Leveraging POMDPs trained with User Simulations and Rule-based Dialogue Management in a Spoken Dialogue System

Sebastian Varges, Silvia Quarteroni, Giuseppe Riccardi, Alexei V. Ivanov, Pierluigi Roberti
Department of Information Engineering and Computer Science
University of Trento
38050 Povo di Trento, Italy
{varges|silviaq|riccardi|ivanov|roberti}@disi.unitn.it

Abstract

We have developed a complete spoken dialogue framework that includes rule-based and trainable dialogue managers, speech recognition, spoken language understanding and generation modules, and a comprehensive web visualization interface.

We present a spoken dialogue system based on Reinforcement Learning that goes beyond standard rule based models and computes on-line decisions of the best dialogue moves. Bridging the gap between handcrafted (e.g. rule-based) and adaptive (e.g. based on Partially Observable Markov Decision Processes - POMDP) dialogue models, this prototype is able to learn high rewarding policies in a number of dialogue situations.

1 Reinforcement Learning in Dialogue

Machine Learning techniques, and particularly Reinforcement Learning (RL), have recently received great interest in research on dialogue management (DM) (Levin et al., 2000; Williams and Young, 2006). A major motivation for this choice is to improve robustness in the face of uncertainty due for example to speech recognition errors. A second important motivation is to improve adaptivity w.r.t. different user behaviour and application/recognition environments.

The RL approach is attractive because it offers a statistical model representing the *dynamics* of the interaction between system and user. This contrasts with the supervised learning approach where system behaviour is learnt based on a fixed corpus. However, exploration of the range of dialogue management strategies requires a simulation environment that includes a simulated user (Schatzmann et al., 2006) in order to avoid the prohibitive cost of using human subjects.

We demonstrate various parameters that influence the learnt dialogue management policy by using pre-trained policies (section 5). The application domain is a tourist information system for accommodation and events in the local area. The domain of the trained DMs is identical to that of a rule-based DM that was used by human users (section 4), allowing us to compare the two directly.

2 POMDP demonstration system

The POMDP DM implemented in this work is shown in figure 1: at each turn at time t, the incoming N user act hypotheses $a_{n,u}$ split the state space S_t to represent the complete set of interpretations from the start state (N=2). A belief update is performed resulting in a probability assigned to each state. The resulting ranked state space is used as a basis for action selection. In our current implementation, belief update is based on probabilistic user responses that include SLU confidences. Action selection to determine system action $a_{m,s}$ is based on the best state (m is a counter for actions in action set A). In each turn, the system uses an ϵ-greedy action selection strategy to decide probabilistically if to exploit the policy or explore any other action at random. (An alternative would be softmax, for example.) At the end of each dialogue/session a reward is assigned and policy entries are added or updated for each state-action pair involved. These pairs are stored in tabular form. We perform Monte Carlo updating similar to (Levin et al., 2000):

$$Q_t(s,a) = R(s,a)/n + Q_{t-1} \cdot (n-1)/n \quad (1)$$

where n is the number of sessions, R the reward and Q the estimate of the state-action value.

At the beginning of each dialogue, a user goal U_G (a set of concept-value pairs) is generated randomly and passed to a user simulator. The user simulator takes U_G and the current dialogue context to produce plausible SLU hypotheses. These

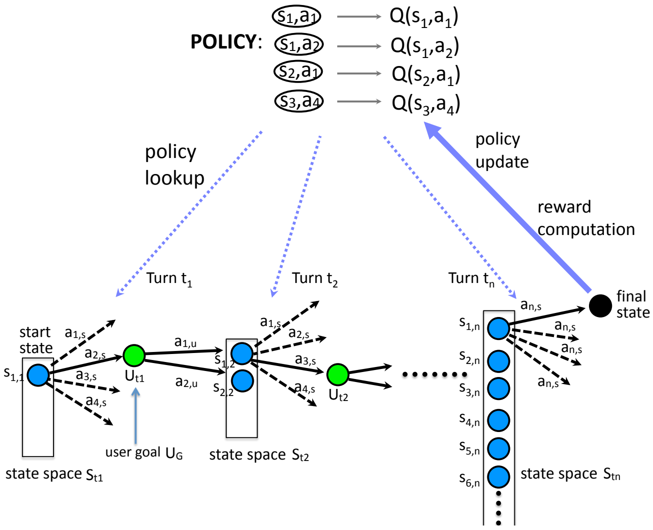


Figure 1: POMDP Dialogue Manager

are a subset of the concept-value pairs in U_G along with a confidence estimate bootstrapped from a small corpus of 74 in-domain dialogs. We assume that the user 'runs out of patience' after 15 turns and ends the call.

