Machine Learning for Interactive Systems: Challenges and Future Trends

Olivier Pietquin, Manuel Lopes

Machine Learning for Interactive Systems : Challenges and Future Trends

Olivier Pietquin[1] Manuel Lopes[2]

[1] Université Lille 1 - LIFL (UMR 8022 CNRS/Lille 1) - France
[2] Inria Bordeaux Sud-Ouest - France

olivier.pietquin@univ-lille1.fr - manuel.lopes@inria.fr

Abstract

Machine learning has been introduced more than 40 years ago in interactive systems through speech recognition or computer vision. Since that, machine learning gained in interest in the scientific community involved in human-machine interaction and raised in the abstraction scale. It moved from fundamental signal processing to language understanding and generation, emotion and mood recognition and even dialogue management or robotics control. So far, existing machine learning techniques have often been considered as a solution to some problems raised by interactive systems. Yet, interaction is also the source of new challenges for machine learning and offers new interesting practical but also theoretical problems to solve. In this paper, we address these challenges and describe why research in machine learning and interactive systems should converge in the future.

Keywords

Machine learning, interactive systems, interaction management

1 Introduction

Communication between humans involves complex signals such as speech, gestures, facial expressions, body movements, written texts, *etc.* These signals convey high-level information such as semantics, emotions, context but can take highly variable forms. For instance, there might be many acoustic realizations for one word sequence, many word sequences for one meaning, *etc.* To enable machines to interact with humans in a natural manner, this variability has to be handled. On another hand, machine learning is the branch of artificial intelligence that addresses the problem of learning intelligent behaviours from data. To deal with communicative signal variability, machine learning has naturally been introduced very early in Human-Machine Interaction (HCI). The first and probably major achievement of machine learning in HCI is the introduction of Hidden Markov Models (HMM) in Automatic Speech Recognition (ASR) in the mid 70's [31, 30] which remains the standard method for ASR today. At the same time, data driven methods for text-to-speech synthesis (TTS) were developed [60]. It is only much later that machine learning has been exploited as a mean to interpret higher-level information such as semantics [63] or facial expression [51]. As for ASR and TTS, high-level analysis methods also gave rise to new synthesis methods like data-driven language generation [76].

In this paper, we are interested in machine learning methods intervening at a higher level : interaction management. Indeed, building an interactive system is not only about putting together all these input and output processing modules. There is a need for a intermediate module for sequencing the interaction. Taking past inputs and outputs into account, the interaction manager is in charge of deciding what should be the next system output. The interaction manager is probably one of the latest components of an interactive system that benefited from machine learning techniques. In order to make the interaction more natural, in a measurable way it is necessary to decide what outputs to make at each step of the interaction. For this, in the late 90's, spoken dialogue management has been cast into a sequential decision making problem [43] that could be solved by machine learning methods such as reinforcement learning [84]. This seminal work led to many other applications of reinforcement learning to Spoken Dialogue Systems (SDS) [81, 68, 42] but also to other types of interacting systems such as tutoring applications [29, 66, 10], museum guides [87], car driving assistance [73], recommender systems [24] and even robotics bar tenders [20].

In the following, we address the different challenges arising when taking the sequential nature of interaction into account. We first describe how interaction can be seen as a sequential decision making problem in Section 2. We then explain why and how this decision making problem has been extended to handle partial observability in Section 3. After 15 years of research in this area, these methods have proven to be efficient in finding good interaction strategies but not to be efficient in terms of data. Data sparsity thus remains a problem addressed in Section 4. Thanks to improvement in data efficiency, there has been a lot of work to enable systems to learn online, from interactions. In Section 5, paradigms to improve efficiency by actively learn new skills will be presented. Recently, going even further

active learning a new trend of research emerged : imitation learning. This will be explained in Section 6. From this, we will see that interaction provides now totally new problems to machine learning and we will summarize these in Section 7 before coming to our conclusion.

