and Breuer, S. (2007). Coding interaction patterns between human and receptionist robot. In *Proceedings of Humanoids*, Pittsburgh, PA, USA. *(145)*

Burghart, C., Mikut, R., Holzapfel, H., and Häußling, R. (2008). Modulare subjektive und objektive Bewertung der Mensch-Roboter-Interaktion. *Zeitschrift "Künstliche Intelligenz". Themenschwerpunkt: Humanoide Roboter*, 4:16–21. *(145)*

Carbonell, J. (1979). Towards a self-extending parser. In *Annual Meeting of the Association for Computational Linguistics*. *(20, 23)*

Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge University Press. *(101)*

Chapelle, O., Weston, J., and Scholkopf, B. (2002). Cluster kernels for semi-supervised learning. *Advances in Neural Information Processing Systems*. *(27)*

Chotimongkol, A. and Rudnicky, A. (2001). N-best speech hypotheses reordering using linear regression. In *Proceedings of Eurospeech*. *(16)*

Choueiter, G., Seneff, S., and Glass, J. (2007). New word acquisition using subword modeling. In *Proceedings of Interspeech 2007*. *(20, 26)*

Chung, G. (2001). *Towards multi-domain speech understanding with flexible and dynamic vocabulary*. PhD thesis, MIT. *(26)*

Chung, G. and Seneff, S. (2002). Integrating speech with keypad input for automatic entry of spelling and pronunciation of new words. In *Proceedings of ICSLP'02*, pages 2061–2064, Denver, CO, USA. *(27)*

Chung, G., Seneff, S., Wang, C., and Hetherington, I. (2004). A dynamic vocabulary spoken dialogue interface. In *Proceedings of ICSLP'04*, Jeju Island, Korea. *(26)*

Core, M. and Allen, J. (1997). Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA. *(40)*

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies – why and how. *Knowledge-Based Systems*, 6(4):258–266. *(18)*

Danieli, M. and Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In *Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAAI Symposium*, pages 34–39, Stanford CA. *(17)*

DARPA (2007). DARPA urban challenge - event guidelines. Defense Advanced Research Projects Agency (DARPA), Arlington. *(14)*

Dautenhahn, K. and Werry, I. (2002). A quantitative technique for analysing robot-human interactions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132 – 1138. *(15)*

Denecke, M. (2000). Object-oriented techniques in grammar and ontology specification. In *The Workshop on Multilingual Speech Communication*, pages 59–64, Kyoto, Japan. *(97, 98)*

Denecke, M. (2002). Rapid prototyping for spoken dialogue systems. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taiwan. *(12, 103)*

Drury, J. L., Scholtz, J., and Yanco, H. A. (2003). Awareness in human-robot interactions. In *IEEE International Conference on Systems, Man and Cybernetics*. *(157)*

Dusan, S. and Flanagan, J. (2002). Adaptive dialog based upon multimodal language acquisition. In *Fourth IEEE International Conference on Multimodal Interfaces (ICMI)*, Pittsburgh, PA, USA. *(20, 23)*

Dusan, S. and Flanagan, J. (2003). A system for multimodal dialogue and language acquisition. In *The 2nd Romanian Academy Conference on Speech Technology and Human-Computer Dialogue*, Romanian Academy, Bucharest, Romania. Invited. *(20, 23)*

Dybkjaer, L. and Bernsen, N. O. (2000). Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6(3-4):243–271. *(17)*

Eckert, W., Levin, E., and Pieraccini, R. (1997). User modeling for spoken dialogue system evaluation. In *Proceedings of the IEEE ASR Workshop*. *(76)*

Ekenel, H. K., Gao, H., and Stiefelhagen, R. (2007). 3-d face recognition using local appearance-based models. *IEEE Transactions on Information Forensics and Security*, 2(3):630–635. *(51)*

Ekenel, H. K. and Jin, Q. (2006). Isl person identification systems in the clear evaluations. In *CLEAR Evaluation Workshop*, Southampton, UK. *(51, 57)*

Ekenel, H. K. and Pnevmatikakis, A. (2006). Video-based face recognition evaluation in the chil project - run 1. In *Proceedings of the 7th Intl. Conf. on Automatic Face and Gesture Recognition (FG 2006)*, Southampton, UK. *(51)*

Ekenel, H. K. and Stiefelhagen, R. (2005a). A generic face representation approach for local appearance based face verification. In *Proceedings of the CVPR IEEE Workshop on FRGC Experiments*, San Diego, CA, USA. *(51)*

