Inverse Reinforcement Learning for Interactive Systems*

[Extended Abstract]

Olivier Pietquin
SUPELEC - UMI 2958 (GeorgiaTech-CNRS)
2 rue Edouard Belin
57070 Metz - France
olivier.pietquin@supelec.fr

ABSTRACT

Human machine interaction is a field where machine learning is present at almost any level, from human activity recognition to natural language generation. The interaction manager is probably one of the latest components of an interactive system that benefited from machine learning techniques. In the late 90's, sequential decision making algorithms like reinforcement learning have been introduced in the field with the aim of making the interaction more natural in a measurable way. Yet, these algorithms require providing the learning agent with a reward after each interaction. This reward is generally handcrafted by the system designer who introduces again some expertise in the system. In this paper, we will discuss a method for learning a reward function by observing expert humans, namely inverse reinforcement learning (IRL). IRL will then be applied to several steps of the spoken dialogue management design such as user simulation and clustering but also to co-adaptation of human user and machine.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Reinforcement learning, Human-Machine Interaction

1. INTRODUCTION

Human machine interaction is a field where machine learning is present at almost any level, from speech [14] or gesture [49] recognition to natural language generation [35] or text-to-speech synthesis [8]. Yet, building an interactive system is not only about putting together all this input and output processing modules. There is a need for a intermediate module for managing 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 the late 90's, sequential decision making algorithms like reinforcement learning [45] have been introduced in the field of spoken dialogue systems (SDS) with the aim of making the interaction more natural in a measurable way [18]. This seminal work led to many other applications of reinforcement learning to SDS [42, 29, 17] but also to other types of interacting systems such as tutoring applications [13, 28], museum guides [46], car driving assistance [34], recommender systems [12] and even robotics bar tenders [10].

Many criticisms have been done to the reinforcement learning approach to interaction management [23, 24] among which the fact that too many dialogues were needed to learn an optimal interaction strategy. This has been addressed by many researchers during more than a decade. They either developed realistic user simulation techniques [7, 40, 38, 32, 31] accompanied with speech recognition/understanding error modeling [39, 27] or applied more efficient learning algorithms [19, 11, 30, 9].

Another criticism, that has been much less 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 [47], this reward is indeed generally handcrafted by the system designer who introduces some expertise in the system [18, 29, 48] but also a strong bias. Very little attention has been paid to the particular problem of defining the best reward function for interactive systems.

In this paper, we will discuss a method for learning a reward function by observing expert humans, namely inverse reinforcement learning (IRL) [36, 21]. IRL will then be applied to several steps of the spoken dialogue management design such as user simulation and clustering but also to co-adaptation of human user and machine.

2. INTERACTION MANAGEMENT AS A SEQUENTIAL DECISION MAKING PROBLEM

Reinforcement learning [45] designates a class of algorithms that solve sequential decision making problems. Such

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 270780.

MLIS'13, August 04 2013, Beijing, China
Copyright 2013 ACM 978-1-4503-2019-1/13/08 ...$15.00.

a problem arises when an agent faces a dynamic system that steps from states to states as an effect of the actions of the agent. The agent therefore learns to perform the sequence of actions that makes the system go through desired states. To assess the quality of a state, the agent perceives 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 which is more and more contested in incremental systems [43]), than 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, $R : S \to \mathbb{R}$ is the reward function, $T : S \times A \to \mathcal{P}(S)$ 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 $\pi : S \to \mathcal{P}(A)$ that maps states to actions in such a way that the cumulative rewards 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 this state. This measure is called the value function $V^\pi : S \to \mathbb{R}$:

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

One can define an order on value functions such as $V^1 > V^2$ if $\forall s \; V^1(s) > V^2(s)$. 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 [45], especially when the transition probabilities are not known, and this is still an active research area.

To optimize human-machine interaction management within this framework, one has to cast this task into an MDP. This has been first proposed in the late 90's [18]. 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 [44] which can be approximated as a linear combination of objective measures that can be gathered during the interaction [47]. Yet, this reward is most often a very simple handcrafted function [18, 29, 48]. Although the reward function is an essential component of the optimisation process, very little attention has been paid to offer a good definition for it.

3. INVERSE REINFORCEMENT LEARNING

Defining the appropriate reward function that will lead to a desired behavior is actually a real problem in the field of reinforcement learning. It is sometimes very hard to define in terms of mathematics although it is easy 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 isolate contexts and associate a reward to each possible action in these contexts.