The system visualizes POMDP-related information live for the ongoing dialogue (figure 2). The visualization tool shows the internal representation of the dialogue manager including the the N best dialogue states after each user utterance and the reranking of the action set. At the end of each dialogue session, the reward and the policy updates are shown, i.e. new or updated state entries and action values. Moreover, the system generates a plot that relates the current dialogue's reward to the reward of previous dialogues.

3 User Simulation

To conduct thousands of simulated dialogues, the DM needs to deal with heterogeneous but plausible user input. We designed a User Simulator (US) which bootstraps likely user behaviors starting from a small corpus of 74 in-domain dialogs, acquired using a rule-based version of the system (section 4). The role of the US is to simulate the output of the SLU module to the DM during the whole interaction, fully replacing the ASR and SLU modules. This differs from other user simulation approaches where n-gram models of user dialog acts are represented.

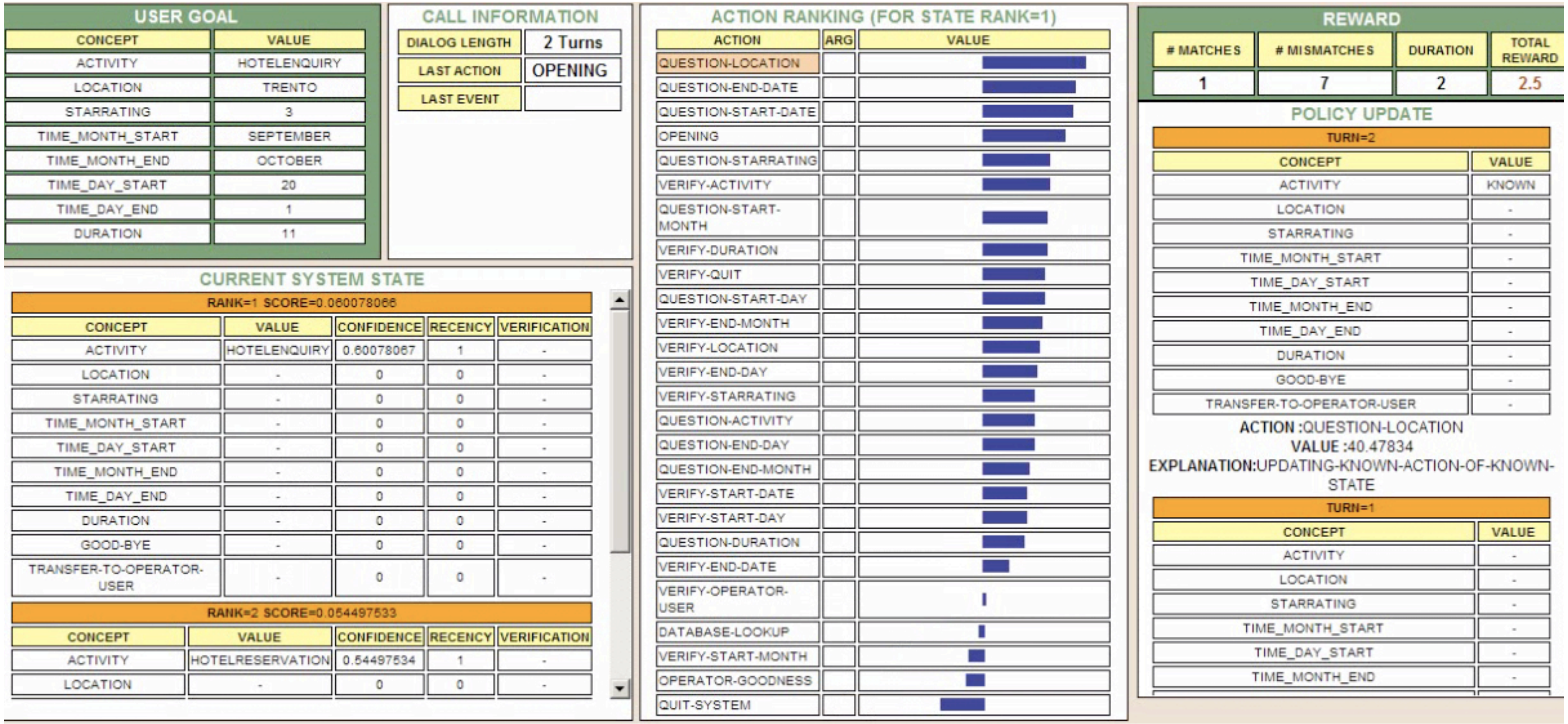
For each simulated dialogue, one or more user goals are randomly selected from a list of possible user goals stored in a database table. A goal is represented as the set of concept-value pairs defining a task. Simulation of the user's behaviour happens in two stages. First, a *user model*, i.e. a model of the user's intentions at the current stage of the dialogue, is created. This is done by mining the previous system move to obtain the concepts required by the DM and their corresponding values (if any) from the current user goal. Then, the output of the user model is passed to an *error model* that simulates the "noisy channel" recognition errors based on statistics from the dialogue corpus. Errors produce perturbations of concept values as well as phenomena such as *noInput*, *noMatch* and *hangUp*. If the latter phenomena occur, they are directly propagated to the DM; otherwise, plausible confidences (based on the dialogue corpus) are attached to concept-value pairs. The probability of a given concept-value observation at time $t+1$ given the system move at time t, $a_{s,t}$, and the session user goal g_u, called $P(o_{t+1}|a_{s,t}, g_u)$, is obtained by combining the outputs of the error model and the user model:

$$P(o_{t+1}|a_{u,t+1}) \cdot P(a_{u,t+1}|a_{s,t}, g_u)$$

where $a_{u,t+1}$ is the true user action. Finally, concept-value pairs are combined in an SLU hypothesis and, as in the regular SLU module, a cumulative utterance-level confidence is computed, determining the rank of each of the N hypotheses output to the DM.

4 Rule-based Dialogue Management

A rule-based DM was developed as a meaningful comparison to the trained DM, to obtain training data from human-system interaction for the US, and to understand the properties of the domain. Rule-based dialog management works in two stages: retrieving and preprocessing facts (tuples) taken from a dialogue state database, and inferencing over those facts to generate a system response. We distinguish between the 'context model' of the first phase – essentially allowing more recent values for a concept to override less recent ones – and the 'dialog move engine' of the second phase. In the second stage, acceptor rules match SLU results to dialogue context, for example perceived user concepts to open questions. This may result in the decision to verify the application parameter in question, and the action is verbalized by language generation rules. If the parameter is accepted, application dependent task



USER GOAL

CONCEPT	VALUE
ACTIVITY	HOTELENQUIRY
LOCATION	TRENTO
STARRATING	3
TIME_MONTH_START	SEPTEMBER
TIME_MONTH_END	OCTOBER
TIME_DAY_START	20
TIME_DAY_END	1
DURATION	11

CALL INFORMATION

DIALOG LENGTH	2 Turns
LAST ACTION	OPENING
LAST EVENT	

CURRENT SYSTEM STATE

RANK=1 SCORE=0.060078066

CONCEPT	VALUE	CONFIDENCE	RECENCY	VERIFICATION
ACTIVITY	HOTELENQUIRY	0.60078067	1	-
LOCATION	-	0	0	-
STARRATING	-	0	0	-
TIME_MONTH_START	-	0	0	-
TIME_DAY_START	-	0	0	-
TIME_MONTH_END	-	0	0	-
TIME_DAY_END	-	0	0	-
DURATION	-	0	0	-
GOOD-BYE	-	0	0	-
TRANSFER-TO-OPERATOR-USER	-	0	0	-