Ekenel, H. K. and Stiefelhagen, R. (2005b). Local appearance based face recognition using discrete cosine transform. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey. *(51)*

Ekenel, H. K. and Stiefelhagen, R. (2006). Analysis of local appearance-based face recognition on frgc 2.0 database. In *Face Recognition Grand Challenge Workshop (FRGC)*, Arlington, VA, USA. *(51)*

Ekenel, H. K. and Stiefelhagen, R. (2007). An un-awarely collected real world face database: The isl-door face database. In *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, Bielefeld, Germany. *(25)*

Fawcett, T. (2003). Roc graphs: Notes and practical considerations for data mining researchers. Technical report, HPL-2003-4, HP Laboratories. *(56)*

Fügen, C., Gieselmann, P., Holzapfel, H., and Kraft, F. (2006). Natural human robot communication. In *Human Centered Robotic Systems (HCRS)*, München, Germany. *(33)*

Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., and Westphal, M. (1997). The karlsruhe-verbmobil speech recognition engine. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany. *(38)*

Fosler-Lusier, E. and Kuo, H. J. (2001). Using semantic information for rapid development of language models within asr dialogue systems. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. *(39)*

Fraser, N. (1997). *Handbook on Standards and Resources for Spoken Language Systems*, chapter Assessment of interactive systems, page 564Ű615. Mouton de Gruyter, Berlin. Gibbon, D., Moore, R., Winski, R. (eds.). *(17, 18)*

Fraser, N. M. and Gilbert, G. . N. (1991). Simulating speech systems. *Computer Speech and Language*, 5:81–99. *(18)*

Fritz, M., Kruijff, G.-J. M., and Schiele, B. (2007). Cross-modal learning of visual categories using different levels of supervision. In *International Conference on Computer Vision Systems (ICVS)*, Bielefeld, Germany. *(22)*

Fügen, C., Holzapfel, H., and Waibel, A. (2004). Tight coupling of speech recognition and dialog management - dialog-context grammar weighting for speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. *(35, 43)*

Gavalda, M. (2000). SOUP: A parser for real-world spontaneous speech. In *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT-2000)*. *(19, 20, 37)*

Gieselmann, P. (2007). *Fehlerbehandlung in Mensch-Maschine-Dialogen*. PhD thesis, Universität Stuttgart. *(14, 33)*

Gieselmann, P. and Holzapfel, H. (2005). Multimodal context management within intelligent rooms. In *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)*, Patras, Greece. *(33, 98)*

Gieselmann, P. and Stenneken, P. (2006). How to talk to robots: Evidence from user studies on human-robot communication. In *Proceedings of the Workshop on How People Talk to Computers, Robots, and other Artificial Communication Partners*, pages 68–79, Bremen, Germany. *(123)*

Glass, J., Polifroni, J., Seneff, S., and Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: The mit experience. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 4, pages 1–4, Beijing, China. *(16, 17)*

Gorin, A., Levinson, S. E., and Sankar, A. (1994). An experiment in spoken language acquisition. *IEEE Transactions on Speech and Audio Processing*, 2(1) Part 2:224–240. *(23)*

Gorin, A., Riccardi, G., and Wright, J. (1997). How may I help you? *Speech Communication - Elsevier*, 23:113–127. *(23)*

Gorin, A. L., Abella, A., Alonso, T., Riccardi, G., and Wright, J. H. (2002). Automated natural spoken dialog. *IEEE Computer Magazine*, 35(4):51–56. *(12)*

Gorin, A. L., Levinson, S. E., Gertner, A., and Goldman, E. (1991). On adaptive acquisition of language. *Computer Speech and Language*, 5(2):101–132. *(22)*

Gorniak, P. and Roy, D. (2005). Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of Seventh International Conference on Multimodal Interfaces (ICMI)*. *(29)*

Grosse, P. (05.2009). Automatische Fehlererkennung und Duplikateliminierung auf interaktiv gelernten Wissensbasen (diplomarbeit). Master's thesis, Universität Karlsruhe (TH). *(109, 114)*

Grosse, P., Holzapfel, H., and Waibel, A. (2008). Confidence based multimodal fusion for person identification. In *Proceedings of ACM Multimedia*, Vancouver, BC, Canada. *(49, 66)*

