The Grand Challenge of
Predictive Empirical Abstract Knowledge

Richard S. Sutton
Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8

Abstract

We survey ongoing work at the University of Alberta on an experience-oriented approach to artificial intelligence (AI) based an reinforcement learning ideas. We seek to ground world knowledge in a minimal ontology of just the signals passing back and forth between the AI agent and its environment at a fine temporal scale. The challenge is to connect these low-level signals to higher-level representations in such a way that the knowledge remains grounded and autonomously verifiable. The mathematical ideas of temporally abstract options, option models, and temporal-difference networks can be applied to begin to address this challenge. This has been illustrated in several simple computational worlds, and we seek now to extend the approach to a physically realized robot. A recent theoretical development is the extension of simple temporal-difference methods to off-policy forms using function approximation; this should enable, for the first time, efficient and reliable intra-option learning, bringing the goal of predictive empirical abstract knowledge closer to achievement.

Introduction

The Predictive Empirical Abstract Knowledge (PEAK) project at the University of Alberta is a radical attempt to understand world knowledge in terms of a minimal ontology of sensori-motor experience. Experience is defined as the time sequence of low-level signals passing back and forth between the AI agent and its world at some relatively fast rate, say 100 times a second. The signals passing from the world to the agent are termed *sensations*, and the signals from the agent to the world are termed *actions*. For concreteness, time is taken to be discrete. The minimal ontology is then exactly these three things: sensations, actions, and time steps. The PEAK project explores the hypothesis that all world knowledge can be precisely characterized as predictions about the relationships among these three things, without reference to any other concepts or entities except insofar as they themselves can be precisely characterized in terms of the minimal ontology.

The primary challenge to the PEAK hypothesis is the mismatch between low-level experience and human-level world knowledge as we normally think of it. The gap between even relatively simple concepts, such as that of a cup or a chair, and low-level 100-times-a-second experience can seem immense. Thus the PEAK project is appropriately focused on the issue of abstraction. Its primary objective is to stretch our imagination through examples and implemented systems until bridging the abstraction gap seems possible and plausible.

Grounding knowledge in experience is extremely challenging, but may bring an equally extreme benefit. Representing knowledge in terms of experience enables it to be *compared with* experience. Knowledge imparted by human experts can be verified or disproved by this comparison. Existing knowledge can be tuned and new knowledge can be created (learned). The overall effect is that the AI agent may be able to take much more responsibility for maintaining and organizing its knowledge.

The ability of an AI system to self-verify its knowledge is indeed a substantial benefit. While large amounts of knowledge is a great strength of AI systems, it is also a great weakness. The problem is that as knowledge bases grow they become brittle and difficult to maintain. There arise inconsistencies in the terminology used by different people or at different times. Errors are inevitably present. When an error becomes apparent, the problem can only be fixed by a human who is expert in the structure and terminology of the knowledge base. This puts an upper bound on the size of the AI system's knowledge base. As long as people are the ultimate guarantors of truth, then the machine cannot become much smarter than its human handlers. In this sense, verifying knowledge by consistency with human knowledge may inevitably be a dead end.

Predictive Knowledge

Much everyday knowledge is clearly predictive. To know that Joe is in the coffee room is to predict that you will see him if you went there, or that you will hear him if you telephoned there. To know what's in a box is to predict what you will see if you open it, or hear if you shake it, feel if you lift it, and so on. To know about gravity is to make predictions about how objects behave when dropped. To know the three-dimensional shape of an object in your hand, say a teacup, is to predict how its silhouette would change if you were to rotate it along various axes. A teacup is not a sin-

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

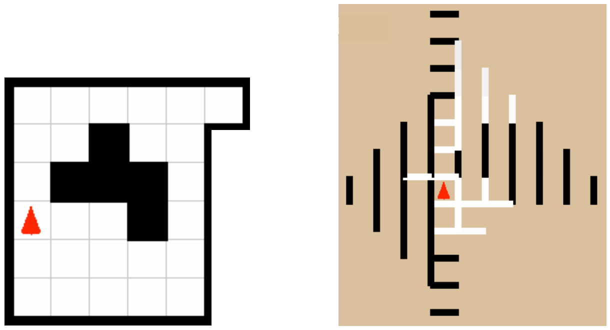

Figure 1: On the left is a bit-to-bit world, a world with one bit of sensation and one of action; the triangle shows the agent's position and orientation. On the right is a subjective, action-conditional representation of the agent's local knowledge. The colored bars each represent a prediction that the agent is making. For example, the agent predicts that if it steps forward three times then it will see a wall.

