The Use of Explanations for Completing and Correcting Causal Models[1]

Joel D. Martin and Michael Redmond
Georgia Institute of Technology
E-mail: joel@gatech.edu, redmond@gatech.edu

Abstract

Causal models describe some part of the world to allow an information system to perform complex tasks such as diagnosis. However, as many researchers have discovered, such models are rarely complete or consistent. As well, the world may change slightly, making a previously complete model incomplete. A computational theory of the use of causal models must allow for completion and correction in the face of new evidence. This paper discusses these issues with respect to the evolution of a causal model in a diagnosis task. The reasoner's goal is to diagnose a fault in a malfunctioning automobile, and it improves its diagnostic model by comparing it with an instructor's. A general process model is presented with two implementations. Related work in explanation based learning and in incorrect causal models is discussed.

Keywords: Learning, Causal Models, Explanations, Diagnosis

INTRODUCTION

A causal model or domain theory is an essential ingredient in understanding complex situations. For example, in order to diagnose a fault in a complex system, the diagnostician must be capable of making guesses about what might be wrong. However, without appropriate heuristic knowledge to guide and make those guesses, the correct hypothesis may never arise. Although many researchers have acknowledged the need for such causal models [Kuipers, 1984] [deKleer & Brown, 1981], very few have been concerned with the possibility that the domain theory may be incomplete or inconsistent. Those researchers who have recognized this problem [Rajamoney & DeJong, 1987] have not yet allowed for modification of the underlying causal theory.

With this in mind, Lancaster and Kolodner [1987] took protocols of the diagnostic behavior of novice, intermediate, advanced, and expert car mechanics. They observed evidence for a working model, a set of symptom fault pairings, and diagnostic strategies. They also observed [Lancaster, personal communication] that less experienced mechanics had inconsistent and incomplete knowledge, as one might expect. The research presented in this paper represents an effort to discover how an incomplete causal model (novice) can evolve to a more complete (experienced) state as a result of problem solving experience coupled with explanations about how those problems are solved.

Our model is implemented in two computer programs called EDSEL-1 and EDSEL-2 (Explanation in Diagnosis: the use of Symptoms, hypotheses, and Explanations for Learning) that each begin with a novice memory and are presented with problems and explanations of how to solve those problems. Specifically, the systems "watch" or attend to an instructor who is diagnosing a fault in an automobile. As they do so, they attempt

[1]This research was supported by the Army Research Institute for the Behavioral and Social Sciences under Contract No. MDA-903-86-C-173. The authors wish to thank Janet Kolodner for her advice and guidance, and Mark Graves and Hong Shinn for helpful comments on earlier versions of the paper.

to identify missing information of various types or to identify whether there is an inconsistency. If one of these problems is discovered, the systems modify their causal model to prevent the difficulty in the future. The recognition and modifications are based upon an attempt by the systems to explain their input.

The paper describes a general algorithm, presents the issues involved in completing and correcting causal models, and compares the two implementations.

GENERAL PROCESS

Completing and correcting a causal model requires a reasoner to recognize when it is missing a piece of information and then to incorporate that information into the model. This notion is complicated by the fact that there are different types of information in the model and that existing knowledge affects how new information is incorporated.

Redmond and Martin [1988] noted in the protocols from Lancaster and Kolodner [1987] that an instructor provides the students with a symptom, a series of hypotheses, and explanations for those hypotheses. We demonstrated that a system may process these inputs by attempting to build causal chains between hypotheses and the symptom, using provided explanations if no causal chain is obvious. If a complete chain can be built, it can be collapsed and be used more efficiently in future similar situations. If a complete chain cannot be built, then the instructor's explanation can be helpful, either by being added directly to the causal model, or by allowing the gap in the chain to be bridged. Our name for this process is Learning by Understanding Explanations (LBUE). The causal model contains frames [Minsky, 1975] for the components of a car, with slots for inputs, outputs, connections, parts, functions, and causal relationships between structures. For example, one piece of the current model is:

```
Starter:
   (isa component)                              ;A starter is a component.
   (part-of starting-system)                    ;Is a part of the starting system.
   (input electricity battery battery-cables)   ;Electricity from battery via cables.
   (parts starter-pinnion-gear starter-motor)   ;PARTS: pinnion gear and starter motor.
   (function spin-action starter-pinnion-gear) ;FUNCTION: spin the pinnion gear.
   (cause (switch-action solenoid on)           ;Solenoid switch causes the two gears to interlock.
       (interlock starter-pinion-gear flywheel-ring-gear))
   (cause (crank starter-pinion-gear)           ;Cranking one gear causes the other to crank.
       (crank flywheel-ring-gear))
```