Gu, L., Li, S., and Zhang, H.-J. (2001). Learning probabilistic distribution model for multi-view face detection. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–116 – II–122. *(24)*

Guo, Y. and Schuurmans, D. (2007). Discriminative batch mode active learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS-07)*. *(109)*

Hajdinjak, M. and Mihelič, F. (2006). The paradise evaluation framework: Issues and findings. *Computational Linguistics*, 32(2):263–272. *(17)*

Hanheide, M. and Sagerer, G. (2008). Active memory-based interaction strategies for learning-enabling behaviors. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Munich. *(20)*

Heckerman, D. (1996). A tutorial on learning with bayesian networks. Msr-tr-95-06, Microsoft Research. *(63)*

Hetherington, L. (1995). *A Characterization of the problem of new, out-of-vocabulary words in continuous speech recognition and understanding.* PhD thesis, MIT. *(26)*

Hild, H. and Waibel, A. (1995). Integrating spelling into spoken dialogue recognition. In *4th European Conference on Speech, Communication and Technology (EUROSPEECH)*, pages 1977–1980. IEEE. *(27)*

Holzapfel, H. (2005). Building multilingual spoken dialogue systems. *Archives of Control Sciences - Special Issue on Human Language Technologies as a Challenge for Computer Science and Linguistics (Part II)*, 15(4):555–566. publisher: Polish Academy of Sciences location: Gliwice, Poland guest editor: Z. Vetulani. *(33)*

Holzapfel, H. (2008). A dialogue manager for multimodal human-robot interaction and learning of a humanoid robot. *Industrial Robots Journal*, 35(6):528–535. *(33)*

Holzapfel, H., Mikut, R., Burghart, C., and Haeussling, R. (2008a). Steps to creating metrics for human-like movements and communication skills (of robots). In *HRI 2008 Workshop on Metrics for Human-Robot Interaction*, Amsterdam, Netherlands. *(145)*

Holzapfel, H., Neubig, D., and Waibel, A. (2008b). A dialogue approach to learning object descriptions and semantic categories. *Robotics and Autonomous*

Systems, Special Issue on Semantic Knowledge in Robotics, 56(11):1004–1013. *(120)*

Holzapfel, H., Nickel, K., and Stiefelhagen, R. (2004). Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI).* *(35, 95)*

Holzapfel, H., Schaaf, T., Ekenel, H. K., Schaa, C., and Waibel, A. (2007). A robot learns to know people - first contacts of a robot. *Lecture Notes in Computer Science - KI 2006: Advances in Artificial Intelligence*, 4314. Freksa, C., Kohlhase, M., Schill, K. (eds.). *(49, 120, 125)*

Holzapfel, H. and Waibel, A. (2006). A multilingual expectations model for contextual utterances in mixed-initiative spoken dialogue. In *Interspeech 2006 - ICSLP*, Pittsburgh PA, USA. *(16, 35, 38, 100)*

Holzapfel, H. and Waibel, A. (2007). Behavior models for learning and receptionist dialogs. In *Proceedings of Interspeech*, Antwerp, Belgium. *(33, 63, 67, 71)*

Holzapfel, H. and Waibel, A. (2008a). Learning and verification of names with multimodal user id in dialog. In *International Conference on Cognitive Systems*, Karlsruhe, Germany. *(71)*

Holzapfel, H. and Waibel, A. (2008b). Modeling multimodal user id in dialogue. In *Proceedings of the 2nd Speech and Language Technology Workshop (SLT)*, Goa, India. *(49)*

Hone, K. and Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3/4):287–305. *(18, 150)*

Hone, K. and Graham, R. (2001). Subjective assessment of speech-system interface usability. In *Proceedings of Eurospeech*, pages 2083–2086. *(18, 150)*

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley. *(54)*

Huang, F., Yang, J., and Waibel, A. (2000). Dialogue management for multimodal user registration. In *Proceedings of the International Conference for Speech and Language Processing (ICSLP).* *(25, 63)*

Jacoff, A., Missina, E., and Evans, J. (2002). Performance evaluation of autonomous mobile robots. *Industrial Robot: An International Journal*, 29(3):259–267. *(14)*

Jin, Q., Schultz, T., and Waibel, A. (Sept. 2007). Far-field speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2023 – 2032. *(52)*

Kahn, P., Ishiguro, H., Friedman, B., and Kanda, T. (2006). What is a human? toward psychological benchmarks in the field of human–robot interaction. In *Proc., IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* *(15)*