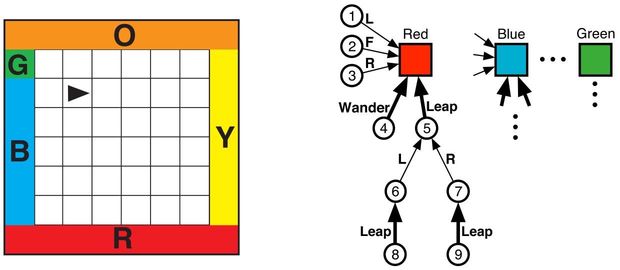

Figure 2: The compass world (left) and a portion of the structure of the corresponding TD network representing the agent's knowledge (right). L, F, and R are elementary actions, and Leap and Wander are temporally abstract options. Node 9, for example, will hold a numerical prediction of the probability of finally sensing Red if the agent were to step Forward until it sensed non-White (the Leap option), turn Right (the R action), and then step Forward again until sensing non-White (the Leap option again).

gle prediction but a pattern of interaction, a coherent set of relationships between action and sensation.

Although these examples of predictive world knowledge are substantially abstracted from the minimal ontology, a formal argument can be made that they all must be reducible to statements about low-level sensations and actions. By world knowledge we mean knowledge about a particular world, not knowledge that is true in any world, such as knowledge of mathematics or logic. Let us denote the action taken at time t as $a_t \in \mathcal{A}$, and the sensation generated at time t as $s_t \in \mathcal{S}$. Experience then is the sequence of intermingled actions and sensations $s_1, a_1, s_2, a_2, s_3, a_3, \ldots,$ each element of which depends only on those preceding it. Define $\mathcal{E} = \{\mathcal{S} \times \mathcal{A}\}^*$ as the set of all possible experiences. Let us call the experience sequence up through some action a *history*. Formally, any world can be completely specified by a probability distribution over next sensations conditional on history, that is, by the probability $P(s|h)$ that the next sensation is s given history h, for all $s \in \mathcal{S}$ and $h \in \mathcal{E}$. To know P exactly and completely is thus to know everything there is to know about the world.

PEAK Systems in Computational Worlds

Figures 1 and 2 show two illustrations of PEAK systems from previous work. These initial systems are simple but still instructive and representative of directions that could be pursued further.

A Bit-to-Bit World

The world shown in Figure 1 (Tanner & Sutton, 2005; Tanner, 2006) is an instance of a *bit-to-bit* world, a world with only one bit of sensation and one bit of action. The agent, shown as a triangle, can be in any cell and oriented in any of four directions, but it can sense only whether the cell in front of it is open or blocked; it cannot directly sense its position in the grid. Actions are similarly limited to a single binary choice: the agent can either step forward, or turn in place in one direction (to the right). This is the simplest possible sensori-motor interface for a decision-making agent. Even so, this small interface does not limit the size of the world that can be represented; by attending to the sequence of action and sensation bits, an agent could have complete knowledge of, and be able localize within, a world of arbitrary size and complexity. The world shown in the figure is sufficiently small that complete knowledge of it was represented and learned using a temporal-difference (TD) network (Sutton & Tanner, 2005), a form of predictive state representation (Littman, Sutton & Singh, 2002; Rosencrantz, Gordon & Thrun, 2004; Jaeger, 2000). Larger worlds would require a more efficient learning algorithm than was used in this work (see below). Other natural extensions of this work would be to consider larger grids, to introduce stochastic elements into the sensation and action bits, and to consider bit-to-bit worlds with non-grid dynamics.

The Compass World

The world shown in Figure 2, the "compass world" (Rafols, 2006; Sutton, Rafols & Koop, 2005), has a slightly larger sensori-motor interface. The agent can sense the color of the cell in front of it, as before, but here that color can take on any of six values: Red, Blue, Green, Orange, Yellow, and White. There are now three actions: step Forward, turn Right, and turn Left. The TD network here involves abstraction not just in state but also in time. In addition to the primitive actions, the agent can also represent knowledge using temporally abstract macro-actions known as *options* (Precup, 2000; Sutton, Precup & Singh, 1999). Options are closed-loop policies defined here entirely in terms of sensations, actions, and predictive state representations. For example, in this system, there are two options, Leap and Wander. Leap is the policy of taking the Forward action until sensing non-White, and Wander is the policy of acting randomly until sensing non-White. The TD network in this illustration (suggested by the right panel of Figure 2) is ca-

