Rapid Simulation-Driven Reinforcement Learning of Multimodal Dialog Strategies in Human-Robot Interaction

Thomas Prommer, Hartwig Holzapfel and Alex Waibel

Interactive Systems Labs
Universität Karlsruhe (TH), Germany
{tprommer|hartwig|alex}@ira.uka.de

Abstract

In this work we propose a procedure model for rapid automatic strategy learning in multimodal dialogs. Our approach is tailored for typical task-oriented human-robot dialog interactions, with no prior knowledge about the expected user and system dynamics being present. For such scenarios, we propose the use of stochastic dialog simulation for strategy learning, where the user and system error models are solely trained through the initial execution of an inexpensive Wizard-of-Oz experiment. We argue that for the addressed dialogs, already a small data corpus combined with a low-conditioned simulation model facilitates learning of strong and complex dialog strategies. To validate our overall approach, we empirically show the supremacy of the learned strategy over a hand-crafted strategy for a concrete human-robot dialog scenario. To the authors' knowledge, this work is the first to perform strategy learning from multimodal dialog simulation.

Index Terms: strategy learning, multimodal human-robot dialogs

1. Introduction

Without any doubt, the dialog strategy and its design plays a crucial role for the quality of any dialog system. Considering the presence of uncertainty about the user dynamics as well as error-prone system components, the dialog manager's decisions - e.g. which question to ask, initiative to use, or information to confirm - are multifaceted and non-trivial. While handcrafting dialog strategies becomes a tedious and non-trivial affair with an increasing complexity of the dialog system, machine learning techniques, and therein particularly reinforcement learning (RL), have become popular for automatic dialog strategy acquisition. Thereto, the dialog scenario is mapped to the formalism of a Markov Decision Process (MDP) [1]. The application of RL algorithms then promises the computation of high-quality, at best optimal dialog strategies.

The pitfall coming along with the application of RL algorithms is their necessity of huge amount of dialog experience to learn the optimal dialog strategy. A popular way to obtain such extent of experience is its artificial generation using dialog simulation (performed e.g. in [2]). So far, the majority of proposed approaches to simulation-driven strategy learning have been exclusively applicable to big-scale dialog systems as their application assume the presence of extensive online-operation experience for training the simulation model. Also all previous approaches have been isolated to speech-only, mostly telephone-based dialogs.

This work carries on the idea of dialog strategy learning to multimodal dialog interactions, hereby particularly focusing on human-robot interactions. We argue that typical human-robot dialog scenarios usually involve the settlement of a task execution by the robot and thereby usually the overall dialog complexity is restricted. Tailored for such interactions, our procedural model proposes training of a simulation model based on a small-sized dialog corpus collected by an initial execution of a Wizard-of-Oz (WOz) experiment. Combined with our propagated design of the user model, the system error models, as well as the MDP's reward function and state space, we claim that for the addressed dialogs such trained simulation model facilitates generation of artificial dialog interactions which are sufficiently accurate to allow learning of powerful strategies. Hereby, we swerve from earlier work within the domain of strategy learning mainly in three points: we (i) facilitate strategy learning already during system development (ii) propagate simulation training on a slim training set (iii) introduce multimodality to strategy learning. We validate our approach for a concrete human-robot dialog scenario by showing the supremacy of the learned strategy over a handcrafted strategy not only in the simulation environment, but more importantly also in the real world domain.

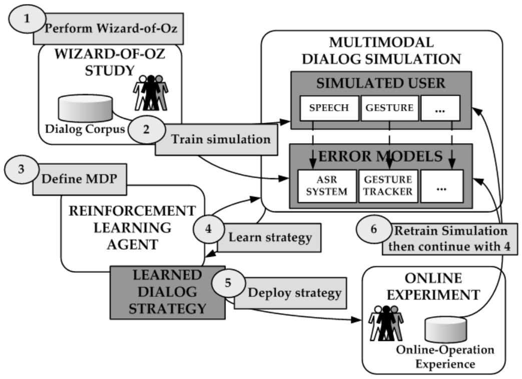

Figure 1: *Procedure model in a step by step manner*

2. Our approach

Figure 1 illustrates our proposed procedure model in a step-by-step manner. In the following, we will conceptually introduce its six different phases. Recall that our design decisions hereby are always targeted to facilitate a rapid strategy learning process, while also addressing multimodality as well as the peculiarities of typical human-robot dialog interactions.