Kaiser, E. C. (2006). Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI)*, pages 347–356, New York, NY, USA. ACM Press. *(20, 26)*

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1-2):61–84. *(16)*

Kasper, A., Becher, R., Steinhaus, P., and Dillmann, R. (2007). Developing and analyzing intuitive modes for interactive object modeling. In *Proceedings of the Ninth International Conference on Multimodal Interfaces (ICMI)*, pages 74–81, Nagoya, Japan. Association for Computer Machinery. *(24)*

Kim, D.-H., Yoon, H.-S., Chi, S.-Y., and jo Cho, Y. (2006). Face identification for human robot interaction: Intelligent security system for multi-user working environment on pc. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, Hatfield, UK. *(25)*

Kirstein, S., Wersing, H., and Koerner, E. (2008). A biologically motivated visual memory architecture for online learning of objects. *Neural Networks*, 21(1):65–77. *(20)*

Kirstein, S., Wersing, H., and Körner, E. (2005). Online learning for object recognition with a hierarchical visual cortex model. *Lecture Notes in Computer Science on Artificial Neural Networks: Biological Inspirations Ű ICANN 2005*, 3696/2005:487–492. *(20)*

Könn, S., Holzapfel, H., Ekenel, H. K., and Waibel, A. (2007). Integrating face-id into an interactive person-id learning system. In *International Conference on Computer Vision Systems (ICVS)*, Bielefeld, Germany. *(49, 54, 55, 57)*

Krsmanovic, F., Spencer, C., Jurafsky, D., and Ng, A. (2006). Have we met? MDP based speaker id for robot dialogue. In *Proceedings of Interspeech.* *(76)*

Kruijff, G.-J. M., Zender, H., Jensfelt, P., and Christensen, H. I. (2007). Situated dialogue and spatial organization: What, where. . . and why? *International Journal of Advanced Robotic Systems*, 4(2). *(21)*

Lampe, A. and Chatila, R. (2006). Performance measure for the evaluation of mobile robot autonomy. In *Proc, IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 4057–4062. *(15)*

Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G. A., and Sagerer, G. (2003). Providing the basis for human-robot-interaction: A multimodal attention system for a mobile robot. In *in Proc. Int. Conf. on Multimodal Interfaces*, pages 28–35. ACM. *(24)*

Lemon, O. (2004). Context-sensitive speech recognition in ISU-dialogue systems: Results for the grammar-switching approach. In *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*. *(35, 39)*

Lemon, O., Bracy, A., Gruenstein, A., and Peters, S. (2001). The WITAS multimodal dialogue system I. In *Proceedings of Eurospeech*. *(12, 13)*

Lemon, O. and Liu, X. (2007). Dialogue policy learning for combinations of noise and user simulation: Transfer results. In *Proceedings of SigDial*. *(14)*

Levin, E., Pieraccini, R., and Eckert, W. (1998a). Using markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. *(13)*

Levin, E., Pieraccini, R., and Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transaction On Speech and Audio Processing*, 8(1):11 – 23. *(13, 76)*

Levin, L., Thymt'e-Gobbel, A., Lavie, A., Ries, K., and Zechner, K. (1998b). A discourse coding scheme for conversational spanish. In *Proc. of Int. Conf. on Speech and Language Processing (ICSLP 98)*. *(39)*

Li, S. (2007). *Multi-modal Interaction Management for a Robot Companion*. PhD thesis, Universität Bielefeld. *(13)*

Lömker, F. (2004). *Lernen von Objektbenennungen mit visuellen Prozessen*. Dissertation, Universität Bielefeld, Technische Fakultät. *(20)*

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, page 1150Ű1517, Corfu, Greece. *(122)*

Maletic, J. I. and Marcus, A. (2000). Data cleansing: Beyond integrity analysis. In *Conference on Information Quality (IQ2000)*. *(27, 28)*

McTear, M. F. (2002). Spoken dialogue technology: Enabling the conversational interface. *ACM Computing Surveys*, 34(1):90–169. *(11, 12)*

McTear, M. F. (2004). *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer Verlag. *(11)*

Meier, U. and Hild, H. (1997). Recognition of spoken and spelled proper names. In *Eurospeech*, Rhodes, Greece. *(27)*

