Acquisition of Procedural Commonsense Knowledge in a 3-D Simulated World

Moin Ahmad
MIT Media Lab
20 Ames St. E15-385
moin@mit.edu

ABSTRACT
In this paper we describe the design of a solution for the acquisition of procedural commonsense knowledge in a situated simulated environment. In this system, users in a simulated game environment, control virtual robots to do various tasks in a real world scenario. In the current system, users are situated in a restaurant scenario and take charge of teaching their apprentice robots actions to complete tasks. The main benefit of situating acquisition of knowledge in a real world environment is that procedures and actions in situations get learnt in many different realms of thought and at many different levels of detail ultimately learning a task in many ways.

ACM Classification Keywords
Commonsense, Knowledge Acquisition, Procedural Knowledge

INTRODUCTION
Giving computers and programs commonsense knowledge has been a major challenge for Artificial Intelligence research for the past half century. Even though computers can do a variety of automated tasks, no current computer program or system can claim to provide resourcefulness and commonsense thinking capacity that humans demonstrate in common everyday life. This lack of commonsense in computers has led to brittleness in machines which are unable to understand high level goals and motivations of its users. For example, a highly successful computer program like Amazon.com cannot suggest options for clothes that people might want to buy to wear to a party. This is due to a lack of understanding of the purposes and goals that humans want to achieve in a party and how different types of clothes can try to fulfill parts of those goals. The problem becomes even harder if the program needs to suggest clothes for parties in multiple cultures.

One of the big reasons for the difficulty in providing such features is lack of commonsense knowledge about typical human social situations. Humans often learn such knowledge by developing a very large memory of procedures and facts of daily life. This knowledge coupled with a large variety of reasoning and memory retrieval methods, humans are able to make fast commonsensical decisions about daily life. For making computers have commonsense we need to replicate human like reasoning and memory organization in computer programs. This project focuses specifically on acquiring procedural knowledge in a simulated world.

LEARNING SOMETHING IN MANY DIFFERENT WAYS
Marvin Minsky has often articulated "You don't understand anything until you learn it more than one way."[1]. Such resourcefulness in humans comes from the enormous variety of ways they learn in a particular situation (or episode), giving rise to a rich episodic memory. In Minsky's view of the human mind as a 'cloud of resources', episodic learning allows rich 'credit-assignment' to resources that get engaged in the learning process. Such resources can then be activated in slightly different or very different, but analogous situations, allowing their reuse in many different problem solving situations. In addition, by using analogy and imagination, humans can extend or simplify situations to hypothesize and reason about very large number of situations many of which they might have never actually encountered.

What does it mean to learn something in many different ways ? There are many meanings to this statement itself.

- *Learning something in multiple representations*: One can sit on a box as well as store something in it. A computer is a social communication device as well as an electrical equipment. An electrical cord is like a rope and can transfer electrical current.
- *Learning at multiple levels of abstractions*: A table is functionally an elevated surface to keep food on. It can also be broken down into a flat surface and its legs. Each of which may consist of their sub parts until they are broken down into molecules and atoms. To grasp something, one first moves the body, then the arms and finally uses the fingers to hold it.
- *Learning many ways to solve the same problem*: To move to another place, one can walk, run, go in a car or be carried by someone. One can solve a problem by divide and conquer or use contradiction to prove it unsolvable. One can count objects one by one, or can count by hierarchically grouping them together. One can talk to

another person by being physically near him or by writing and email or talking on phone.

- *Learning in multiple realms of thought*: Food can be eaten by buying it (economic), stealing or begging it (ownership) or growing it (biological). A restaurant can be a place for satisfying hunger (bodily), a place for social interaction (social) or a place for work (economic).
- *Assigning credit to relevant mental resources*: The hardest of all is learning multiple things at the same time. You can recognize a face by a line diagram, a full color photograph or a 3d model. Our visual system has multiple resources that get associated with the same object allowing us to activate such resources at various levels of detail.