2 Interaction as a sequential decision making process

Interaction management is the problem of deciding on what to do in a given context, knowing that this context will be influenced by the decision. It is thus a sequential decision making problem where present decisions influence future ones and the success of the interaction. To optimize this process, planning algorithms [21] were first proposed. Yet, planning makes a lot of assumptions such as being able to enumerate all the possible contexts or knowing transition probabilities between states given actions. Also, the objective has to be known in advance so that the optimal path in the graph can be computed. Such approach is not robust to model uncertainty and does not have a proper solution in realistic stochastic scenarios. Once the plan is computed, it can hardly be modified even though the interaction goes wrong.

The machine learning answer to the sequential decision making optimisation problem is Reinforcement Learning [84]. Although model-based approach have been studied for a long time [2], it's only in the 90's that is has been applied to real world problems where there is no knowledge about the model, and so the system as to simultaneously optimize and estimate the model. In this paradigm, an agent (*e.g.* interactive systems) faces a dynamic system (*e.g.* humans) that steps from state to state as an effect of the actions of the agent. The agent has to learn which is the sequence of actions that makes the system go through desired states. To assess the quality of a state, the agent receives rewards after each action it performs in the environment. It thus tries to follow a path in the state space that offers the best cumulative reward. If one assumes that human-machine interaction is a turn-taking process (which is a strong assumption that is more and more contested in incremental systems [82]), then interaction management becomes such a sequential decision making problem.

Using reinforcement learning requires casting the task into the Markov Decision Processes (MDP) paradigm [2]. An MDP is formally a t-uple $\{S, A, R, T, \gamma\}$ where S is the state space, A is the action space, $R : S \rightarrow \mathbb{R}$ is the reward function, $T : S \times A \rightarrow \mathcal{P}(S)$ is a set of Markovian transition probabilities and γ is a discount factor to be defined later. The optimisation of the decision making problem consists in finding a policy $\pi : S \rightarrow \mathcal{P}(A)$ that maps states to actions in such a way that the cumulative reward obtained by following this policy is maximized. To do so, the quality of a policy is measured in every state as the expected cumulative reward that can be obtained by following the policy starting from that state. This measure is called the value function $V^\pi : S \rightarrow \mathbb{R}$:

$$V^\pi(s) = E\left[\sum_{i=0}^{\infty} \gamma^i R(s_i) | s_0 = s, a_i = \pi(s_i)\right] \quad (1)$$

One can define an order on value functions such as $V^{\pi_1} > V^{\pi_2}$ if $\forall s \; V^{\pi_1}(s) > V^{\pi_2}(s)$. The optimal policy π^* is the one that maximizes the value function for every state : $\pi^* = \arg\max_\pi V^\pi$. Many algorithms have been proposed in the literature to attempt at solving this problem [84], especially when the transition probabilities are not known, and this is still an active research area.

We can now cast human-machine interaction management as an MDP (first proposed in the late 90's [44]). The state space is the set of all possible interaction contexts and actions are the communicative acts the system can perform. The transition probabilities are usually unknown and several definitions for the reward function can be found in the literature. It is generally argued that the user satisfaction should be used as a reward [83] which can be approximated as a linear combination of objective measures that can be gathered during the interaction [88]. Yet, this reward is most often a very simple handcrafted function [44, 68, 89]. To define such a reward is very task-dependent. If the system is devoted to goal-oriented dialogues, social chat, emotion control etc. it of course has to be different.

3 Partial Observability and non-Markovian processes

The MDP framework makes several strong assumptions. For instance, the dialogue contexts cannot be perfectly observed due to the recognition error introduced by the speech and the semantic analysers. The task is therefore non-Markov in the observation space. To meet the Markov assumption made by the MDP framework, the underlying states have to be inferred from observations using what is called a belief tracker. For example, the *Hidden Information State* [90] paradigm builds a list of the most probable current situations given the past observations, which is supposed to be a Markovian representation allowing for taking decisions in the MDP framework.