RANK=2 SCORE=0.054497533

CONCEPT	VALUE	CONFIDENCE	RECENCY	VERIFICATION
ACTIVITY	HOTELRESERVATION	0.54497534	1	-
LOCATION	-	0	0	-

ACTION RANKING (FOR STATE RANK=1)

ACTION	ARG	VALUE
QUESTION-LOCATION		
QUESTION-END-DATE		
QUESTION-START-DATE		
OPENING		
QUESTION-STARRATING		
VERIFY-ACTIVITY		
QUESTION-START-MONTH		
VERIFY-DURATION		
VERIFY-QUIT		
QUESTION-START-DAY		
VERIFY-END-MONTH		
VERIFY-LOCATION		
VERIFY-END-DAY		
VERIFY-STARRATING		
QUESTION-ACTIVITY		
QUESTION-END-DAY		
QUESTION-END-MONTH		
VERIFY-START-DATE		
VERIFY-START-DAY		
QUESTION-DURATION		
VERIFY-END-DATE		
VERIFY-OPERATOR-USER		
DATABASE-LOOKUP		
VERIFY-START-MONTH		
OPERATOR-GOODNESS		
QUIT-SYSTEM		

REWARD

# MATCHES	# MISMATCHES	DURATION	TOTAL REWARD
1	7	2	2.5

POLICY UPDATE

TURN=2

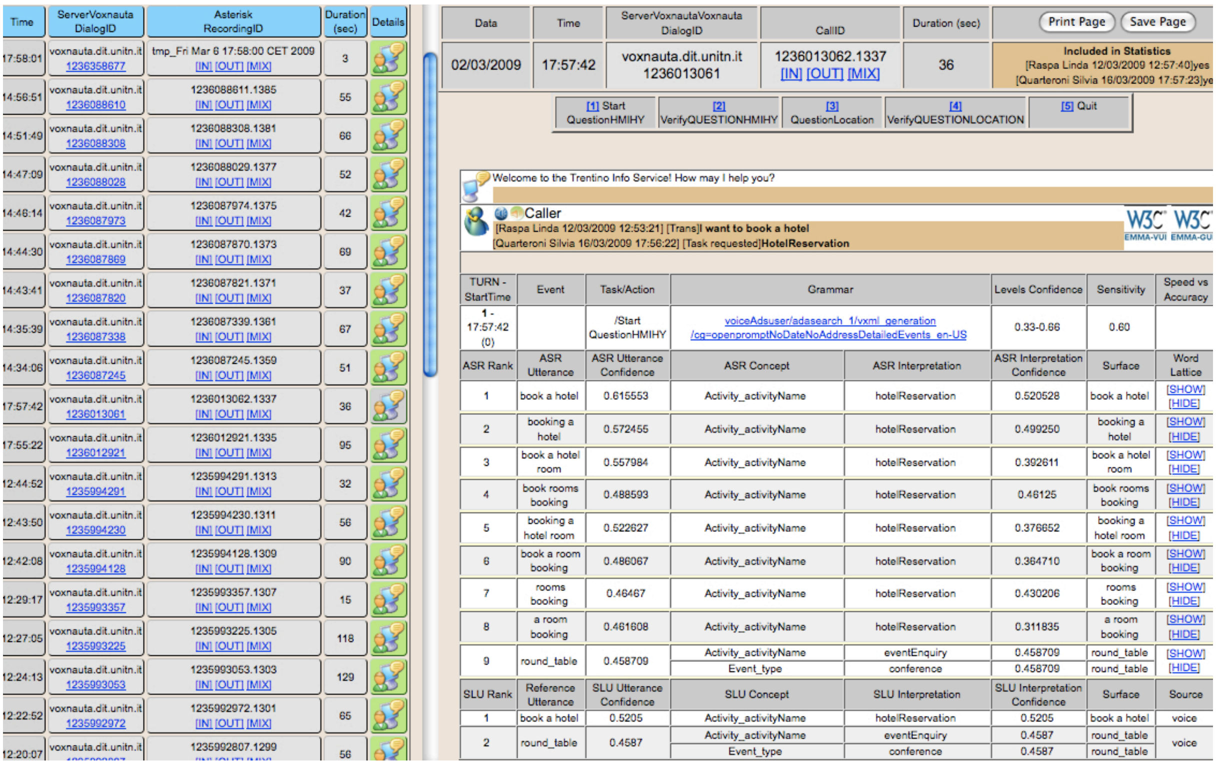
CONCEPT	VALUE
ACTIVITY	KNOWN
LOCATION	-
STARRATING	-
TIME_MONTH_START	-
TIME_DAY_START	-
TIME_MONTH_END	-
TIME_DAY_END	-
DURATION	-
GOOD-BYE	-
TRANSFER-TO-OPERATOR-USER	-