SITUATED ACQUISITION OF COMMONSENSE

In contrast to declarative knowledge present in a commonsense knowledge database like Cyc [3], which does not address any of the issues presented above, knowledge acquired in a simulated world can have many of the above characteristics. Since actions are situated in the world, resources can recognize many different features of the scenario and assign relevant credit. In a simulated world, particularly attractive is the ability to re-enact or imagine the scene and do mental simulations of situations encountered. This can also be helpful in extending or modifying previous knowledge when better abstractions or representations are learnt, without the need for re-teaching the scenario.

The first implementation of Minsky's Emotion Machine architecture EM-ONE [6] was implemented by Push Singh. The knowledge base in his implementation at its core consists of episodes of problem solving narratives which are used, using case based reasoning and a set of mental critics, to generalize to other previously un-encountered situations. The larger the number of narratives that EM-ONE has, the better its reasoning capabilities will be. This system extends EM-ONE to allow addition of narratives by allowing users to teach virtual robots to accomplish a variety of tasks in a restaurant environment.

APPROACH

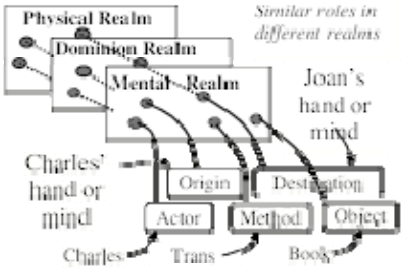
To allow EM-ONE to work in a restaurant environment, we need to make it work with a game engine that can take care of the rendering, physics and other constraints.

The Torque Game Engine

We use the Torque game 3-D engine for developing the simulated world. We are using this engine instead of the original Roboverse environment used in EM-ONE because of its ability to import an extensive variety of objects and its ability to work over the network—which in particular is important so that multiple robots can be controlled by separate EM-ONE clients. In the interface we are exploring how we might be able to give users options to create new high level actions using low level actions and classify the changes into different realms. So users might be able to extend the actions already available in the system.

Realms of Thought

Marvin Minsky in The Society of Mind describes that thinking and problem solving happens in multiple realms of thought.



These realms enable organization of commonsense knowledge in separate clusters in which actions affect changes.

Realms and Representations

In the restaurant environment, we need multiple realms of thought in which interactions take place. Reasoning in realms of thought are done in different representations. Actions typically will have frame-slots that span one or more realms. For instance, the eating food in a restaurant script will change dominion (food transfer), economic (money exchange), bodily (food inside body, satisfy-hunger) etc.

We use the following realms in our implementation:

- *Visual*: Objects in view, containment (in, out), relative-position (on, under).
- *Spatial*: Abs-location, relative-location, near, far.
- *Physical*: Lift an object, support an object on another.
- *Bodily*: Body-location, needs, position.
- *Social*: Facing, communicating, order request.
- *Dominion*: Object possession.
- *Economic*: Monetary.
- *BDI:* Beliefs, desires, intentions, goals of actors.

Trace of Actions

Procedures in this system are essentially, detailed traces of actions done and the changes in the differences observed in various representations. An example trace of actions is (realm changes in rectangular brackets):

Walk 3 steps [visual, spatial]
Pick up food [physical, dominion]
Sit on chair [bodily, visual, physical]

Order to waitress [social, bdi]
Eat food [bodily, visual]

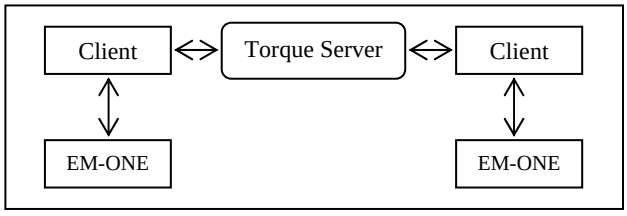
After having a large number of scenarios, using credit assignment algorithms robots will be able to select the most likely action that will reduce a certain difference that will help achieve the current goal. For instance, the act of walking can change the containment relationship (being in and out of a restaurant). The act of eating will change hunger and not position. Some higher level tasks may use tasks from other realms to get accomplished. In a simulated world, one can possibly run automated simulations of tasks to organize realms and task hierarchies.