2.1. Step 1: Perform Wizard-of-Oz experiment

We propose the execution of an WOz experiment as a first step within the strategy learning process. In WOz experiments the system's functionality is simulated to a certain extent by a hidden human wizard, whereas the test subjects are left in the belief of interacting with a finalized system. As a valuable addition, WOz experiments can also be used to obtain direct feedback from the test subjects about the system. The execution of Wizard-of-Oz (WOz) experiments within the domain of strategy learning has been proposed earlier, e.g. by Williams et al. [3] to support the MDP design. We promote the utilization of the collected WOz dialog corpus as followed: **(i)** training of the dialog simulation (user model and the system error models) **(ii)** domain-specific porting and tuning of the system's language model **(iii)** collection of user feedback for its incorporation in the MDP's reward function.

2.2. Step 2: Train dialog simulation

While in some cases there might be options to adapt task-independent data for simulation training, we exclude such possibility in this work. Instead, we consider our WOz experiment as sole data source. By using simplistic and low-conditioned statistical methods for user and error modeling, we successfully challenge data sparsity.

2.2.1. User model

For the user simulation we adapt a bi-gram model approach where the user action is solely conditioned on the last system action.

$$p = P(\mathsf{action}_{\mathsf{user}}|\mathsf{action}_{\mathsf{system}})$$

With respect to earlier proposed and more sophisticated user models (e.g. the Levin model [1]), we prefer the bi-gram user model due to its simplicity which allows us to uniformly and easily model multimodality of user actions, responses to mixed- and system-initiative actions as well as multiplicity of user actions. Note, that the quality of the bi-gram model is highly dependent on the defined abstraction granularity of simulated user actions. Within our concrete setup, we define a particular user action for the modality of speech on intentional level as the utterance of a semantic concept (SC). For the modality of gesture, a simulated action is represented by the user's execution of a 3D pointing gesture towards a particular point of interest. Generally, we assume that the user is always acting fully cooperative and goal-directed, i.e. he always only provides correct information and is motivated of fulfilling an initially fixed goal as quickly as possible.

2.2.2. ASR error model

For our applied ASR error model we adopt and extend an approach proposed by Pietquin [4] which specifies recognition probabilities for a finite set of recognition tasks. Especially for slot-based human-robot dialogs such ASR error model can be rapidly designed. Hereby, we simply define the recognition of the semantic concept of each crucial piece of information within the dialog interaction - usually represented by the information slots - as individual recognition task.

As we do not restrict our language model for particular system prompts as Pietquin, in addition to the recognition rate of a recognition task we train probabilities expressing how often confusions appear inside (In-Domain) or outside (Out-Domain) the same semantic class of the recognition task. Note that such three defined probability values not necessarily have to accumulate to one. Instead, the potential missing probability fractions model recognition errors where the utterance of a semantic concept is completely lost.

A further aspect by which we extend the ASR model proposed by Pietquin is the modeling of talker-dependent recognition performance variations. We claim that divergences e.g. in language capability, pronunciation clarity or input device adjustment (e.g. proper setup of head-mounted microphone) of different speakers cause significant deviations in recognition performance, which should not be ignored in dialog simulation. In order to model such phenomena, we measure the standard deviation of recognition rates (RR) ($\sigma_{RR(SC)}$) for each defined recognition task (RT) over all test subjects. Within the simulation, we then compute the recognition rate for each recognition task individually for every simulated dialog as follows:

$$RR_{\mathsf{SimRun}}(RT) = \mathsf{Sample}(N(0,1)) * \sigma_{RR(RT)} + RR_{\mathsf{Overall}}(RT)$$

By using the same sample value for the computation of each $RR(RT)$ within a simulation run, we create a more consistent ASR simulation: in case a high value was drawn from the normal distribution, speech recognition in the particular simulated dialog works comparable well for all recognition tasks. For a low sample value the effect is correspondingly contrary.

