TOWARDS A MODEL OF CONCEPTUAL KNOWLEDGE ACQUISITION THROUGH DIRECTED EXPERIMENTATION

Shankar Rajamoney
Gerald DeJong
Boi Faltings

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana,IL 61801

ABSTRACT

Most current Artificial Intelligence systems require a complete and correct model of their domain of application. However, for any domain of reasonable size, it is not feasible to construct such a model. The main thrust of this project is to build a system that can continuously update its model through a constant monitoring of the real world. The project involves the development of a system that starts with an incomplete and incorrect model of the world. While performing its tasks the system is occasionally confronted by observations which are inconsistent with its current beliefs. It attempts to explain these observations by hypothesizing reasons for the inconsistencies and devising experiments to pinpoint the flawed belief. Based on the results of the experiments the system revises its beliefs to accommodate the previously inconsistent observations.

I INTRODUCTION

Machine Learning is of increasing importance in Artificial Intelligence [Carbonell82, Michalski83, Mitchells3, SchankS2, Winston83j. In this paper, we shall describe a new form of explanation-based learning which involves designing and conducting experiments in the "rear world to help with the explanation of some observation. When human learning behavior is considered, it seems that there are two classes of explanatory learning to be distinguished. The first is the acquisition of *schematic* knowledge, U. the high-level learning of plans and scripts (e.g„ how to behave at a restaurant). This form of learning does not, in principle, enhance the power of the knowledge that is already there, but makes its application more efficient via chunking with schemata. The second is the acquisition of *conceptual* knowledge. This form of learning adds new capabilities that the previous knowledge did not provide.

Note that there are marked differences between the two forms. Schematic knowledge may be learned from a single observation. For example, the basic structure of a script for kidnapping might be clear from a single observation, provided there is enough knowledge to understand the dependencies between the individual events in that situation. This has been demonstrated in work on explanatory schema acquisition [DeJong82l

Acquiring new concepts, on the other hand, is more difficult. When people discovered radioactivity, it took them a long time to formulate the proper concept. This is a case where the previous knowledge is not sufficient to explain the observations and thus *reasoning has to be replaced by experimentation.* Another point to observe in this example is that new concepts are not discovered by a *search* for them, but by noticing *discrepancies* between the world model predictions and the way the world behaves. New concepts arise out of the necessity to explain these discrepancies while maintaining a knowledge base consistent with the earlier observations.

This work wis supported in part by the Air Force Office of Scientific Research under gram F49620-82-K-0009 and in part by the National Science Foundation under gwnt NSF-IST-8M7889.

• IBM Fellow

II THE SYSTEM

A. System Overview

Our system has a world model of its domain that drives its reasoning. A number of beliefs are implicit in the structure of the world model. When situations arise that cannot be explained with the current world model thereby leading to a contradiction, the system starts questioning its beliefs. It first questions beliefs directly linked to the contradiction. Only if these fail to give a consistent explanation does it question secondary causes of the contradiction, beliefs on which the primary beliefs rested, and the beliefs behind the questioning and investigation process. The underlying assumption is that the number of errors in the model is small. The system performs a series of experiments directed at finding the belief at fault. Once the belief has been identified it is revised to give a model consistent with the current observations.

The planner used by the system predicts the observations the system will make when the plans are executed. The system constantly compares these predictions with the actual observations. Whenever it finds changes in the world that were not a consequence of some plan, it tries to explain them as an effect of processes running independently from the system (for example, evaporation).

One of the system's implicit beliefs is that liquids cannot pass through solids. While carrying out its normal activities, the system inadvertently encounters an example of osmosis, a process unknown to the system. Osmosis is a natural phenomenon that occurs when two solutions are separated by a permeable solid. If the concentrations of the two solutions are different then a flow of solvent through the solid takes place in such a way as to minimize the difference in concentrations.

The system's observation of osmosis cannot be explained with its current knowledge of the world. The information deduced from the contradiction enables the system to devise experiments to determine which of its beliefs is wrong and to discover some characteristics of the osmosis process.

The observation of the phenomenon of osmosis results in the system revising its world model. It modifies its belief to allow for liquids to pass through solids under special circumstances. In the process, it discovers the concept of permeability of solids and a new form of flow which we know as osmosis.

B. The Osmosis Example

The system is given a goal of forming a mixture of specific amounts of solutions, #$Solutionl and #$Solution2, of known concentrations, concl and conc2 (concl > conc2). In order to achieve this goal it generates a subgoal of temporarily storing the specified amounts of the two solutions in containers, As it happens the only suitable container has two compartments separated by a permeable membrane. The planner, not realizing the significance of the partition, plans to pour the two solutions of differing concentrations into the two compartments. As a part of its monitoring of the real world, it expects a number of observations including a decrease in

the amount of solutions in the two original containers, and the appearance of specified amounts of solution in the two compartments of the selected container.

C. Contradiction Detection

The system verifies the predictions made by the planner by comparing the predictions with its input observation stream. Immediately alter execution of the plan it finds that all the predictions made by the planner are confirmed However the next time it examines the container it finds that the amounts of the two solutions have changed. This change was not predicted by the planner.