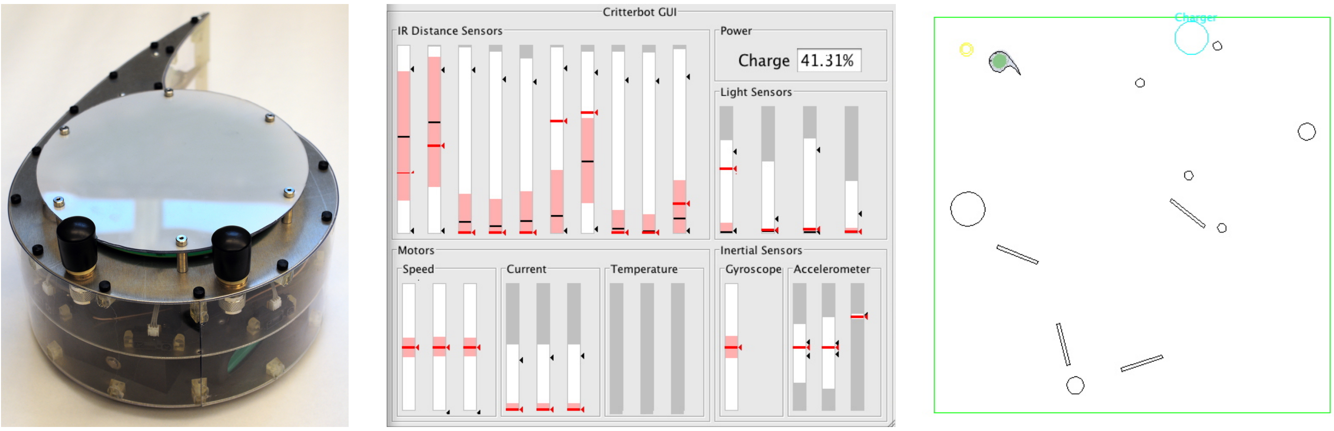

Figure 3: Left) the Critterbot, a sensor-rich mobile robot approximately one foot in diameter; middle) a GUI visualization of its sensor values; right) a 2D Critterbot simulator, with the simulated Critterbot in the upper left of a bounded field with various objects. Agent programs and the GUI can be used interchangeably with the physical Critterbot and the simulator.

pable of representing and learning abstractions in state and time that can be composed to build abstractions upon abstractions. For example, in the state shown in the left panel of Figure 2, the agent predicts that Leaping (stepping Forward until non-White) will result finally in sensing Yellow. This prediction will form part of its state representation for the state shown. These state abstractions and action abstractions can be flexibly interrelated and composed. For example, the agent knows that, in the state shown, if it were to turn Right, then it would be in a state where it will predict that Leaping would produce Red. In other words, if you are 'facing a yellow wall' then, if you turn Right, you will be 'facing a red wall', where the quoted phrases are fully, explicitly grounded in sensations and actions. Moreover, such temporally abstract knowledge can be maintained without constant sensory clarification. In fact, the agent can make many turns in the center of the compass world, without sensing any of the walls, and keep track indefinitely of which wall it is facing.

A Robot for Exploring PEAK

Computational worlds such as the bit-to-bit world and the compass world have the advantage that everything about them can be precisely controlled and understood. However, this is also a disadvantage in that we might deceive ourselves by designing worlds that are easy for our PEAK methods but not reflective of the real worlds of interest. An alternative approach that partially alleviates this problem, and which we are pursuing in parallel with studies of computational worlds, is to use physically realized robotic systems. A real robot forces one to come to grips with temporal issues such as sensing and acting delays, asynchrony of perception and action, and the need for real-time responses. It can also provide a focus for group activity, a single system that can be addressed in multiple ways while reflecting a consensus about objectives.

To these ends, we have over the last year and a half designed and constructed a small, sensor-rich mobile robot, the *Critterbot*, shown in the left panel of Figure 3. The sensors include infrared proximity detectors going out to about one meter in ten different directions, light sensors in four directions, binaural microphones, a three-axis accelerometer, a gyroscope, a radio-spectrum sensor that can detect WiFi base-station signal strengths, a compass, sensors for the battery level, and three sensors each for wheel velocity, motor current, and motor temperature. A real-time display of the instantaneous sensor values and various statistics of their recent values is shown in the second panel of Figure 3. Bump or contact sensors are currently being designed, and a camera is planned for the future. We have deliberately avoided laser range-finders and other sensors that would tempt us to think in cartesian rather than experience-centered terms. The only physical actuators are the three motors driving the wheels of a holonomic-drive system that enables the robot to translate and rotate independently and simultaneously in any direction. The robot can also express itself through a speaker and a circle of bright color LEDs around its upper surface.

The robot runs at a natural time cycle of 100Hz. Every 1/100th of a second, all sensor values are read and transferred to the control program, and an output is taken from the control program that directs the control of the motors. The control program may run directly on the robot, which includes a 500MHz x86 processor with 1GB of RAM and a full-function Gentoo unix environment, or it can run off-robot using the wireless interface, with some additional delay. Overall, there may be a delay of 3–5 cycles between emitting a motor command and sensing its consequences as motor velocities. We have also built a 2D Critterbot simulator (see right panel of Figure 3) which supports the same sensori-motor interface as the physical robot, though the dynamics are inevitably slightly different. A standard RL-Glue interface is available to both the robot and simulator. In programming the Critterbot we must directly and immediately face the challenge of a dense stream of low-level sensorimotor information. The Critterbot will be a focus of much of our future work with PEAK systems.