2.2.3. Multimodal error models

In general we propose that each single modality processing unit should be represented by an own dedicated error model within the dialog simulation framework. However, due to the different characteristics and recognition mechanisms of imaginable modalities (e.g. gaze, pointing gestures, face expressions etc.), it is impossible to provide a generic approach for error modeling fitting each modality. For our concrete experiment presented in chapter 3, we will exemplarily present an error model for the recognition of 3D pointing gestures.

2.3. Step 3: Define MDP

An adequate state space design is a crucial point for any reinforcement learning algorithm. Hereby, the exponential growth of the state space with every additional state feature requires a well-thought selection and encoding of such features. Yet, using dialog simulation facilitates the generation of a theoretically infinite amount of dialog interactions. Thereby, a relative rich state space design is not a major concern, but instead presumably facilitates learning of more complex and successful strategies. Within our procedure model, we explicitly encourage the use of an extensive state space, within the limit of a reasonable computation duration of the the learning process. Particularly interesting within the dynamic environment of human-robot dialogs, we also recognize the utility of state features to facilitate the system's intelligent adaptation to changes of the recognition conditions for different modalities within or between dialog interactions. As one example of a feature facilitating such adaptation, within our experiment a state feature is established which identifies if the recognition of pointing gestures is available or not within a particular dialog state (correct gesture recognition requires correct tracking of the user's hands and head).

A second essential key design issue within the MDP setup is the reward function. Next to the obvious evaluation criteria of success and length of the dialog, an additional optimization criterion we apply is dialog naturalness. While a system developer might be forced to allow potentially unnatural system behavior in order to avoid deadlocks, the reward function can be designed in a way to punish and avoid such behavior. However, representing the subjective criterion of naturalness within a numerical reward function is a non-trivial task. We propose the WOz experiment as an adequate scenario to collect user feedback about the naturalness of the system which subsequently can be incorporated within the reward function. Our experiment will provide an example for such an approach.

2.4. Step 4 and 5: Learn and deploy strategy

For these two steps we do not significantly swerve from other work. It should be noticed that we use the popular Watkins(λ) method for strategy learning.

2.5. Step 6: Retrain simulation

The final step six is an essential part of our approach. If further strategy improvement is desired or changes to the experiment's environmental conditions occurred, we propose the use of transcribed online-operation data to retrain the simulation model in order to subsequently learn an improved strategy from the refined simulation environment. We believe that such an approach is superior to performing online-learning out of the following reasons: **(i)** Efficient online-learning requires a small and fixed state space, while simulation-based strategy (re)learning allows alterations to the state space at any time as well as the use of a richer set of state features. **(ii)** Online-learning requires immediate knowledge if a dialog interaction ended successfully or not. Unlike in the simulation, it is non-trivial to derive this information securely online without supervision. **(iii)** In contrast to our simulation-based approach, online-learning requires a continuous exploration in the live experiment which has a negative effect on the system's online performance as well as the user's amenity.

3. Experimental setup

The exemplary human-robot dialog interaction is installed within the context of the collaborative research center SFB588 at the University of Karlsruhe, developing cooperative and multimodal humanoid robots. In our particular scenario, our robot acts in the role of an early-stage bartender. Twenty objects, diverging in color, shape and location are placed on a table in front of the robot and represent the user's order options. Before each dialog, our test subjects silently choose a particular item of interest from the table setting. Our robot then initiates a multimodal dialog with the goal to identify the user's object of interest and serve the corresponding item. The robot's dialog capabilities are mainly represented by its speech and gesture recognition system. Detailed information about the system's conversational interface can be found in [5].