To take into account the perceptual aliasing problem introduced by error-prone speech and language understanding modules, Partially Observable MDP (POMDP) have been proposed to model the dialogue management task [77] and the tutoring task [75]. Yet, solving the POMDP problem requires the transition and observation models to be known which also requires a lot of assumptions and engineering work. There has been a lot of work to make this approach tractable and suitable for learning online making this approach very promising [89, 12].

There has been some attempts to either learn a Markov state representation online [13] or to learn a policy without making the Markov assumption [14].

4 Data sparsity

The data required to create complete and accurate models of interactive systems is often impossible to obtain. To alleviate this problem, interaction simulation based on user modeling [80, 64, 45] together with error modeling (ASR *etc.*) [72, 67, 86] is most often used to artificially expand training datasets. However, the learnt strategies are sensible to the quality of the user model which is very difficult to assess [79, 71].

An alternative to this bootstrapping method is to use generalization frameworks adapted to RL such as approximate dynamic programming. Although this idea was first proposed very early [3] it took a long time before it has been studied in the field of reinforcement learning [25, 41, 84]. Because it was very new in machine learning at the time RL was first introduced in interactive systems, very few attempts to apply generalization methods in the framework of interaction management can be found in the literature. In [28], the authors use the SARSA(λ) algorithm [84] with linear function approximation which is known to be sample inefficient. In [46], LSPI [41] is used with feature selection and linear function approximation. Recently, Fitted Value Iteration (FVI) [25] has also been applied to dialogue management [6, 70]. All these studies report *batch* learning of dialog policies from fixed sets of data and thus learn in an *off-policy* manner, meaning that they learn an optimal policy from observations generated with another policy (which is mandatory for learning from fixed sets of data). It also means that, once a strategy is learnt from these datasets, it doesn't evolve anymore while, of course, one cannot expect to have a representative enough dataset for complex tasks.

5 Online and active learning

To alleviate the problem of incompleteness and inconsistency of data collected offline, online learning of interaction management strategies has recently been made possible. These systems optimize the policy while interacting with a user. This requires permanently changing the policy to be learnt, and a trade-off must be made between trying new actions to learn their effects (exploration) and use actions whose effects are already known (exploitation).

Examples are Gaussian Processes [22], Natural Actor Critic [34] or Kalman Temporal Differences [69]. The two former [22, 34] report the use of *online* and *on-policy* algorithms that change the policy frequently. These changes to the policy made during learning are visible to the user which may cause problems in real applications at the early stage of learning where the changes in the policy can lead to very bad behaviors. Thus, user simulation is still required. The later [69] makes possible *online* and *off-policy* learning which means that the system can learn online by observing a non-optimal policy in action (e.g. an handcrafted safe but suboptimal strategy). To make online learning safer (to avoid the online learner to take very bad action), active learning has been proposed [11]. This method estimates the uncertainty about the outcomes of actions and decides to explore the most uncertain but promising actions. This approach has shown to perform very efficiently online in simulation [12] and in real world [23]. Similar approaches could be made even under a POMDP framework [17].

6 Learning from Demonstrations

Being the optimization of the behavior of an interactive system such a hard problem, it has been suggested to learn such behaviors from humans. Indeed, it is not a strong assumption to say that humans are experts in interaction that should be used as model for machines. Here again, several approaches can be envisioned.

Many criticisms have been done to the reinforcement learning approach to interaction management [61, 62]. Especially, one criticism that has not been much addressed, is that these algorithms require providing the learning agent with a reward after each interaction. Although there have been attempts to define objective reward functions such as the PARADISE framework [88], this reward is indeed generally handcrafted by the system designer who introduces some expertise in the system [44, 68, 89] but also a strong bias. Very little attention has been paid to the particular problem of defining the best reward function for interactive systems.

