in: Proc. 55th Int. Scientific Colloquium, Ilmenau, Germany, 2010, pp. 635-640, Verlag ISLE 2010

ASPECTS OF USER SPECIFIC DIALOG ADAPTATION FOR AN AUTONOMOUS ROBOT

Steffen Müller, Christof Schröter, Horst-Michael Gross

Ilmenau, University of Technology
Neuroinformatics and Cognitive Robotics Lab

ABSTRACT

The paper is giving a survey on multimodal dialog technology and highlight some specifics of human machine dialog on an autonomous companion robot especially for elderly, cognitively impaired people, to be developed in the CompanionAble [1] project. The central aspect is adaptation to the user and multi-modality of inputs and outputs, which is essential for a natural and intuitive interaction. The paper at first introduces a prototypical dialog system and figures out issues of possible user adaptation. Then, a new concept for a dialog system, realizing three aspects of adaptation is described. Learning the meaning of user inputs, adaptation of dialog strategy according to the user's experience and user specific timing of proactive behaviors like reminders

Index Terms— multi-modal human machine dialog, input interpretation, fusion, timing, daytime management

1. INTRODUCTION

Within the CompanionAble consortium we are developing a mobile service robot for elderly people with mild cognitive impairments (MCI), which aims to assist them in their daily life. Besides the embodied interaction robot, the system also consists of a smart home environment able to recognize the user's situation and interact to her via a touch screen and verbally. Caused by the target group of elderly, special emphasise has been put on a natural communication allowing an intuitive interaction. Multimodal design of such an interface is obvious but comes along with some hard problems, which are not solved sufficently yet.

The most useful medium for commanding and interaction with a robot, which most of the time is located at some distance to the user, is speech. Speech is an intuitive medium with a great explanatory power but on the other hand, it is difficult to recognize speech from a distant microphon in a noisy environment. Also the interpretation of natural language is a hard problem, which often is done by means of grammars, as we do, too. Because of the mentioned problems using speech recognition, the communication with our system always includes further channels which are a GUI on a touch screen – a very robust and reliable input channel –, as well as a simple gesture recognition (only head gestures). Furthermore, the presence and activity of the user is recognized and modeled by the system. This allows a multimodal grounding in the dialog.

Multimodality as later detailed helps reducing misunderstandings and besides speeds up dialogs (up to 10%), while it is preferred by 95%-100% of interviewed users as Oviatt [2] could show.

A further aspect of a system like the robot developed in CompanionAble is an adaptation to the user's specifics. Since our system will be in contact to a person over a long period of time, adaptation of the user to the system can be expected and the user's attitude to the system has to be tracked and modulated in a positive way. The robot should not be seen as a tool convincing by its functional benefit, but it shall be a real companion having some personal relationship to the care recipient.

In the remainder of the paper, a brief overview on the state of the art in multimodal dialog management and a survey on user adaptivity is given, followed by an overview on our robot's realization and adaptation mechanisms for dialog content selection, input fusion and timing of proactive behaviors.

2. MULTIMODAL DIALOG MANAGEMENT

Multimodality means that multiple channels are used for communication between the dialog partners. Trung in his detailed survey [3] distinguishes *active* and *passive* input modes, depending on whether it is used consciously like speech or unconsciously like head gestures and facial expressions.

Various analytical models exist for describing the multimodal interaction, where a *dialog model* and a *dialog management model* have to be distinguished. Dialog models are psychological and sociological or linguists descriptions of human human or human computer interactions with a rather technical focus. On the other hand, the implementation and the methods for realizing an interactive system are following some

This work is supported by EU-FP7-ICT Grant #216487 to CompanionAble.

dialog management model, where they fall into the folowing classes: *finite-state, frame-based, information state and plan-based* and *collaborative agend-based* [3]. These are not mutually exclusive, since often mixed forms appear.