The explanation module tries to relate the change to an effect of one of the known processes for changing the amount of a solution. The relevant processes are evaporation, condensation, absorption (in which the solid absorbs liquid like a sponge), release (in which the solid releases the absorbed liquid), and flow. The system tries to explain the observations as a result of these processes by activating the corresponding schemata and checking if the change caused by running each process is compatible with the observations. However, it finds that none of the processes can be activated since each of them has preconditions which cannot be satisfied. Evaporation and condensation require exposure to the atmosphere and since the compartments are closed by lids they cannot be activated. Absorption and release require the container to have an absorbent property. Flow requires a free path from the source to the destination. These requirements are not met in the present situation. So the system is left with a contradiction in which it has observations for which its current world model has proven inadequate (Figure 1).

D. Beliefs Tested

The explanation structure that resulted in the contradiction involves a number of processes. Each process seems inapplicable because one or more preconditions are not satisfied in the current situation. The beliefs on which the contradiction rests are obtained from the reasons for the failure of each process to run. The beliefs are :

(1) The procedure for classifying solids into absorbent and non-absorbent classes is right,

(2) Liquids require a clear, solid-free path to flow, and

(3) Evaporation and condensation require exposure to the atmosphere.

For the observation to be valid one of these beliefs must be wrong and to find out which one the system performs a series of experiments. Note that on the first attempt at finding an explanation the system tests only the immediate causes of failure of the primary processes.

E. The Experiment Designer

There are two stages of experimentation:

(1) The first stage involves distinguishing among the five processes that would explain the contradiction,

(2) The second stage involves constructing a series of experiments which are used to generalize the specific instance of the process observed.

The partially instantiated processes, the failed preconditions, and the unexplained observations form the input to the experiment design module. The ideal experiment would reproduce the setup in the original observation in such a manner that only one of the processes is active and all the other competing processes are eliminated totally. However, it is impossible to eliminate the possibility of some processes (for example, absorption, since one cannot build a container without walls). Furthermore, this may be undesirable because the process, which is itself suspect, may continue to influence the observations in some hidden manner. For example, one might think that flow can be eliminated by separating the two compartments and moving one away from the other, but the belief being questioned, namely, liquids cannot flow through solids, is still in effect; this time with a longer path through more solids like tables and floors.

Some means of discriminating all the instantiations of the five processes simultaneously is required Note that the basis for distinguishing the processes cannot depend on the preconditions because they are suspected of not activating processes when they should have. One characteristic of a process is that the time rate at which the process progresses depends on geometrical parameters (length, area etc.), state variables (temperature, pressure etc), and the properties of the participants. The system bases its experiments upon the rate at which each process progresses. Note that a secondary belief, which forms the basis for the investigation, is that the proportionalities of the rates on the above parameters do not change.

Experiments are designed such that one of the processes is allowed to dominate the rest by having an environment in which its rate is enhanced and the other competing processes' rates inhibited. If the original observations are reproduced in a much shorter time period then the evidence points to the dominating process as the cause of the observation. The flow process rate depends on the

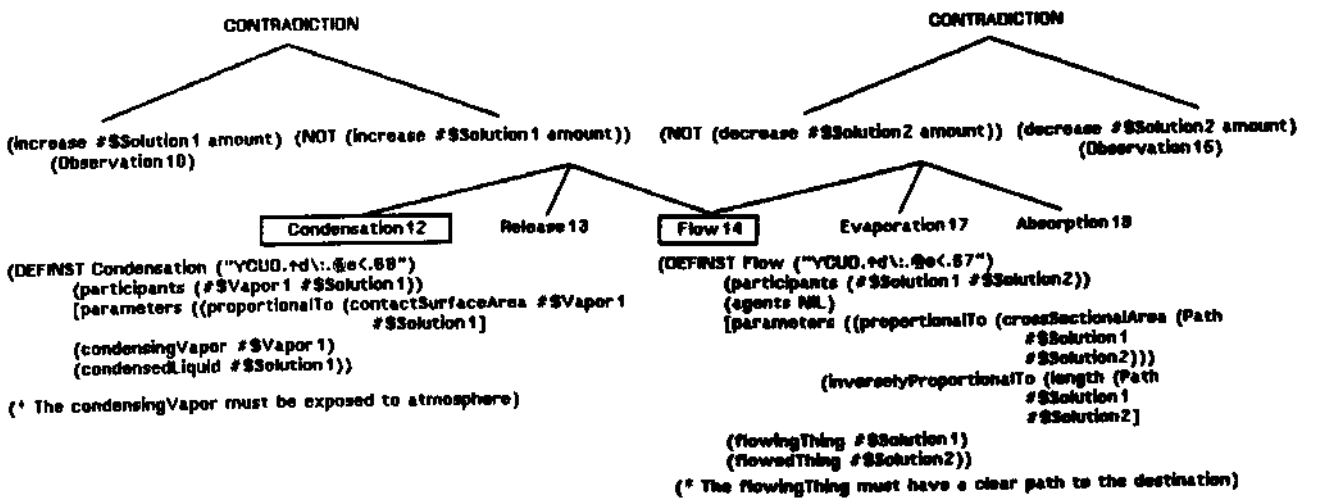

Figure 1. The contradiction diagram, a brief description of two processes and the preconditions that failed.

cross-sectional area and the length of the path from the source to the destination, the evaporation and condensation rates depend on the surface area of contact between the liquid and its vapor, and the absorption and release rates depend on the surface area of contact between the absorbing solid and the absorbed liquid. By manipulating the geometry of the containers it is possible to build containers in which everything else remains as in the original setup, but parameters like contact surface area, cross-sectional area and length are maximized or minimized to allow one process to dominate the other competing ones.