A formal approach that tries to learn a reward function from human behavior is Inverse Reinforcement Learning (IRL) [78, 56]. It is of major importance in human-machine interaction where naturalness of the interaction is a desired feature. Indeed, since quantifying naturalness and user satisfaction is tricky, imitating the behavior of human operators can be a solution as suggested in [62]. This solution has been used to model human behaviours [4, 65] or for learning the reward of a dialogue system [9].

Nevertheless, IRL is not without problems. It is an ill-posed problem since the zero-reward is a solution whatever the expert policy (in other words, if you receive a zero-reward whatever you do, every policy is optimal). Also, most algorithms suppose that the direct RL problem can be solved as many time as needed or that any number of random samples of interactions can be generated [1, 55, 35]. Yet, this is not true since solving the direct RL problem or gathering random data requires interacting with humans with whom the system cannot be random. New paradigms that do not make these assumptions have been recently proposed [36, 74, 53] and applied to Embodied Conversional Agents applications [58]

Defining the appropriate reward function that will lead to a desired behavior is actually a real problem and sometimes it is easier to demonstrate examples of optimal behaviors. Giving driving lessons is such a task where demonstrating a good behavior is easier than associating a reward to each couple of contexts and actions. Interaction management is also such a task since it is very natural for human beings to interact with each other although it is much harder to iso-

late contexts and associate a reward to each possible action in these contexts. Humans can thus help machines to learn policies during an interactive process [16, 48, 15]. Under this approach the human user is considered as a teacher that interacts with the machine and provides extra feedback. Approaches have considered extra reinforcement signals [85], action requests [27, 50], disambiguation among actions [8], preferences among states [52], iterations between practice and user feedback sessions [33, 39] and choosing actions that maximize the user feedback [37, 38], expert judgement [18].

No matter what formalism being used, while learning from humans, it is possible to rely on active learning approaches that ensure that the data provided by the human is the most relevant. Such approaches can be applied when learning a reward function [50, 32] or a policy directly [8, 53]. Another reason to learn from humans is that when the users train the system they might become more comfortable with using it and accept it. See the work from [59] for a study on this subject. The queries of the machine will have the dual goal of allowing it to deal with its own limitations and give the user information about the its uncertainty on the task being learned [19, 7].

7 New challenges for machine learning

As shown before, interactive systems have many properties that require innovative machine learning techniques such as the sequential nature of interaction, the partial observability of inputs or the non-deterministic behaviour of users. Although these fields are still under active research (like imitation learning), there are many other big challenges brought by interactive systems to machine learning that will undoubtedly generate fundamental research in this field.

A first one is to clearly understand the theoretical properties of such systems. Machine learning has became a very theoretical field with time which can create a big gap between the interests of different communities. But on another hand, using machine learning in human-computer interaction requires theoretical proofs since empirical ones are hard to obtain. For instance, guarantees about security are often required before using robots in an inhabited area. Having a human on the loop we have to consider the risks involved by a decision or the cost in terms of tiredness of making many queries in an interactive learning setting. Estimating risks in a sequential decision making process is a real machine learning challenge [54]. Studies and algorithms have also addressed the problem of deciding when to ask. Most approaches will just ask to user whenever the information is needed [57] or when there is high uncertainty [8]. A more advanced situation considers making queries only when it is too risky to try experiments [17].

Another challenge is to take into account the fact that human users may also change their behaviour with time. Its not only that this makes the environment of the machine-learning agent non-stationary but adversarial. Indeed, the users adapt their behaviour to the one of the machine which itself learns from the observations they make from the human behaviour. This *co-adaptation* phenomenon is very poorly addressed in the HCI literature [5] (although it is also known in brain-computer interaction [40]) but it is also not common in the machine learning community because it brings very tricky problems to solve [47].

A related challenging aspects of this co-adaptation is when parts of interactions are not understood by one of the agents. Here the machine must be able to learn the meaning of such unknown symbols [26, 49].

These are only few examples of unsolved challenges, but there are many others such as scalability, weakly supervised learning, transfer learning, cold start and so on.