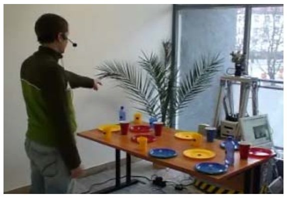

Figure 2: *Experiment setup*

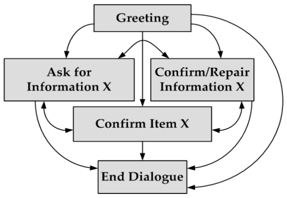

Figure 3: *Dialog automaton*

3.1. The dialog structure

Figure 3 shows the dialog automaton of our experiment. Our robot always initiates the dialog with a self-introduction and the mixed-initiative action "What can I serve you?". Afterwards the robot has the option to prompt for the three available information slots: object type, object color and object location. For the case of object location, our robot prompts explicitly for a pointing gesture of the user towards the desired item. Further options exist in confirming information slots either individually or jointly. At any point within the dialog, the robot is able to confirm/exclude the so far best matching item as desired item or respectively end the dialog and serve such item. As a general constraint we restricted the number of dialog turns to a maximum of ten.

3.2. Wizard-of-Oz experiment

Within our Wizard-of-Oz experiment 15 test subjects were engaged and a total of 82 dialogs (314 utterances) were completed. The human wizard performed the function of speech and gesture recognition, as well as carried out the dialog strategy during runtime. We interviewed all test subjects after task completion about the naturalness of the system. Many test subjects perceived the direct repetition of system actions as well as a second prompt for the same information slot (e.g. in case of recognition gaps) as particularly unnatural. We incorporated such feedback in the reward function of our reinforcement algorithm as we will introduce later.

3.3. Building the simulation

The simulation of our dialog scenario comprises three models: the user model, the ASR error model and the gesture recognition error model. Table 1 shows three exemplary entries of the user model trained on the WOz corpus using the bi-gram model introduced earlier. Table 2 shows the indicated recognition tasks and trained ASR error model parameters for close-speech communication.

Action	Type		Color		Gesture		Confirm	
Greeting	67/82	82%	51/82	62%	21/82	26%	—	
AskType	30/35	86%	2/35	6%	1/35	3%	—	
ConfColor	0/62	0%	1/62	2%	0/62	0%	62/62	100%

Table 1: *Trained user model using a bi-gram approach*

Semantic Concept	Correct Recognition		In-Domain Confusion		Out-Domain Confusion		Standard Deviation
Type	72/99	72.7%	1/99	1.0%	8/99	8.1%	0.12
Color	69/92	75.0%	2/92	2.2%	12/92	13.0%	0.18
Confirm	219/250	87.6%	4/250	1.6%	5/250	2.0%	0.12

Table 2: *Close-speech ASR error model*

For the gesture tracker error model, we configured and trained the model on pointing gestures recorded during the WOz experiment, using three parameters: the average error angle (21.1) , its standard deviation (4.54) and the recall rate (79.5%). Within the simulation process, a draw from the recall rate determines if a simulated user gesture is detected and if so, an error angle drawn from a normal distribution (based on the average error angle and its deviation) is added to the assumed perfect direction vector of the gesture.

3.4. MDP design

Within our MDP design we defined eight actions, derivable by the earlier dialog description. Furthermore, we used the following eight state features F1 to F8:

[F1-F3] Status of information slots object type, color and position (0:empty & never asked, 1:empty & already asked, 2:filled, 3:confirmed)
[F4] Number of candidates given the current slot information (0:no candidates, 1:one candidate, 2: two to four candidates, 3: more than four candidates)
[F5] Speech condition (0:close speech, 1:distant speech)
[F6] ASR history (0: bad, 1: good) - the history is set too bad if no speech input was understood for a system action or the confirmation of an information slot was declined
[F7] Gesture tracking condition (0:bad, 1:good) - condition was determined by observing the component's internal state (e.g. hands of test subject not traceable -> set condition to bad)
[F8] Last system action

The design of our reward function was intended to facilitate the optimization of the dialog strategy towards the criteria length, success and naturalness. In order to motivate short dialogs, every system action was penalized with a reward of -0.2. An unsuccessful dialog - i.e. the robot served the wrong item - was heavily punished with a reward of -5, a successful finish rewarded by +1.0. Integrating the user feedback from the WOz experiment, we additionally penalized the second prompt for an information slot as well as the direct repetition of the same system action also by a value of -0.2.