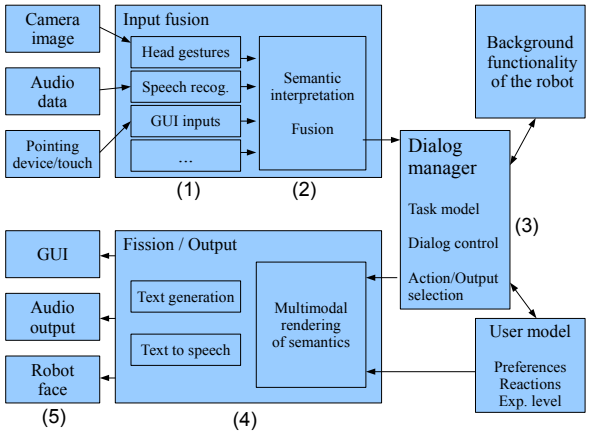

Fig. 1. Structure of a multimodal dialog system

A prototypical multimodal dialog system consist of mainly five parts (compare fig. 1). Various input modules (1) are extracting features from the different modalitis. These features are integrated by a fusion module (2) in order to extract some common semantics from the stream of features, which is sent to the dialog manager.

The dialog manager (DM) (3) is responsible for planning and executing the next actions of the system. It can take into account different models of knowledge, like some form of a task model, general and domain specific knowledge, a dialog model, a user model and the dialog state itself. Depending on the dialog management approach, the task model and the dialog model are realized in an explicit or implicit manner. E.g. in a frame-based approach the task model is explicit, while the dialog model is contained implicitly.

In order to fulfill the current tasks, the DM also needs access to the background functionality of the system, which can be a data base or the functions of a robot. Once the DM has selected the content for the message to the user, the fission module (4) is responsible for selection of the best modalities to express the content for the user. At last, output modalities (5) will render and transmit the message to the user.

In contrast to unimodal GUI interaction as well as to pure language based interaction, a multimodal dialog system comes along with the inherent need of continuous interpretation of inputs and parallel processing of the different modalities. While unimodal inputs realize as a sequence of atomic events, in a multimodal system various channels have to be analyzed in parallel. This mostly is done in a probabilistic manner. Furthermore, the different inputs have to be associated to indivual communicative acts based on the time line and the development of the dialog state. Time sensitivity is crucial in order to decide whether to interpret commands in parallel (e.g. a command and a pointing gesture) or sequential. There exist formal models [4] (CASE [5] and CARE) describing the combination of inputs. The CASE model e.g. cathegorizes the modalities into *sequential* or *parallel* and the fusion into *independant* or *combined*. Thus it finds four different ways of modality combination. Only when the modalities are independent (e.g. voice command for "come here" and doing respective hand gestures meanwhile) and used in parallel a synergy effect occures helping to reduce ambiguity and gain robustness. In contrast in the combined case (e.g. command "go there" and pointing pose giving the target) the correct interpretation of the semantic is dependent on two systems, thus true positive rate of both modalities multiplies, reducing the overall robustness. Cohen et al. [6] already in 1997 introduced a tool called quickset for multimodal dialog design, their modality integration scheme considers temporal statistical and semantical properties of the modalities.

Fusion of multimodal inputs as central element of multimodality is subdivided into *Data-level, Feature-level* and *Decision-level fusion* [4], where Data-level fusion refers to combination of different audio channels or camera images before features are extracted. Feature-level fusion combines different features before the semantic classification takes place, but it only helps when applied for closely coupled modalities, which are syncronized very well. An example for this is the combination of speech and mouth movements prior to of speech recognition. Often probabilistic models like HMMs or temporal neural networks provided with a combined feature vector are used for that kind of fusion coming along with difficulties due to complex models, which depend on extensive datasets. Despite expensive training, these methods often do not generalize well for less closely coupled modalities as we intend to use. The alternative is fusion at decision level. Here the different input channels each have extracted their own interpretation of the semantics in the user's expression, which are combined in the dialog manager.

Depending on the dialog management model, decision-level fusion by Dumas [4] are classified as *frame-based, unification-based or symbolic/statistical fusion*. The latter group also allows adaptation to a user's charakteristics as we will discuss later. The interpretation of combined inputs, which only can be interpreted in conjunction explicitely can be realized by multimodal associative maps. Here for each combination of two modal classes the multimodal semantic class is stored. This mapping can be handcrafted or learned from a labeled corpus. In our case, we apply a kind of frame-based dialog management model, the combination is done in the dialog manager, thus our research on fusion only captures the beneficial case of

parallel combined inputs.