Learning a reward function from demonstrations of the optimal behavior is known as the inverse reinforcement learning (IRL) [36, 21] problem. 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). It also suffers from the same scaling-up problem as reinforcement learning when dealing with large state spaces. For these reasons, many recent algorithms in the literature [1, 20, 15] make the assumption that the reward can be approximated by a linear combination of n features:

$$R(s) = \sum_n \theta_n \phi_n(s) = \theta^T \Phi(s) \qquad (2)$$

Replacing this reward in Eq. 1, we have:

$$\begin{aligned} V^\pi(s) &= E\left[\sum_{i=0}^{\infty} \gamma^i \theta^T \Phi(s_i) | s_0 = s, a_i = \pi(s_i)\right] \\ &= \theta^T E\left[\sum_{i=0}^{\infty} \gamma^i \Phi(s_i) | s_0 = s, a_i = \pi(s_i)\right] \\ &= \theta^T \mu^\pi(s), \end{aligned}$$

where

$$\mu^\pi(s) = E\left[\sum_{i=0}^{\infty} \gamma^i \Phi(s_i) | s_0 = s, a_i = \pi(s_i)\right], \qquad (3)$$

is called the *feature expectation* of policy π which can be seen as a probability of the agent occupying a given state when following the policy π.

Most of the algorithms in the literature try to find an approximated reward $\hat{R}(s) = \theta^T \Phi(s)$ that leads to a policy $\pi_{\hat{R}}$ for which the feature expectation is close to the one of the expert policy π_E. More formally, the algorithms try to minimize, in a more or less direct way, a cost function of the following form:

$$||\mu^{\pi_{\hat{R}}}(s) - \mu^{\pi_E}(s)||_p$$

where $||x||_p$ is the p-norm of vector x.

IRL can be seen as a way to transfer the behavior of an expert to an artificial agent. 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 [24].

4. USER SIMULATION

In the field of spoken dialogue systems, user simulation has been widely studied in the last decade [40, 31]. The necessity of user simulation is often discussed [26] and its effect on policy learning is also criticized [37], but it remains largely used in practice. The main difficulties when developing a user simulation is to make it consistent all along the interaction and according to a goal [25, 38] and to make it trained on real data [33, 41].

In [5], it is proposed to alleviate these two problems thanks to IRL. Using IRL, the user goal (in terms of evaluation of the dialogue flow) is encoded in the reward. Moreover, the behavior of the user is modeled as a policy and not as some probability of selecting actions given the context like in most simulation methods. Thus, the consistency of the behavior is really taken into account. In addition, the reward is learnt on data which also allows transferring actual users' behavior to the simulated agent.

Several experiments have shown that this approach can lead to consistent user simulators and could be used to optimise the dialogue policy. The choice of simulating users via IRL instead of learning a policy for the dialogue management module is motivated by the fact that the user is placed in the same position during data collection than in real application. Learning the reward of a human operator in interaction with users will not lead to an optimal behavior for the system since the role and the perceptions are different. Indeed, the human operator doesn't suffer from speech recognition errors and doesn't target the same goal has the machine.

5. USER BEHAVIOUR CLUSTERING

When collecting a dataset with several users, it is rare to have an homogeneous population of testers. For instance, some of them are novices and others are expert users of interactive systems. Considering the data as homogeneous for user simulation or direct policy training on the data would lead to a policy adapted to an average user which actually doesn't exist in reality.

Considering user simulation as an IRL problem also leads to an original method for behavior clustering [4]. Indeed, Eq. 3 provides a way for describing a policy by a fixed-size vector in the form of the feature expectation. Seeing each user behavior as a policy π_i, one can use μ^{π_i} in a vector quantification method so as to exhibit different homogeneous clusters.

Given this, one can build several user simulations (one for each cluster) and train a dialogue policy for each prototypic behavior [3].

6. COADAPTATION

One other problem arising when training a dialogue policy with users simulation or from a fixed-set of data is to take into account to phenomenon of coadaptation. This phenomenon occurs when a human-machine interface adapts to the user behavior who, in turns, adapt his/her behavior to the new interface policy. This can result in unstable equilibria which cannot be simulated by standard methods. Using IRL, describing the behavior of users by a reward instead of rules or probabilities may bring a solution to this specific problem [6].

7. CONCLUSION AND PERSPECITVES

Other applications can benefit from imitation learning via IRL in the field of human-machine interaction. For instance, developing avatars able to display emotions or being able to laugh at the appropriate time and in the appropriate manner is hard to do with standard methods in artificial intelligence or machine learning [22].

Yet, IRL is still a very young field of research and most of existing algorithms still suffer from the curse of dimensionality and the fact that, once a reward is obtained, it has to be optimized. Yet it is an active research field and solutions are being proposed [16].

8. REFERENCES

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

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

[3] S. Chandramohan, M. Geist, F. Lefèvre, and O. Pietquin. Behavior specific user simulation in spoken dialogue systems. In *Proceedings of the ITG Symposium of Speech Communication*, pages 1–4, 2012.