Metze, F., Gieselmann, P., Holzapfel, H., Kluge, T., Rogina, I., Waibel, A., Wolfel, M., Crowley, J., Reignier, P., Vaufreydaz, D., Berard, F., Cohen, B., Coutaz, J., Rouillard, S., Arranz, V., Bertran, M., and Rodriguez, H. (2005). The "fame" interactive space. In *MLMI*, Edinburgh. *(33)*

Möller, S., Smeele, P., Boland, H., and Krebbera, J. (2007). Evaluating spoken dialogue systems according to de-facto standards - a case study. *Computer Speech & Language*, Volume 21, Issue 1:26–53. *(17, 18, 150)*

Möller, S. (2005). *Quality of Telephone-Based Spoken Dialogue Systems.* Springer, New York, NY. *(17)*

Montemerlo, M., Pineau, J., Roy, N., Thrun, S., and Verma, V. (2002). Experiences with a mobile robotic guide for the elderly. In *Proceedings of the National Conference on Artificial Intelligence (AAAI).* *(13)*

Nabe, S., Cowley, S. J., Kanda, T., Hiraki, K., Ishiguro, H., and Hagita, N. (2006). Robots as social mediators: coding for engineering. In *Proc. of the International Symposium on Robot and Human Interactive Communication (Ro-Man)*, Hatfield, UK. *(16)*

Nakano, M., Hoshino, A., Takeuchi, J., Hasegawa, Y., Torii, T., Nakadai, K., Kato, K., and Tsujino, H. (2006). A robot that can engage in both task-oriented and non-task-oriented dialogues. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids-06)*, pages 404–411, Genova, Italy. *(12, 103)*

Nardi, D. and et al. (2007). *RoboCup@Home: Rules and Regulations (Draft, Version 1.0, Revision 16).* RoboCup Federation. *(14, 21)*

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences.* *(158)*

Nickel, K. and Stiefelhagen, R. (2007). Fast audio-visual multi-person tracking for a humanoid stereo camera head. In *IEEE-RAS Intl. Conference on Humanoid Robots*, Pittsburgh, USA. *(24, 50)*

Niels Ole Bernsen, Hans Dybkjær, L. D. (1996). Principles for the design of cooperative spoken human-machine dialogue. In *Proc. ICSLP.* *(17)*

Park, A. and Glass, J. R. (2006). Unsupervised word acquisition from speech using pattern discovery. In *ICASSP 2006*, Toulouse, France. *(26)*

Pietquin, O. (2004). *A Framework for Unsupervised Learning of Dialogue Strategies, PhD thesis edited.* Presses universitaires de Louvain, SIMILAR,, Louvain-la-Neuve, Belgium. *(1)*

Pietquin, O. and Renals, S. (2002). Asr system modelling for automatic evaluation and optimization of dialogue systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, Orlando, Florida. *(13, 76)*

Polifroni, J., Hirschman, L., Seneff, S., and Zue, V. (1992). Experiments in evaluating interactive spoken language systems. In *Proceedings of the HLT workshop on Speech and Natural Language*, pages 28–33, Morristown, NJ, USA. Association for Computational Linguistics. *(17)*

Price, P., Hirschman, L., Shriberg, E., and Wade, E. (1992). Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the HLT workshop on Speech and Natural Language*, pages 34–39, Morristown, NJ, USA. Association for Computational Linguistics. *(17)*

Prommer, T., Holzapfel, H., and Waibel, A. (2006). Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human-robot interaction. In *Proceedings of Interspeech - ICSLP*, Pittsburgh PA, USA. *(33, 42, 72, 73, 76, 87)*

Putze, F. (2008). Social user model acquisition through network analysis and interactive learning (diploma thesis). Master's thesis, Universität Karlsruhe. *(157, 160, 166)*

Putze, F. and Holzapfel, H. (2008). Islenquirer: Social user model acquisition through network analysis and interactive learning. In *Proceedings of the 2nd Speech and Language Technology Workshop (SLT)*, Goa, India. *(157)*

Roy, D. (1999). *Learning from Sights and Sounds: A Computational Model*. PhD thesis, MIT. *(20, 22, 23)*

Roy, D. (2005). Grounding words in perception and action: Insights from computational models. *Trends in Cognitive Science*, 9(8):389–96. *(22)*

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146. *(22)*

Roy, D. K. (2003). Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209. *(20, 22, 23, 133)*

Roy, N., Pineau, J., and Thrun, S. (2000). Spoken dialog management for robots. In *Association for Computational Linguistics (ACL)*, Hong Kong. *(12)*