[7] gives an overview of the development of fusion techniques for multimodal dialoges and concludes that classical fusion in contrast to machine learning based approaches is well understood nowadays. [8] noticed that machine learning based fusion at decision level on the other hand is still in its infancy and needs further research. One example for application of Bayesian inference applied to decision level fusion is presented in [9]. A problem for machine learning approaches in general is the amount of necessary training datasets, not to be underrated.

The other important part where multimodality plays a special role is the output or fission module. Once the dialog manager has decided what content to send to the user, the Fission module is responsible for selecting the way to express the message. This consists of selecting the modalitiy and if necessary generating a natural language expression. Different types of approaches for output generation can be found in literature. *template-based generation*, *conventional NLG* (Natural Language Generation) and recently *trainable generation*. Template-based methods are often used but come along with limited abilities for adaptation to user preferences and dialog context. The conventional NLG [10] consists of three steps, which are a content selection and discours planner, a sentence planner and a renderer for transforming the plans into natural language. The results of that process are satisfying, but adatation is limited by the designer of the system, which requires very domain specific tuning. For face to face dialog, such systems are often too complex and slow. The trainable NGL refines the stages of conventional NGL by means of statistical models gained from train data. Here a couple of possible solutions is generated with a conventional approach, which later are ranked by means of a learned language model. The reinforcement based methods for strategy selection in the dialog manager are also extended to include the way of information presentation in the domain of information retrieval [11].

3. THE COMPANIONABLE ROBOT'S DIALOG SYSTEM

The robot companion to be developed in our project has no manipulators an thus concentrates on hands-off services like communication, entertainment, information, observation, agenda management, and as a special aspect of CompanionAble, the cognitive stimulation of people with MCI. This shows that it intends not to be a passive tool but has a proactive component, actively approaching the user in order to encourage some activities and deliver reminders and other social interactions.

Considering the complexity and effort for developing plan and information state based implementations, we concentrate on a frame-based approach, which models the communicative acts as described below. Finite State approaches are not sufficient for the robot, since we have to handle mixed initiative dialoges. Furthermore, the poor quality of speech recognition in a noisy indoor environment makes generic dialog systems (plan based / collaborative agent-based approaches) seem to be inadequate for the domain of the robot applications.

The development of the dialog management model was inspired from Speech Act Theory [12], the turn taking model of a sequence of communication acts resulting from analysis of human human interaction. We have a high level dialog situation which incoorporates the requirements arising from a mobile robot. This level is describing the state of the conversation like *the user is absent*, *repetition*, *non matching inputs*, *help necessary or dialog ok*. At the task level topics of conversation are described by frames, e.g. navigation commands, reminder, greeting, information request, ..., which can also be hierarchically dependent.

The frames consist of a couple of communicative acts, more precisely user acts and machine acts (see fig. 2) which describe the semantic classes of possible communication. In order to allow disambiguation of inputs, the history of the dialog is represented implicitly by means of expectation values for the communicative acts. Whenever a question is asked, the possible input acts with the answers are assigned some expectation according to the graph (as depicted by the arrows in fig. 2). So a simple "yes" e.g. can be handled by the most expected input act, even if it would also match others. The data communicated in the dialog is stored in slots, having a value and a reliability which results from the inference process as described below. The reliability also indicates whether an input has been confirmed already or not. The output generation in our system is controlled by a handcraftet grammar handling each machine output act. Grammar rules can be activated or suppressed by means of rules, which will consider the reliability of slots to decide whether to confirm assumptions implicitely or explicitely. Furthermore, different possible personalities are realizable due to the activation of other grammar parts, as used for adaptation to user preferences. In order to take into account the variability of the dialogs, considerable effort has to be put into the specification of the robot's expressions. This leads to the central topic of adaptivity.