In our particular example the system comes up with experiment specifications to distinguish each process. The set of specifications to distinguish the process flow is shown in Figure 2 and an experiment that meets these specifications relative to the original setup is shown in Figure 3. The system requests that the experiments be carried out and the results returned. Based on the combined results the system concludes that some form of the process flow caused the original observation.

```
#$Experiment 112
Process being tested
  Flow 14
Specifications
((maximize (crossSectionalArea (Path
                        #$Solution 1
                        #$Solution 2)))
 (minimize  (length (Path #$Solution 1
                    #$Solution 2)))
 (minimize (contactSurfaceArea #$Vapor 2
                        #$Solution 2))
 (minimize (contactSurfaceArea #$Container 3
                        #$Solution 2))
 (minimize (contactSurfaceArea #$Vapor 1
                        #$Solution 1))
 (minimize (contactSurfaceArea #$Container 2
                        #$Solution 1)))
Competing Processes
(Condensation 12 Release 13 Evaporation 17
                              Absorption 18)
```

Figure 2. Experiment specifications to distinguish flow

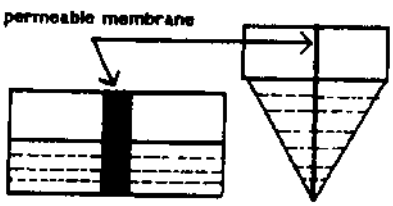

Figure 3. The original setup and an experiment designed from the specifications in Figure 2.

F. Revising the World Model

The most specific revision of the world model would be to create a new process that resembles the process flow in most respects. However this process will have preconditions that permit flow when the two specific solutions are separated by the specific partition as in the example above. Our current efforts are concentrated on generalizing this specific case using explanation-based learning [DeJong85] in conjunction with a further series of experiments aimed at discovering the properties of the participants that played a crucial role in the example. In any case, the important thing to note is that the system no longer believes that liquids cannot pass through solids and has a specific example to disprove it.

HI CONCLUSIONS

We have explored a model for making experiments to explain inconsistencies between the system's knowledge and the real world. In some respects, this is similar to work described in [Langley8l] but our approach uses the world model to drive the experimentation. Also, some results on testing hypotheses using experiments have been reported in [Shapiro8l]. Shapiro's approach is, however, of a more theoretical nature, and is impractical because it uses a large number of experiments.

The project is still in its infancy. The next step is to improve the experiment design, where the project addresses serious philosophical issues about knowledge and the closed world hypothesis. Our approach to this so far has been to copy people's behavior and try the simple explanations first. Only if these fail are the more basic assumptions that underlie them retracted.

REFERENCES

[Carbonell82] J. Carbonell, "Experiential learning in Analogical Problem Solving," *Proceedings of the National Conference on Artificial Intelligence,* Pittsburgh, PA, 1982. 168-171.

[DeJong82] G. DeJong, "Automatic Schema Acquisition in a Natural Language Lnvironment," *Proceedings of the National Conference on Artificial Intelligence,* Pittsburgh, PA, 1982, 410-413.

[DeJong85] G. DeJong, B. Faltings, R. Mooney, P. O'Rorke, S. Rajamoney, A. M. Segre and J. Shavlik, "A Review of Lxplanation-Based Learning," Technical Report in preparation, Coordinated Science Laboratory, Urbana, IL, 1985.

[Langley8l] P. Langley, "Data-driven Discovery of Physical Laws," *Cognitive Science* 5, (1981), 31-54.

[Michalski83] R. Michalski and R. Stepp, "Learning from Observation: Conceptual Clustering," in *MacHne Learning,* Ryszard Michalski, Jaime Carbonell, Tom Mitchell (ed.), Tioga Publishing Company, Palo Alto, 1983, 331-363.

[Mitchell83] T. Mitchell, "Learning and Problem Solving," *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* Karlsruhe, West Germany, 1983,1139-1151.

[Schank82] R. Schank, *Dynamic Memory,* Cambridge University Press, Cambridge, 1982.

[ShapiroSl] E Y. Shapiro, "Inductive Inference of Theories from Facts," Research Report 192, Yale University, Yale, February, 1981.

[Winston83] P. H. Winston, T. 0. Binford and M. Lowry, "Learning Physical Descriptions from Functional Definitions, Examples, and Precedents," Artificial Intelligence Laboratory Memo No. 679, M.I.T., Cambridge, MA, January 1983.