Sakaue, F., Kobayashi, M., Migita, T., Shakunaga, T., and Satake, J. (2006). A real-life test of face recognition system for dialogue interface robot in ubiquitous environments. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. *(25)*

Schaaf, T. (2001). Detection of oov words using generalized word models and a semantic class language model. In *Proceedings of Eurospeech.* *(26, 27, 124)*

Schaaf, T. (2004). *Erkennen und Lernen neuer Wörter.* PhD thesis, Universität Karlsruhe (TH). *(19, 20, 26, 27, 38)*

Scharenborg, O. and Seneff, S. (2005). Two-pass strategy for handling oovs in a large vocabulary recognition task. In *Proceedings of Interspeech'05*, pages 1669–1672. *(20, 26)*

Schatzmann, J., Georgila, K., and Young, S. (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal. *(17)*

Schatzmann, J., Thomson, B., and Young, S. (2007). Statistical user simulation with a hidden agenda. In *SigDial.* *(13)*

Schatzmann, J., Weilhammer, K., Stuttle, M. N., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review, Cambridge University Press*, 21(2):97–126. *(13)*

Scheffler, K. and Young, S. (2002). Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of the Human Language Technology Workshop (HLT).* *(12, 13)*

Schneiderman, H. and Kanade, T. (2000). A statistical model for 3D object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE. *(24)*

Schultz, T., Stüker, S., Soltau, H., Metze, F., and Fügen, C. (2003). Efficient handling of multilingual language models. In *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU).* *(44)*

Shibata, T., Kato, N., and Kurohashi, S. (2007). Automatic object model acquisition and object recognition by integrating linguistic and visual information. In *Proceedings of the 15th ACM International Conference on Multimedia (ACM Multimedia 2007)*, Augsburg, Germany. *(22)*

Simpson, A. and Fraser, N. M. (1993). Black box and glass box evaluation of the sundial system. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93), Berlin*, volume 2, pages 1423–1426. *(17)*

Singh, S., Kearns, M., Litman, D., and Walker, M. (1999). Reinforcement learning for spoken dialogue systems. In *Proceedings of the Conference on Neural Information Processing Systems.* *(13)*

Skantze, G. (2007a). *Error Handling in Spoken Dialogue Systems.* PhD thesis, KTH, Stockholm, Sweden. *(14)*

Skantze, G. (2007b). Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds. In *Proceedings of SigDial*, pages 206–210, Antwerp, Belgium. *(17)*

Slobada, T. and Waibel, A. (1996). Dictionary learning for spontaneous speech recognition. In *Proceedings of ICSLP.* *(26)*

Solsona, R. A., Fosler-Lussier, E., Kuo, H.-K. J., Potamianos, A., and Zitouni, I. (2002). Adaptive language models for spoken dialogue systems. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02).* *(39)*

Soltau, H., Metze, F., Fuegen, C., and Waibel, A. (2001). A one pass- decoder based on polymorphic linguistic context assignment. In *Proceedings of ASRU'01*, Madonna di Campiglio, Trento, Italy. *(38)*

Spexard, T., Li, S., Wrede, B., Hanheide, M., Topp, E. A., and Hiittenrauch, H. (2007). Interaction awareness for joint environment exploration. In *Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication.* *(21)*

Stallard, D. (2000). Talk'n'travel: a conversational system for air travel planning. In *Proceedings of the sixth conference on Applied natural language processing*, pages 68–75, Morristown, NJ, USA. Association for Computational Linguistics. *(12)*

Steels, L. and Baillie, J.-C. (2003). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems*, 43(2-3):163–173. *(22)*

Steels, L. and Kaplan, F. (2001). Aibo's first words. the social learning of language and meaning. *Special issue of Evolution of Communication: The Evolution of Grounded Communication*, 4(1):3–32. *(22)*

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proc., ACM SIGCHI/SIGART Human-Robot Interaction*, pages 33–40. ACM Press New York, NY, USA. *(15)*

Stent, A., Dowding, J., Gawron, J. M., Bratt, E. O., and Moore, R. (1999). The commandtalk spoken dialogue system. In *Proceedings of the 37th Annual Meeting of ACL.* *(35)*

Stiefelhagen, R., Ekenel, H. K., Fügen, C., Gieselmann, P., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., and Waibel, A. (2007). Enabling multimodal human - robot interaction for the karlsruhe humanoid robot. *IEEE Transactions on Robotics*, 23 Issue 5. *(33)*