3.5. Experimental validation

In order to empirically validate the benefit of our overall approach, we compared the performance of our learned strategy with a handcrafted and rule-based baseline strategy for the simulation and real world domain. The designed baseline strategy proceeds in a way that it first collects and jointly confirms the object type and color information, before asking once for a user gesture. Thereafter, the robot explicitly tries to confirm the best matching item within the candidate collection, until the referenced item is confirmed as correct by the user. As a general rule the dialog is ended as soon as only one candidate item remains. 18 test subjects engaged within the validation experiment of which only one test subject also participated within the WOz experiment. A total of 94 dialogs (576 utterances) were collected in sequential runs of four to six dialogs of each test subject. Hereby, in order to fairly balance a potential learning effect of the user, we evenly switched between use of the two strategies.

4. Results

Table 3 shows the results of the comparison of the performance of our learned strategy to the introduced handcrafted baseline strategy. As we can see, our learned strategy performs significantly superior to the handcrafted strategy, not only in the simulation (SIM) but also in the real world (REAL) with respect to the central optimization criterion of collected reward, the task completion rate as well as dialog brevity.

As a further positive aspect, despite the small corpus of collected real dialog interactions, the strategies' performance figures resemble nicely for the simulation and real world domain and thereby indicate an adequate accuracy of our simulation model.

	Baseline Strategy		**RL Strategy**	
	REAL	SIM	REAL	SIM
Dialogs	47	10^5	47	10^5
Task completion	80,4%	83.3%	86.9%	91.3%
∅ Utterances	5.924	5.956	4.943	5.007
∅ Reward	-1.252	-1.067	-0.782	-0.688

Table 3: *Comparison baseline to learned strategy*

After the validation experiment, we also retrained our simulation framework based on the 94 collected and transcribed online-operation dialogs. We could observe that the newly learned optimal strategy changed insignificantly to the one learned from the WOz corpus: only in 2.7% of the states the computed optimal action diverged for both strategies. Such observation shows that the performed WOz experiment represented a valid data source for the simulation training, but also indicates a moderate stability of the optimal learned strategy, reachable already by a small simulation training corpus.

5. Conclusions

In this paper we introduced a holistic procedure model for rapid learning of multimodal dialog strategies within human-robot interactions. We empirically validated that despite the use of a small training corpus and low-conditioned simulation model our approach is capable of learning dialog strategies which significantly outperform typical handcrafted strategies. In contrast to earlier work, our approach represents an efficient and inexpensive proceeding to automatic dialog strategy learning during system development. Future work will have to show how well our approach can be applied to dialog scenarios with a more extensive information space as well as additional modalities other than speech and pointing gestures.

6. Acknowledgments

This work was supported by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots" as well as by an interACT and Baden-Württemberg scholarship.

7. References

[1] Levin, E., Pieraccini, R., Eckert, W. '*A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies*' Proc. IEEE Trans. on Speech and Audio Processing, 2000

[2] Scheffler, K., Young, S., '*Automatic Learning of Dialogue Strategy Using Dialogue Simulation and Reinforcement Learning*' Proc. Human Language Technology, 2002

[3] Williams, J. D., Young, S. '*Using Wizard-of-Oz Simulations to Bootstrap Reinforcement-Learning-Based Dialog Management Systems*' SIGdial Workshop, 2003

[4] Pietquin, O., Renals, S. '*ASR System Modeling for Automatic Evaluation and Optimization of Dialogue Systems*' Proc. IEEE Conf. Acoustics, Speech and Signal Proc., 2002

[5] Stiefelhagen, R. et al. '*Natural Human-Robot Interaction using Speech, Gaze and Gestures*' Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2003