4. ADAPTIVITY

With the new fields for robot applications, a need for adaptivity comes along. One reason is the growing variety of potential users. Formerly application of IT concentrate on experienced technically versed users, while nowadays a wide public community is addressed. Additionally, the growing complexity of system functionality, the amount of information and objects of in-

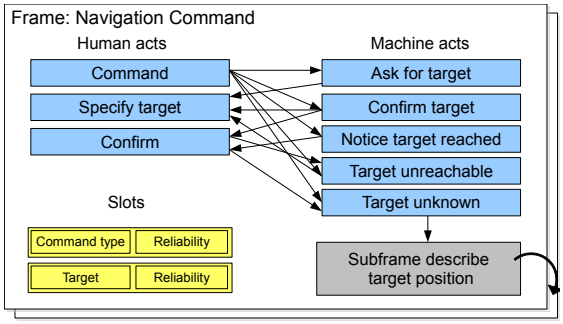

Fig. 2. Frame in the companionable dialog manager. A sequence of human input acts and machine acts is modelled. Necessary information on the dialog state are stored in slots with value and reliability also indicating the confirmation level (used for grounding)

terest to be dealt within a service application has grown compared to a few years ago, making a user specific selection and filtering necessary. In [13] a survey of adaptivity functions for adaptive user interfaces in general is given, where a useful subset is:

- taking over parts of routine tasks,
- adapting the interface to fit better the user's way of working with it (smart menus),
- giving advice on the system usage,
- controlling the dialog,
- supporting information aquisition, and prevent from information overload by means of filtering and finding information

In presence of all the adaptivity capabilities, a small set of fundamental design aspects has to be considered carefully: *Predictability, Controllability, Unobtrusiveness, Privacy*

For example, a proactive presentation of information like reminders, which falls into the last category, needs to take care of the unobtrusivness. This can be done by means of a good situation analysis and a strategy for timing active behaviors as described later on.

For designing the interface, one has to distinguish between *adaptability* and *adaptivity*. Behaviour of the system, which is setup manually by the user consciously, should be called adaptability. When the system itself finds the best settings for behavior parameters according to the user's interactions and rewards adaptivity is present. Since adaptability is indeed a serious field but is less challenging methodically, we now concentrate only on adaptivity. In the field of spoken dialog systems, mostly used for phone services, a variety of reinforcement learning approaches has been presented in the last years. Thus offline optimization of dialog strategies is a well studied topic. [14] learned from a WoZ study the optimal strategy in a markow decision process (MDP). Unfortunately, such approaches are limited to quite low complexity of dialog problems (only 4 slots in their example). For information retrieval optimizing the dialog strategy in order to add or remove restiction for selecting from a huge set of possibilities has been solved by means of reinforcement learning and user studies. Others tried to optimize duration (number of turns) necessary for task completion in a noisy input channel by means of a POMDP (also reinforcement learning).

Spitters et al. [15] are optimizing a dialog strategy for encouraging users to take up some exercises. The system can choose between different social elements like greeting, smalltalk, apologies, jokes and relational questions, which are optimized by means of reinforcement learning based on user rewards. This kind of strategy selection for an encouraging dialog for us is also of interest in the CompanionAble project.

Besides the dialog strategy for information state models, recent research concentrates on the output generation of a dialog system. [11] show an example for learning the way, how to express messages to certain users in an information retrieval scenario.

All of these optimization approaches are applied offline in the design phase of the dialog system and are limited to a quite low complexity. Furthermore, they all have to solve the training data problem, which grows inherently with the complexity of the task. A method to mitigate that problem is user modelling. Here some model users are learned based on a limited set of real interactions and then the necessary training data for the dialog adaptation is generated in simulations.

In the distinct domain of long term interaction in a personal environment, offline approaches as used for the multi-user short term dialogs are not sufficient. Here the system has to change its behavior during operation time in order to adapt to the specific user. Also in short term dialogs some user specific behavior can be realized by means of classifying the user into stereotypes and act according to a predefined or adaptable policy. The classification can be done based on information gathered so far inplicitely or explicitely. Typical stereotypes are distinguishing gender, age, or preferences in item selections.

Within the CompanionAble project different aspects of adaptation are planned or still partially realized. In particular we intent to adapt the output presentation, the input interpretation and the timing of proactive robot behaviour.

4.1. Output and dialog content adaptation

The way the system has to talk to the user inherently depends on the level of experience with and expectations to the system. Because we intent to have a long term interaction of many weeks or month, three phases

of living together have been identified: A) getting known to each other, B) stable, C) changing behavior. In the phase A the system has to guide the user to learn about its functionality and introduce the different possibilities while observing the users reaction. Further, user's preferences are acquired explicitely or inplicitely by building histograms on the users selections in menus and other media like TV programm and news channels. In the getting known phase, the dialog strategy of the system is more passive, while thresholds for self explaining help messages are lower. Passive in that context means that the robot is more commanded by the user instead of proactively making choices by it's own.