8 Conclusion

In this paper, we described a list of challenges induced by interactive systems that were addressed by means of machine learning. Especially, we were interested in the problem of managing interactions which is intrinsically sequential. Although interactive systems were at the origin of major signal processing and machine learning achievements initially (as for HMMs), they became consumers of machine learning techniques in the last decades in the field of sequential decision making. It is now again a source of big challenges for the machine learning community and, especially, it offers a panel of killing applications that has the potential to increase the visibility of machine learning. For these reasons, we believe that links between communities will be tighter than ever in the near future.

Références

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of ICML 2004*, page 1, 2004.

[2] R. Bellman. *Dynamic Programming*. Dover Publications, sixth edition, 1957.

[3] R. Bellman and S. Dreyfus. Functional approximation and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13 :247–251, 1959.

[4] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, et al. User simulation in dialogue systems using inverse reinforcement learning. *Proceedings of Interspeech 2011*, pages 1025–1028, 2011.

[5] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, M.-I. Supelec, and F. Metz. Co-adaptation in spoken dialogue systems. In *Proceedings of the IWSDS 2012*, Ermenonville, France, 2012.

[6] S. Chandramohan, M. Geist, and O. Pietquin. Optimizing Spoken Dialogue Management with Fitted Value Iteration. In *Proceedings of Interspeech'10*, Makuhari (Japan), 2010.

[7] C. Chao, M. Cakmak, and A. Thomaz. Transparent active learning for robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE Inter. Conf. on*, pages 317–324, 2010.

[8] S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *J. Artificial Intelligence Research*, 34 :1–25, 2009.

[9] H. R. Chinaei and B. Chaib-draa. Learning dialogue pomdp models from data. In *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, Canadian AI'11, pages 86–91, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] B. Clement, P.-Y. Oudeyer, D. Roy, and M. Lopes. Online optimization of teaching sequences with multi-armed bandits. *International Conference on Educational Data Mining*, 2014.

[11] L. Daubigney, M. Gasic, S. Chandramohan, M. Geist, O. Pietquin, and S. Young. Uncertainty management for on-line optimisation of a POMDP-based large-scale spoken dialogue system. In *Proceedings of Interspeech 2011*, page 1301âĂŞ1304, Florence (Italy), August 2011.

[12] L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin. A Comprehensive Reinforcement Learning Framework for Dialogue Management Optimisation. *IEEE Journal of Selected Topics in Signal Processing*, 6(8) :891–902, December 2012. pdf.

[13] L. Daubigney, M. Geist, and O. Pietquin. Model-free POMDP optimisation of tutoring systems with echo-state networks. In *Proceedings of SIGDial 2013*, pages 102–106, Metz (France), August 2013.

[14] L. Daubigney, M. Geist, and O. Pietquin. Particle Swarm Optimisation of Spoken Dialogue System Strategies. In *Proceedings of Interspeech 2013*, Lyon (France), August 2013.

[15] R. Dillmann. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems*, 47(2) :109–116, 2004.

[16] R. Dillmann, O. Rogalla, M. Ehrenmann, R. Zollner, and M. Bordegoni. Learning robot behaviour and skills based on human demonstration and advice : the machine learning paradigm. In *Inter. Symposium on Robotics Research (ISRR)*, volume 9, pages 229–238, 2000.

[17] F. Doshi, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement : using bayes risk for active learning in pomdps. In *25th Inter. Conf. on Machine learning (ICML'08)*, pages 256–263, 2008.

[18] L. El Asri, R. Laroche, and O. Pietquin. Reward Shaping For Statistical Optimisation Of Dialogue Management. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP 2013)*, volume 7978 of *Lecture Notes in Computer Science*, pages 93–101, Tarragona (Spain), July 2013. Springer.

[19] T. Fong, C. Thorpe, and C. Baur. Robot, asker of questions. *Robotics and Autonomous systems*, 42(3) :235–243, 2003.