The causal chaining process uses the causal relationships and some of the related knowledge. Besides the causal model of the domain, the LBUE approach also includes a set of likely symptom-fault pairings, as observed by Lancaster and Kolodner [1987] . These sets of pairings associate a symptom with a problem, and are used to derive initial hypotheses during diagnosis, and to index into the causal model at the appropriate place. The general algorithm for the process is as follows:

1. From the symptom (presented by the instructor), chain backward, inferring possible findings that could lead to the symptom.
2. From each hypothesis (presented by the instructor), chain forward, inferring possible effects that could be caused by the hypothesized fault.
3. If the symptom chain meets a hypothesis chain, then the reasoner has an explanation for the hypothesis, and the generalization that (CAUSE HYPOTHESIS SYMPTOM) is added to the symptom fault table and to the causal model.

4. If the chains do not meet - the reasoner does not have enough information to explain the hypothesis. In this case, it uses an explanation presented by the instructor,
   (a) Chain backwards from the explanation toward the hypotheses chain.
   (b) Chain forward from the explanation toward the symptom chain.
   (c) If both directions can be linked, then the most general relationship (CAUSE HYPOTHESIS SYMPTOM) can be learned.
5. Add explanation to the causal model.

The LBUE process results in an updated causal model. As discussed in following section, the things that may be learned are,

1. new objects
2. new relationships between objects
3. new causal information
4. new symptom fault knowledge

After learning, diagnosis is more efficient and more powerful for the same or similar problems, because the symptom-fault set can provide more reasonable hypotheses more quickly and the causal model is more capable of verifying an explanation. In addition, what the reasoner learns depends on what it already knows, since the reasoner's ability to chain back from the symptom and forward from the hypothesis is affected by the knowledge in the causal model. This means that the chains could meet given one version of the causal model, and have an unbridgable gap given another version.

The alternative to the LBUE approach is simply to remember symptom-hypothesis pairs. However, this would require a system to have already experienced a fault in order to diagnose it; no general knowledge is retained.

ISSUES FOR COMPLETING AND CORRECTING CAUSAL MODELS

As outlined above, a good diagnostic reasoner tries to explain why an hypothesis causes a symptom. It is this process that allows for the recognition of different types of missing information, and mediates the addition of knowledge to the causal model. The process of explaining hypotheses identifies missing information that might be useful for diagnosis because diagnosis is itself explanation, and hence requires the same information.

TYPES OF MISSING KNOWLEDGE

In general, a causal model may be missing many causal relations necessary for diagnosis. A reasoner will recognize that a causal relationship is missing if an explanation of a symptom cannot be formed, either while watching an instructor or while doing diagnosis. As well, there are situations in which an unknown causal relationship will be presented to the reasoner. Both possibilities are simple to detect, the former when causal chaining fails or no reasonable hypothesis is generated, and the latter, when the reasoner is actually told that something is missing.

Another form of knowledge whose absence is easily detected consists of referred-to facts. In other words, when an object or general relationship between objects is asserted, but is not known, then it is missing from the model. Somewhat more interesting are implied facts. The reasoner guesses it is missing an implied fact when a causal relationship is stated or implied by an instructor that the reasoner believes requires a mediating fact. For example, a reasoner may know,

```
(INTERLOCKED gear1 gear2) & (SPIN gear1 'clockwise) --> (SPIN gear2 'c-clockwise)
```

and an instructor may state,

```
(SPIN starter-gear 'clockwise) --> (SPIN flywheel-ring-gear 'c-clockwise)
```

From this, the reasoner will recognize that it is missing a fact (i.e., that the two gears are interlocked).