New Gradient-Based Learning Algorithms

The systems described earlier in this paper were handicapped by their use of learning algorithms that do not scale well. The natural learning algorithms are those that can learn about an option even from fragments of its execution, known as *intra-option learning* algorithms (Sutton, Precup & Singh, 1998, 1999). Intra-option learning holds out the possibility of learning simultaneously about a great many different options at once from a single rich stream of sensorimotor experience. However, this plan runs afoul of the problem of *off-policy learning*, one of the greatest standing problems in reinforcement learning. Off-policy learning refers to learning about one policy while following another. Off-policy learning most commonly arises in algorithms such as Q-learning that learn about an optimal or greedy policy while actually following a more explorative policy such as softmax or ϵ-greedy (Sutton & Barto, 1998). It also arises here as we try to learn simultaneously about many different options each with their own internal policy, while actual behavior is generated by at most one of those policies. The problem of off-policy learning is that conventional TD learning procedures, such as Q-learning and TD(λ), are not completely sound when trained off-policy and when using linear function approximation; for some problems and policies their parameters diverge to infinity (Baird, 1995). There have been numerous partial solutions (some of which were used in the example systems above) but none has been completely satisfactory. Some require a restricted form of function approximator, some scale quadratically rather than linearly in the number of features, and some have high variance.

Very recently, however, new algorithms have been developed which may solve essentially all of these problems (Sutton, Maei, Precup, Bhatnagar, Silver, Szepesvári & Wiewiora, 2009). These new algorithms are variations on conventional TD learning procedures that adhere more thoroughly to the idea of gradient descent. As such, it is straightforward to prove that they converge under general on- or off-policy training. Empirically, the latest gradient TD methods appear to learn at a rate comparable to conventional TD methods on on-policy problems, while converging reliably on off-policy problems. The new methods are more complex than conventional methods, but only by a factor of two (memory and computation). Although it has not yet been done as of this writing, it should be straightforward to extend these methods to intra-option learning methods that can be used off-policy, and thus learned in parallel from a single stream of experience. This would be a significant step toward solving the learning part of the PEAK challenge.

Conclusion

We have illustrated in computational worlds that PEAK systems can build non-trivial abstract concepts from the minimal ontology of sensations, actions, and time steps. Extensions to a physical robot and to the new more-scaleable gradient-based learning algorithms are currently being developed. Although the gap between low-level experience and human-level knowledge remains immense, perhaps it can be bridged bit by bit. If so, it might remake our entire conception of knowledge in knowledge-based systems, enabling them to self-maintain, and thus to become much larger than has hitherto been possible. This is a large prize and deserves a sustained effort. That is the grand challenge of predictive empirical abstract knowledge.

References

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37.

Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural Computation 12*(6):1371–1398.

Koop, A. (2007). *Investigating Experience: Temporal Coherence and Empirical Knowledge Representation*. University of Alberta MSc. thesis.

Littman, M., Sutton, R. S., Singh, S. (2002). Predictive representations of state. *Advances In Neural Information Processing Systems 14*, pp. 1555–1561. MIT Press.

Precup, D. (2000). *Temporal Abstraction in Reinforcement Learning*. University of Massachusetts PhD thesis.

Rafols, E. J. (2006). *Temporal Abstraction in Temporal-difference Networks*. University of Alberta MSc. thesis.

Rosencrantz M., Gordon G. J., Thrun S. (2004). Learning low dimensional predictive representations. *Proceedings of the 21st International Conf. on Machine Learning*.

Sutton, R. S., Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, Cs., Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. *Proceedings of the 26th International Conference on Machine Learning*.

Sutton, R. S., Precup, D., Singh, S. (1998). Intra-option learning about temporally abstract actions. *Proceedings of the 15th International Conference on Machine Learning*, pp. 556–564.

Sutton, R. S., Precup, D., Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence 112*, pp. 181–211.

Sutton, R. S., Rafols, E. J., Koop, A. (2006). Temporal abstraction in temporal-difference networks. *Advances in Neural Information Processing Systems 18*.

Sutton, R. S., Tanner, B. (2005). Temporal-difference networks. *Advances in Neural Information Processing Systems 17*.

Tanner, B., Sutton, R. S. (2005). Temporal-difference networks with history. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*.

Tanner, B. (2006). *Temporal-Difference Networks*. University of Alberta MSc. thesis.