[20] M. E. Foster, S. Keizer, Z. Wang, and O. Lemon. Machine learning of social states and skills for multi-party human-robot interaction. In *Proceedings of the workshop on Machine Learning for Interactive Systems (MLIS 2012)*, page 9, Montpellier, France, 2012.

[21] R. Freedman. Atlas : A plan manager for mixed-initiative, multimodal dialogue. In *Proceedings of the AAAI-99 Workshop on Mixed-Initiative Intelligence*, pages 1–8. Citeseer, 1999.

[22] M. Gasic, F. Jurcicek, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of SIGDIAL'10*, Tokyo, Japan, 2010.

[23] M. Gašić, F. Jurčiček, B. Thomson, K. Yu, and S. Young. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *Proceedings of ASRU*, pages 312–317, 2011.

[24] N. Golovin and E. Rahm. Reinforcement learning architecture for web recommendations. In *Proceedings of the International Conference on Information Technology : Coding and Computing (ITCC 2004)*, volume 1, pages 398–402, Las Vegas, Nevada, USA, 2004.

[25] G. Gordon. Stable Function Approximation in Dynamic Programming. In *ICML'95*.

[26] J. Grizou, I. Iturrate, L. Montesano, P.-Y. Oudeyer, and M. Lopes. Calibration-free bci based control. In *AAAI'14*, 2014.

[27] D. Grollman and O. Jenkins. Dogged learning for robots. In *Robotics and Automation, 2007 IEEE Inter. Conf. on*, pages 2483–2488, 2007.

[28] J. Henderson, O. Lemon, and K. Georgila. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 2008.

[29] A. Iglesias, P. Martínez, R. Aler, and F. Fernández. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, 31(1) :89–106, 2009.

[30] F. Jelinek. *Statistical Methods for Speech Recognition*. Language, Speech and Communications Series. Mit Press, 1997.

[31] F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3) :250–256, May 1975.

[32] K. Judah, A. Fern, and T. Dietterich. Active imitation learning via reduction to iid active learning. In *UAI*, 2012.

[33] K. Judah, S. Roy, A. Fern, and T. Dietterich. Reinforcement learning via practice and critique advice. In *AAAI Conf. on Artificial Intelligence (AAAI-10)*, 2010.

[34] F. Jurcicek, B. Thomson, S. Keizer, M. Gasic, F. Mairesse, K. Yu, and S. Young. Natural Belief-Critic : a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems. In *Proceedings of Interspeech'10*, Makuhari (Japan), 2010.

[35] E. Klein, M. Geist, B. Piot, and O. Pietquin. Inverse reinforcement learning through structured classification. pages 1–9, South Lake Tahoe, Nevada, USA, 2012.

[36] E. Klein, B. Piot, M. Geist, and O. Pietquin. A cascaded supervised learning approach to inverse reinforcement learning. In H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezny, editors, *Proceedings of ECML/PKDD 2013*, volume 8188 of *Lecture Notes in Computer Science*, pages 1–16, Prague (Czech Republic), September 2013. Springer.

[37] W. Knox and P. Stone. Interactively shaping agents via human reinforcement : The tamer framework. In *fifth Inter. Conf. on Knowledge capture*, pages 9–16, 2009.

[38] W. Knox and P. Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *9th Inter. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 5–12, 2010.

[39] P. Korupolu, V.N., M. Sivamurugan, and B. Ravindran. Instructing a reinforcement learner. In *Twenty-Fifth Inter. FLAIRS Conf.*, 2012.

[40] S. Koyama, S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. Comparison of brain–computer interface decoding algorithms in open-loop and closed-loop control. *Journal of computational neuroscience*, 29(1-2) :73–87, 2010.

[41] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 2003.

[42] O. Lemon and O. Pietquin. Machine learning for spoken dialogue systems. In *Proceedings of Interspeech'07*, pages 2685–2688, Anvers, Belgium, 2007.