ACTION :QUESTION-LOCATION
VALUE :40.47834
EXPLANATION:UPDATING-KNOWN-ACTION-OF-KNOWN-STATE

TURN=1

CONCEPT	VALUE
ACTIVITY	-
LOCATION	-
STARRATING	-
TIME_MONTH_START	-
TIME_DAY_START	-
TIME_MONTH_END	-

Figure 2: A screenshot of the online visualization tool. Left: user goal (top), evolving ranked state space (bottom). Center: per state action distribution at turn t_i. Right: consequent reward computation (top) and policy updates (bottom). See video at `http://www.youtube.com/watch?v=69QR0tKKhCw`.



Time	ServerVoxnauta DialogID	Asterisk RecordingID	Duration (sec)	Details
17:58:01	voxnauta.dit.unitn.it 1236358677	tmp_Fri Mar 6 17:58:00 CET 2009 [IN] [OUT] [MIX]	3	
14:56:51	voxnauta.dit.unitn.it 1236088610	1236088611.1385 [IN] [OUT] [MIX]	55	
14:51:49	voxnauta.dit.unitn.it 1236088308	1236088308.1381 [IN] [OUT] [MIX]	66	
14:47:09	voxnauta.dit.unitn.it 1236088028	1236088029.1377 [IN] [OUT] [MIX]	52	
14:46:14	voxnauta.dit.unitn.it 1236087973	1236087974.1375 [IN] [OUT] [MIX]	42	
14:44:30	voxnauta.dit.unitn.it 1236087869	1236087870.1373 [IN] [OUT] [MIX]	69	
14:43:41	voxnauta.dit.unitn.it 1236087820	1236087821.1371 [IN] [OUT] [MIX]	37	
14:35:39	voxnauta.dit.unitn.it 1236087338	1236087339.1361 [IN] [OUT] [MIX]	67	
14:34:06	voxnauta.dit.unitn.it 1236087245	1236087245.1359 [IN] [OUT] [MIX]	51	
17:57:42	voxnauta.dit.unitn.it 1236013061	1236013062.1337 [IN] [OUT] [MIX]	36	
17:55:22	voxnauta.dit.unitn.it 1236012921	1236012921.1335 [IN] [OUT] [MIX]	95	
12:44:52	voxnauta.dit.unitn.it 1235994291	1235994291.1313 [IN] [OUT] [MIX]	32	
12:43:50	voxnauta.dit.unitn.it 1235994230	1235994230.1311 [IN] [OUT] [MIX]	56	
12:42:08	voxnauta.dit.unitn.it 1235994128	1235994128.1309 [IN] [OUT] [MIX]	90	
12:29:17	voxnauta.dit.unitn.it 1235993357	1235993357.1307 [IN] [OUT] [MIX]	15	
12:27:05	voxnauta.dit.unitn.it 1235993225	1235993225.1305 [IN] [OUT] [MIX]	118	
12:24:13	voxnauta.dit.unitn.it 1235993053	1235993053.1303 [IN] [OUT] [MIX]	129	
12:22:52	voxnauta.dit.unitn.it 1235992972	1235992972.1301 [IN] [OUT] [MIX]	65	
12:20:07	voxnauta.dit.unitn.it	1235992807.1299	56	

Data	Time	ServerVoxnautaVoxnauta DialogID	CallID	Duration (sec)	Print Page Save Page
02/03/2009	17:57:42	voxnauta.dit.unitn.it 1236013061	1236013062.1337 [IN] [OUT] [MIX]	36	**Included in Statistics** [Raspa Linda 12/03/2009 12:57:40]yes [Quarteroni Silvia 16/03/2009 17:57:23]ye

[1] Start QuestionHMIHY	[2] VerifyQUESTIONHMIHY	[3] QuestionLocation	[4] VerifyQUESTIONLOCATION	[5] Quit

Welcome to the Trentino Info Service! How may I help you?