[4] S. Chandramohan, M. Geist, F. Lefevre, and O. Pietquin. Clustering behaviors of spoken dialogue systems users. In *Proceedings of the Intenational Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, Kyoto, Japan, 2012.

[5] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, et al. User simulation in dialogue systems using inverse reinforcement learning. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 1025–1028, 2011.

[6] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, M.-I. Supelec, and F. Metz. Co-adaptation in spoken dialogue systems. In *Proceedings of the Fourth International Workshop on Spoken Dialog Systems*, Ermenonville, France, 2012.

[7] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira. Human-computer dialogue simulation using hidden markov models. In *Procedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2005)*, pages 290–295, San Juan, Mexico, 2005.

[8] R. Damper. *Data-Driven Techniques in Speech Synthesis.* Telecommunications Technology & Applications Series. Springer, 2001.

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

[10] 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.

[11] M. Gašić, F. Jurčíček, 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 the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 201–204. Association for Computational Linguistics, 2010.

[12] 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.

[13] 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.

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

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

[16] E. Klein, B. PIOT, M. Geist, and O. Pietquin. A cascaded supervised learning approach to inverse reinforcement learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2013)*, Prague (Czech Republic), September 2013.

[17] O. Lemon and O. Pietquin. Machine learning for spoken dialogue systems. In *Proceedings of the European Conference on Speech Communication and Technologies (Interspeech'07)*, pages 2685–2688, Anvers, Belgium, 2007.

[18] E. Levin, R. Pieraccini, and W. Eckert. Using Markov decision process for learning dialogue strategies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 98)*, volume 1, pages 201–204, Seattle, Washington, USA, 1998.

[19] L. Li, S. Balakrishnan, and J. Williams. Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech 2009)*, volume 9, Brighton, United Kingdom, 2009.

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

[21] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning*, pages 663–670, Stanford, CA, USA, 2000.

[22] 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 the Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS2013)*, Saint Paul, USA, May 2013.

[23] 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.

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

[25] O. Pietquin. Consistent goal-directed user model for realisitc man-machine task-oriented spoken dialogue simulation. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 425–428, Amsterdam, Netherlands, 2006.

[26] O. Pietquin. Statistical user simulation for spoken dialogue systems: what for, which data, which future? In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 9–10, Montreal, Canada, 2012.

[27] O. Pietquin and R. Beaufort. Comparing ASR Modeling Methods for Spoken Dialogue Simulation and Optimal Strategy Learning. In *Proceedings of the 9th European Conference on Speech Communication and Technologies (Interspeech/Eurospeech)*, pages 861–864, Lisbon (Portugal), September 2005. ISCA.

[28] 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.

[29] 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.

[30] 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 (TSLP)*, 7(3):1–21, 2011.

[31] O. Pietquin, H. Hastie, et al. A survey on metrics for the evaluation of user simulations. *Knowledge Engineering Review*, 15, 2011.

[32] O. Pietquin, S. Rossignol, and M. Ianotto. Training Bayesian networks for realistic man-machine spoken dialogue simulation. In *Proceedings of the 1rst International Workshop on Spoken Dialogue Systems Technology (IWSDS 2009)*, Irsee (Germany), December 2009. 4 pages.

[33] O. Pietquin, S. Rossignol, and M. Ianotto. Training Bayesian networks for realistic man-machine spoken dialogue simulation. In *Proceedings of the 1rst International Workshop on Spoken Dialogue Systems Technology (IWSDS 2009)*, Irsee (Germany), December 2009. 4 pages.

[34] 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.

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

[36] S. Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103, Madison, Wisconsin, USA, 1998.

[37] 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 workshop on Automatic Speech Recognition and Understanding (ASRU'05), San Juan, Puerto Rico*, December 2005.

[38] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for*

Computational Linguistics; Companion Volume, Short Papers, pages 149–152. Association for Computational Linguistics, 2007.

[39] J. Schatzmann, B. Thomson, and S. Young. Error simulation for training statistical dialogue systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2007)*, pages 526–531, Kyoto, Japan, 2007.

[40] 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, 2006.

[41] J. Schatzmann and S. Young. The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):733–747, 2009.

[42] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of the second international conference on Human Language Technology Research*, pages 12–19, San Diego, Californie, USA, 2002. Morgan Kaufmann Publishers Inc.

[43] D. Schlangen and G. Skantze. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718, 2009.

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

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

[46] 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.

[47] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

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

[49] M. Yang and N. Ahuja. *Face Detection and Gesture Recognition for Human-Computer Interaction.* International Series in Engineering and Computer Science. Kluwer Academic pub., 2001.