[43] E. Levin, R. Pieraccini, and W. Eckert. Learning dialogue strategies within the markov decision process framework. In *Proceedings of ASRU 1997*, pages 72–79. IEEE, 1997.

[44] E. Levin, R. Pieraccini, and W. Eckert. Using Markov decision process for learning dialogue strategies. In *Proceedings of ICASSP 98*, volume 1, pages 201–204, Seattle, Washington, USA, 1998.

[45] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1) :11–23, 2000.

[46] L. Li, S. Balakrishnan, and J. Williams. Reinforcement Learning for Dialog Management using Least-Squares Policy Iteration and Fast Feature Selection. In *InterSpeech'09*, Brighton (UK), 2009.

[47] M. L. Littman. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

[48] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ Inter. Conf. on*, volume 4, pages 3475–3480, 2004.

[49] M. Lopes, T. Cederborg, and P.-Y. Oudeyer. Simultaneous acquisition of task and feedback models. In *IEEE - International Conference on Development and Learning (ICDL'11)*, Frankfurt, Germany, 2011.

[50] M. Lopes, F. S. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'09)*, 2009.

[51] K. MASE. Recognition of facial expression from optical flow. *IEICE transactions*, 74(10) :3473–3483, 1991.

[52] M. Mason and M. Lopes. Robot self-initiative and personalization by learning through repeated interactions. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, 2011.

[53] F. S. Melo and M. Lopes. Learning from demonstration using mdp induced metrics. In *Machine learning and knowledge discovery in databases (ECML/PKDD'10)*, 2010.

[54] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49(2-3) :267–290, 2002.

[55] G. Neu and C. Szepesvári. Training parsers by inverse reinforcement learning. *Machine learning*, 77(2-3) :303–337, 2009.

[56] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of ICML 2000*, pages 663–670, Stanford, CA, USA, 2000.

[57] M. Nicolescu and M. Mataric. Learning and interacting in human-robot domains. *Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on*, 31(5) :419–430, 2001.

[58] R. Niewiadomski, J. Hofmann, J. Urbain, T. Platt, J. Wagner, B. Piot, H. Cakmak, S. Pammi, T. Baur, S. Dupont, M. Geist, F. Lingenfelser, G. McKeown, O. Pietquin, and W. Ruch. Laugh-aware virtual agent and its impact on user amusement . In *Proceedings of AAMAS2013*, pages 619–626, Saint Paul, USA, May 2013.

[59] T. Ogata, N. Masago, S. Sugano, and J. Tani. Interactive learning in human-robot collaboration. In *Intel-*

ligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ Inter. Conf. on, volume 1, pages 162–167, 2003.

[60] J. P. Olive and N. Spickenagel. Speech resynthesis from phoneme-related parameters. *The Journal of the Acoustical Society of America*, 59(4) :993–996, 1976.

[61] T. Paek. Reinforcement learning for spoken dialogue systems : Comparing strengths and weaknesses for practical deployment. In *Proceedings of the Interspeech Dialog-on-Dialog Workshop (2006)*, 2006.

[62] T. Paek and R. Pieraccini. Automating spoken dialogue management design using machine learning : An industry perspective. *Speech Communication*, 50(8) :716–729, 2008.

[63] R. Pieraccini, E. Levin, and E. Vidal. Learning how to understand language. In *Proceedings of Eurospeech'93*, pages 1407–1412, 1993.

[64] O. Pietquin. Consistent goal-directed user model for realisitc man-machine task-oriented spoken dialogue simulation. In *Proceedings of ICME 2006*, pages 425–428, Amsterdam, Netherlands, 2006.

[65] O. Pietquin. Inverse Reinforcement Learning for Interactive Systems. In *Proceedings of the IJCAI workshop on Machine Learning for Interactive Systems (MLIS 2013)*, pages 71–75, Beijing (China), August 2013. Invited Speaker.