Caller

[Raspa Linda 12/03/2009 12:53:21] [Trans]**I want to book a hotel**

[Quarteroni Silvia 16/03/2009 17:56:22] [Task requested]**HotelReservation**

TURN - StartTime	Event	Task/Action	Grammar		Levels Confidence	Sensitivity	Speed vs Accuracy
1 - 17:57:42 (0)		/Start QuestionHMIHY	voiceAdsuser/adasearch_1/vxml_generation /cg=openpromptNoDateNoAddressDetailedEvents_en-US		0.33-0.66	0.60	
ASR Rank	ASR Utterance	ASR Utterance Confidence	ASR Concept	ASR Interpretation	ASR Interpretation Confidence	Surface	Word Lattice
1	book a hotel	0.615553	Activity_activityName	hotelReservation	0.520528	book a hotel	[SHOW] [HIDE]
2	booking a hotel	0.572455	Activity_activityName	hotelReservation	0.499250	booking a hotel	[SHOW] [HIDE]
3	book a hotel room	0.557984	Activity_activityName	hotelReservation	0.392611	book a hotel room	[SHOW] [HIDE]
4	book rooms booking	0.488593	Activity_activityName	hotelReservation	0.46125	book rooms booking	[SHOW] [HIDE]
5	booking a hotel room	0.522627	Activity_activityName	hotelReservation	0.376652	booking a hotel room	[SHOW] [HIDE]
6	book a room booking	0.486067	Activity_activityName	hotelReservation	0.364710	book a room booking	[SHOW] [HIDE]
7	rooms booking	0.46467	Activity_activityName	hotelReservation	0.430206	rooms booking	[SHOW] [HIDE]
8	a room booking	0.461608	Activity_activityName	hotelReservation	0.311835	a room booking	[SHOW] [HIDE]
9	round_table	0.458709	Activity_activityName Event_type	eventEnquiry conference	0.458709 0.458709	round_table round_table	[SHOW] [HIDE]
SLU Rank	Reference Utterance	SLU Utterance Confidence	SLU Concept	SLU Interpretation	SLU Interpretation Confidence	Surface	Source
1	book a hotel	0.5205	Activity_activityName	hotelReservation	0.5205	book a hotel	voice
2	round_table	0.4587	Activity_activityName Event_type	eventEnquiry conference	0.4587 0.4587	round_table round_table	voice

Figure 3: Left Pane: overview of a selection of dialogues in our visualization tool. Right Pane: visualization of a system opening prompt followed by the user's activity request. All *distinct* SLU hypotheses (concept-value combinations) deriving from ASR are ranked based on concept-level confidence (2 in this turn).

rules determine the next parameter to be acquired, resulting in the generation of an appropriate request. See (Varges et al., 2008) for more details.

5 Visualization Tool

In addition to the POMDP-related visualization tool (figure 2), we developed another web-based dialogue tool for both rule-based and POMDP system that displays ongoing and past dialogue utterances, semantic interpretation confidences and distributions of confidences for incoming user acts (see dialogue logs in figure 3).

Users are able to talk with several systems (via SIP phone connection to the dialogue system server) and see their dialogues in the visualization tool. They are able to compare the rule-based system, a randomly exploring learner that has not been trained yet, and several systems that use various pre-trained policies. The web tool is available at `http://cicerone.dit.unitn.it/DialogStatistics/`.

Acknowledgments

This work was partially supported by the European Commission Marie Curie Excellence Grant for the ADAMACH project (contract No. 022593) and by LUNA STREP project (contract No. 33549).

References

E. Levin, R. Pieraccini, and W. Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1).

J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowledge Engineering Review*, 21(2):97–126.

Sebastian Varges, Giuseppe Riccardi, and Silvia Quarteroni. 2008. Persistent information state in a data-centric architecture. In *Proc. 9th SIGdial Workhop on Discourse and Dialogue*, Columbus, Ohio.

J. D. Williams and S. Young. 2006. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):393–422.