Later in the stable phase with a raising experience level of the user, the system can switch to active offering of services if a specific context situation has been noticed. Further, selections from lists of options can be suggested by the system in order to speed up the choice. Comparable to adaptive menus the problem of reduced predictability comes along with that. On the other hand current user trials with a robot assisting elderly people showed, the users want more companion like behavior instead of task fulfilling automatism, which has been suggested to be reached by introducing some unpredictability.

The last phase of tracking the changes in the user's behavior alternates with the stable phase. Implications on that are that tracking of user's preferences and rewards is necessary over the complete time of operation online and results are to be used for a continuous reselection of options (adaptation).

4.2. Adpative multimodal input interpretation

The second aspect for adaptation is a user specific interpretation of input in order to reduce annoying confirmation correction cycles in the dialog. On the one hand the individual detectors for each modality have to adapt themselfs to the user, and on the other hand, as introduced in sec. 2, learning of parallel occuring observations helps to benifit from a synergy effect.

In our application, because of the frame-based management, combined inputs are not to be considered by the fusion module. If one frame needs inputs from different modalities, they can be aggregated in the dialog manger.

The domain of home robotics is characterized by large interindividual, cultural and situation depending variances in expression of gestures (head gestures), activities and different quality of speech command recognition. Therefore, the focus of our research lies on the online learning of user specific probabilistic mappings from modality specific detections to semantic classes for the dialog management system.

In contrast to the multimodal associative maps [4], where the cross product table of semantic classes of all input modalities is built up and filled with the semantic class for the dialog, we have the assumption, that modalities are conditional independent given the user's original intention, which is the semantic class S for the dialog. By means of that, we can built up a Bayesian Network allowing to infere the semantics S from a given set of observed input modalitiy classes, knowing that combined semantics is not noticable that way. (see fig. 3)

The resulting model is depending on a couple of probability distributions $p(S|M_i)$, where S is the set of dialog semantic classes and M_i are the sets of detectable classes from the input modalities. For learning and adapting these models we do some bootstrapping from the known semantics of touch based GUI interactions. There is a continuos cycle of first inferring the S given all the current observations and sending it to the dialog manager. Once the dialog manager confirms the correctness of the result, the probability distributions are updated by a simple Maximum A Posteriory (MAP) estimation process.

Unfortunately experiments with that method are outstanding yet.

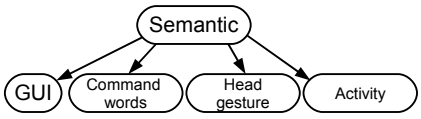

Fig. 3. Bayesian Network for inference of semantic class for a set of parallel observed independent inputs

4.3. Adaptive timing of proactive robot behavior

A third very important aspect for a personalized robot assistant is the timing of its behaviors like reminders for aganda items, but also for smalltalk related to special activities like cooking or a weather forecast service offered before the person is about to leave the home. All these services require starting the actions before the user does something or at least before he is not available anymore. An example is a reminder for an appointment, which is at 4:00 PM (see fig. 4), if the user typically is out from 3:00 PM to 6:00 PM. In this case the reminder to that appointment has to be delivered before he leaves home, which in bad conditions could not be observed directly or the user goes out earlier than usual.

The solution to that problem is a model of the user's availability over the day and for each behavior to be triggered a necessity function is defined, which charakterizes the type of the behavior. See fig. 5 for examples.

The system observes the user's presence and aktivities and builds up a statistical model of the availability. In our case a Gaussian Process is used as model mapping the daytime to a probability function

on availablity $P(A|t)$. By means of the model, the unobtrusiveness and privacy design constraints can be achieved when triggering interaction. Once a model for a day exists, the system can predict the availability and check, whether the behavior can be expected to be delivered later with a better rating. The prediction is permanently updated with new observations. E.g. if an "about to leave" activity is observed, the availability will be overlayed with an absent blob in the next hours (case (b) in the figure). For decision, the product of availability and necessity for each behavior is computed for the complete day. The point of 90% of the maximum determines the execution time. If it is in the past or currently, then the behavior will be triggered. If it is in the future the robot will wait and update the prediction meanwhile.