Suhm, B. (2003). Towards best practices for speech user interface design. In *Proceedings of Eurospeech*, pages 2217– 2220. *(17)*

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA. *(73)*

Terrill L. Frantz, K. M. C. (2006). Communication networks from the enron email corpus. In *SCTPLS Conference*, Baltimore, USA. *(158)*

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.-E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A., and Mahoney, P. (2006). Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, 23(9):661–692. *(14)*

Topp, E. A., Christensen, H. I., and Eklundh, K. S. (2006). Acquiring a shared environment representation. In *HRI'06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 361–362, New York, NY, USA. ACM. *(21)*

Toptsis, I., Li, S., Wrede, B., and Fink, G. (2004). A multi-modal dialog system for a mobile robot. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP).* *(12)*

Traum, D. (1999). Speech acts for dialogue agents. *Foundations of Rational Agency*, pages 169–201. *(39)*

Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., and Poesio, M. (1999). A model of dialogue moves and information state revision. Technical report, Trindi Report D2.1. *(13)*

Traum, D. R. and Hinkelman, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599. *(39)*

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Science*, pages 71–86. *(51)*

Turunen, M. and Hakulinen, J. (2003). Jaspis2 - an architecture for supporting distributed spoken dialogues. In *Proceedings of Eurospeech.* *(12, 103)*

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518. *(24)*

Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3-4):363–377. *(17)*

Walker, M. A., Litman, D. J., Kamm, A. A., and Abella, A. (1998). Evaluating spoken dialogue agents with PARADISE: two case studies. *Computer Speech and Language*, 12:317–347. *(17)*

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (ACL)*, pages 271–280, Morristown, NJ, USA. Association for Computational Linguistics. *(15, 17)*

Walker, M. A. and Shannon, A. T. T. (2000). An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:2000. *(13)*

Walters, M. L., Dautenhahn, K., Woods, S. N., and Koay, K. L. (2007). Robotic etiquette: Results from user studies involving a fetch and carry task. In *Proc. of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–324, New York, NY, USA. ACM. *(15)*

Wasserman, S. and Faust, K. (1997). *Social Network Analysis.* Cambridge University Press. *(160)*

Wersing, H., Kirstein, S., Götting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., and Körner, E. (2006). A biologically motivated system for unconstrained online learning of visual objects. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, volume 2, pages 508–517. *(20)*

Williams, J. and Young, S. (2003). Using wizard-of-oz simulations to bootstrap reinforcement-learning-based dialog management systems. In *Proceedings of the 4th SigDial Workshop on Discourse and Dialogue.* *(13)*

Woods, S., Walters, M., Koay, K., and Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *Proc. of the International Workshop on Advanced Motion Control*, Istanbul, Turkey. *(15)*

Wrede, B., Kleinehagenbrock, M., and Fritsch, J. (2006). Towards an integrated robotic system for interactive learning in a social context. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1817–1823. *(20)*

Wu, B. and Nevatia, R. (2007). Improving part based object detection by unsupervised, online boosting. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR '07*, pages 1–8, Minneapolis, MN, USA. *(21)*

Xu, W. and Rudnicky, A. (2000a). Language modeling for dialog system. In *Proc. of the Int. Conf. of Speech and Signal Processing (ICSLP'00).* *(39)*

Xu, W. and Rudnicky, A. I. (2000b). Task-based dialog management using an agenda. In *Proceedings of the ANLP/NAACL Workshop on Conversational Systems.* *(11)*

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, vol. 14, no. 4. *(28)*

Yao, L., Tang, J., and Li, J. (2007). A unified approach to researcher profiling. In *Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence.* *(158)*

Young, S., Schatzmann, J., Weilhammer, K., and Ye, H. (2007). The hidden information state approach to dialog management. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–149–IV–152. *(12)*

Young, S. R. (1993). Learning new words from spontaneous speech: A project summary. In *CMU Tech Report, CMU-CS-93-223.* *(26)*

Ziesemer, S. (2007). Namenserkennung bekannter und unbekannter namen. Master's thesis, Universität Karlsruhe (TH). *(141, 177)*

Zipf, G. K. (1965). *Human behavior and the principle of least effort : an introduction to human ecology.* Hafner, New York [u.a.], facs. of 1949 ed. edition. *(177)*

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. J., and Hetherington, L. (2000). Jupiter: A telephonebased conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):100–112. *(12)*