[66] O. Pietquin, L. Daubigney, and M. Geist. Optimization of a tutoring system from a fixed set of data. In *Proceedings of the ISCA workshop on Speech and Language Technology in Education*, pages 1–4, Venice, Italy, 2011.

[67] O. Pietquin and T. Dutoit. Dynamic bayesian networks for nlu simulation with applications to dialog optimal strategy learning. In *Proceedings of ICASSP 2006*, volume 1, pages 49–52, Toulouse, France, 2006.

[68] O. Pietquin and T. Dutoit. A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2) :589–599, 2006.

[69] O. Pietquin, M. Geist, and S. Chandramohan. Sample Efficient On-line Learning of Optimal Dialogue Policies with Kalman Temporal Differences. In *Proceedings of IJCAI 2011*, pages 1878–1883, Barcelona, Spain, July 2011. Oral Presentation.

[70] O. Pietquin, M. Geist, S. Chandramohan, and H. Frezza-Buet. Sample-Efficient Batch Reinforcement Learning for Dialogue Management Optimization. *ACM Transactions on Speech and Language Processing*, 2011.

[71] O. Pietquin and H. Hastie. A survey on metrics for the evaluation of user simulations. *Knowledge Engineering Review*, 28(01) :59–73, February 2013. first published as FirstView.

[72] O. Pietquin and S. Renals. ASR System Modeling For Automatic Evaluation And Optimization of Dialogue Systems. In *Proceedings of ICASSP 2002*, volume I, pages 45–48, Orlando, (USA, FL), May 2002.

[73] O. Pietquin, F. Tango, and R. Aras. Batch reinforcement learning for optimizing longitudinal driving assistance strategies. In *Proceedings of the IEEE Symposium on Computational intelligence in vehicles and transportation systems (CIVTS 2011)*, pages 73–79, Paris, France, 2011.

[74] B. Piot, M. Geist, and O. Pietquin. Boosted and reward-regularized classification for apprenticeship learning. In *Proceedings of AAMAS2014*, Paris (France), May 2014.

[75] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by pomdp planning. In *Artificial intelligence in education*, pages 280–287. Springer, 2011.

[76] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[77] N. Roy, J. Pineau, and S. Thrun. Spoken dialogue management using probabilistic reasoning. In *Proceedings of ACL 2000*, pages 93–100. Association for Computational Linguistics, 2000.

[78] S. Russell. Learning agents for uncertain environments. In *Proceedings of COLT 1998*, pages 101–103, Madison, Wisconsin, USA, 1998.

[79] J. Schatzmann, M. N. Stuttle, K. Weilhammer, and S. Young. Effects of the user model on simulation-based learning of dialogue strategies. In *Proceedings of ASRU 2005*, December 2005.

[80] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2) :97–126, June 2006.

[81] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of HTL 2002*, pages 12–19, San Diego, Californie, USA, 2002. Morgan Kaufmann Publishers Inc.

[82] D. Schlangen and G. Skantze. A general, abstract model of incremental dialogue processing. In *Proceedings of EACL 2009*, pages 710–718, 2009.

[83] S. Singh, M. Kearns, D. Litman, and M. Walker. Reinforcement learning for spoken dialogue systems. In *Proceedings of NIPS99*, 1999.

[84] R. Sutton and A. Barto. *Reinforcement Learning : An Introduction*. Cambridge Univ Press, 1998.

[85] A. Thomaz and C. Breazeal. Teachable robots : Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7) :716–737, 2008.

[86] B. Thomson, M. Gasic, M. Henderson, P. Tsiakoulis, and S. Young. N-best error simulation for training spoken dialogue systems. In *Proceedings of SLT 2012)*, pages 37–42. IEEE, 2012.

[87] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, et al. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research*, 19(11) :972–999, 2000.

[88] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE : A framework for evaluating spoken dialogue agents. In *Proceedings of EACL 1997*, pages 271–280. Association for Computational Linguistics, 1997.

[89] J. D. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2) :393–422, 2007.

[90] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model : A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2) :150–174, 2010.