In a simulation of a reminder scenario for an office assistant, the method could prove to be satisfying. Tests in the CompanionAble scenario are outstanding yet.

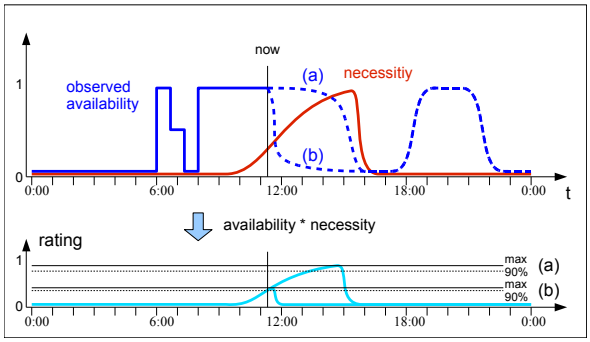

Fig. 4. Decision model for proactive behavior like delivery of reminders. (a) prediction based on a daytime model of user's presence at home, (b) overlay of currently observed activity "prepare for leaving home", lower graph: rating, product of predicted availability and need for delivery of message, For case (a) the intersection with the 90% threshold is not in the past, thus execution will be shiftet to future. In case (b) the predicted availability decreases and the 90% is reached in the past causing immediate delivery of message.

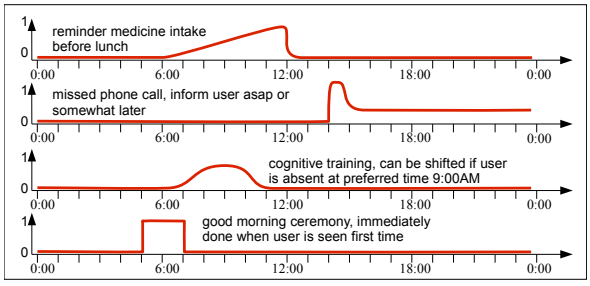

Fig. 5. Exemplary necessity curves for different behaviors: Raising ramps will shift execution to last possible point, while falling curves try to execute tasks to the first possible time.

5. CONCLUSION AND FUTURE WORK

In the present work, some essential aspects of personalized multimodal dialogs especially for companion robots have been identified and our solutions for the adaptation has been briefly sketched. In the future, these approaches will be implemented on the CompanionAble robot and tested in long term studies with users living alone in their home. For the dialog strategy selection, some further research is necessary in order to find a automatic machine learning based solution substituting handcrafted grammars as described in section 4.1.

6. REFERENCES

[1] www.companionable.net

[2] S. Oviatt, "Mulitmodal interactive maps: Designing for human performance," *Human-Computer Interaction*, vol. 12, no. 1, pp. 93–129, 1997.

[3] H. Trung, "Multimodal dialogue management-state of the art," *Human Media Interaction Department, University of Twente*, 2006.

[4] B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal Interfaces: A Survey of Principles, Models and Frameworks," *Human Machine Interaction*, pp. 3–26, 2009.

[5] L. Nigay and J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 1993, p. 178.

[6] P.R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "QuickSet: Multimodal interaction for distributed applications," in *Proceedings of the fifth ACM international conference on Multimedia*. ACM, 1997, pp. 31–40.

[7] D. Lalanne, L. Nigay, et al., "Fusion engines for multimodal input: a survey," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 153–160.

[8] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[9] S. Oviatt, R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, and L. Carmichael, "Toward a theory of organized multimodal integration patterns during human-computer interaction," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, p. 51.

[10] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[11] O. Lemon, "Adaptive natural language generation in dialogue using Reinforcement Learning," *Proceedings of SEMdial*, 2008.

[12] W.P. Alston, *Illocutionary acts and sentence meaning*, Cornell Univ Pr, 2000.

[13] A. Jameson, "Adaptive interfaces and agents," *Human-Computer Interaction: Design Issues, Solutions, and Applications*, p. 105, 2009.

[14] V. Rieser and O. Lemon, "Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation," in *Proceedings of ACL*, 2008.

[15] M. Spitters, M. de Boni, J. Zavrel, and R. Bonnema, "Learning effective and engaging strategies for advice-giving human-machine dialogue," *Natural Language Engineering*, vol. 15, no. 03, pp. 355–378, 2008.