USING EM-ONE FOR REFLECTIVE THINKING

The main contribution of EM-ONE is the ability to write critics that can critique on problem solving predicaments and use previously learnt narratives to generate better hypotheses. However, in its implementation, EM-ONE lacks an environment where a robot engages in an actual real-world task, due to which its commonsense narratives are not useful for daily human scenarios. One of the objectives of our system is to allow EM-ONE to connect and control robots in the restaurant. This will allow development of a thinking engine that is able to reason in such a real world environment as well as acquire commonsense procedural knowledge in it.

ARCHITECTURE

The system has a simple architecture typically used by multiplayer gaming environment. A server maintains the game state and synchronized clients do actions in the game. EM-ONE is a rule based system that polls for sensory information. We will implement representational extraction in the Torque client environment and EM-ONE will be modified to work on the network and poll for representational information.



A TYPICAL SCENARIO

In this section I will describe a scenario to illustrate what kind of knowledge is acquired in the simulated environment. The screenshot below is of the environment built by Jeff Orkin in the Torque 3D engine using objects imported from the SIMS game. The main scenario is based in a restaurant where four actors are present—guest, waitress, cook and bartender. The scenario also contains a number of objects like table, chair, menu, glass, plate, food that can be manipulated using various actions that are available on the objects like pick-up, eat, talk-to, sit etc. Various high-level representational realms will be implemented using the game engine. For example, a list of objects in the current viewpoint can be generated, or low level information like the distance between objects can be calculated.

Scenario: Eating food in a restaurant.

The scenario illustrates the acquisition of script for eating food in a restaurant. Some changes are not reported. The scenario starts with one of the actors outside the restaurant.

Initial state of the guest actor: {hungry, outside restaurant}

Action	Some predicates
Turn	in-view (tree, restaurant (cash register))
Walk	near-restaurant
Walk	in-restaurant, in-view(restaurant objects)
Turn	in-view (many objects)
Walk	near-chair
Sit	on-chair
Ask order	(knowledge, belief, expectations change)
Eat food	satisfy-hunger
Ask check	(knowledge, belief, expectations change)
Pay check	monetary-transfer
Get up	~on-chair
Turn	in-view and ~in-view for many objects
Walk	~in-restaurant

In the similar way we can illustrate a scenario when the waitress waits on the guest. It would be interesting to explore the area how a guest can use the knowledge of the waiting process to engage social reflection to infer what the waitress is possibly doing.

CONCLUSION

As work in progress, this project is still in the planning and hacking phase. Commonsense acquisition is an enormously complex activity and requires joint working of many different artificial intelligence algorithms and representations. Most commonsense acquisition in the previous years has been focused on passive acquisition of declarative commonsense knowledge. In this project we focus on actively acquiring commonsense in a real world scenario. The main benefit of this approach is the detail in script acquisition which will eventually allow formation of richer episodic memories and enable learning tasks in many different ways.

ACKNOWLEDGMENTS

I thank Henry Lieberman and Junia Anacleto for teaching the Commonsense for Interactive Applications class at MIT. I also thank Dustin Smith, Bo Morgan, Jeff Orkin, Marvin Minsky and Barbara Barry for many fruitful discussions on architectures for commonsense reasoning.

REFERENCES

1. Minsky, M. *The Society of Mind*. Simon & Schuster, New York, NY, 1988.
2. Minsky, M. *The Emotion Machine*. Simon & Schuster, New York, NY, 2006.
3. Lenat, D. *CYC: A large-Scale Investment in Knowledge Infrastructure*, Communications of the ACM, Nov 1995, Vol. 38, No. 11.
4. Singh, P., and Barbara, B. *Collecting commonsense experiences*. (K-CAP 2003). Sanibel Island, FL.
5. Singh, P., *The public acquisition of commonsense knowledge*. Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access, Palo Alto, CA
6. Singh, P., *EM-ONE: An Architecture for Reflective Commonsense Thinking*, Ph.D. thesis, MIT, 2005.
7. Singh, P. *The Panalogy architecture for commonsense computing - Brief Description*, Report for the Institute for Defense Analysis, 2003.
8. Lockerd, A. *Acquiring Commonsense Through Simulation* (unpublished).
9. Mueller, Erik T., *ThoughtTreasure*, http://www.signiform.com.
10. Torque Game Engine. http://www.garagegames.com