The final type of information that a reasoner may be missing is essentially efficiency information. The reasoner must be able to arrive at a reasonable or correct hypothesis quickly. If it cannot, the causal model must be modified to ensure timely and correct diagnoses in the future. The reasoner can recognize that it is missing this kind of information if it arrives at an incorrect hypothesis during diagnosis or if its hypotheses differ from the instructor's.

METHODS OF HANDLING INCOMPLETE KNOWLEDGE

An instructor's explanation of a given hypothesis can lead to information being added in three different ways. The explanation itself could be an unknown causal relationship which can be added to the model directly. For example, if the instructor explained

```
(cause (corroded battery-terminals) (not (connect battery battery-terminals)))
```

and this relationship was not associated with either battery or battery-terminals in the causal model, then it can be added there. A second way that the instructor's explanation can be used is to enable filling a gap in a causal chain. Either the explanation filled the gap, or it was a better cue to information that was not being accessed in the causal model. If the causal chain that can be built from the symptom (NOT (RUN ENGINE)) is:

```
(not (run engine)) -->  (not (spin crankshaft)) -->
    (not (down-stroke cylinder)) --> (not (combustion cylinder))
```

and the causal chain that can be built from the associated hypothesis (NOT (MOVABLE BUTTERFLY-VALVE)) is:

```
(not (movable butterfly-valve)) --> (flow air carburetor low)
```

then there is a gap in the causal chain — the hypothesis is not fully explained. If the instructor provides the explanation that low air flow into the carburetor leads to a low air/gas mixture as the air passes the fuel float bowl then the following results:

```
(not (movable butterfly-valve)) --> (flow air carburetor low) -->
  (mix air gas less)  -->  (not (combustion cylinder)) -->
     (not (down-stroke cylinder)) --> (not (spin crankshaft)) --> (not (run engine))
```

Not only is the causal relationship given in the explanation used in filling the gap, but the relationship that (MIX AIR GAS LESS) causes (NOT (COMBUSTION CYLINDER)) is accessible when it hadn't previously been accessible, since the cue of carburetor is now available. Between the two, the gap has been filled.

The third way in which the instructor's explanation can be used is to infer a relationship that would fill a gap in a causal chain. If the explanation doesn't allow bridging the gap as discussed above, causal relationships which bridge the gap, which are implied by the expert instructor, can be inferred. The instructor implies that there is a causal relationship

between the hypothesis and symptom and that the explanation lies along this causal chain. Gaps will be filled with inferences if there is some general knowledge that indicates a cause is possible. For example, a cracked wire can cause low electricity because (a) wires conduct electricity, and (b) a conduit affects what it conducts. This would be given lower credibility than other learned relationships. There may be several plausible but inconsistent inferences that might fill a gap; the one chosen could depend on confirmation from a human observer.

Knowledge can be added to an incomplete causal model by inferring facts from a cause. This would occur as a result of the starter-gear ring-gear example mentioned above. In this case, the reasoner will infer that the starter gear and ring gear are interlocked.

In a sense, inefficiently represented knowledge is a type of incomplete knowledge. The information that is needed is in the causal model, but is not useful because it cannot be accessed, or it is given insufficient credibility, or it leads to slow processing. For instance, the explanation can allow the access of knowledge that couldn't previously be accessed. Additionally, filling a gap in a causal chain, as discussed above, is a way of dealing with some inefficient knowledge. This allows collapsing the chain into a single causal relationship, which can be used for more efficient processing. In collapsing the chains, the LBUE method has some similarities to Explanation-Based Learning (EBL) [Mitchell, Kellar, & Kedar-Cabelli, 1986] [DeJong & Mooney, 1986]. In order to allow for proper generalization of variables [DeJong & Mooney, 1986], a substitution list is kept that indicates to what categories each feature in the example was matched in order to instantiate the causal relationships. The collapsed chain then uses the most general category for a feature as the variable name in the antecedent or consequent of the new causal relationship.

INCONSISTENT KNOWLEDGE

Since new information is being added to the causal model, there is a possibility that a contradiction may occur. A few types of contradiction are possible. In one case, the same condition could be believed to cause contradictory effects such as:

```
   (corroded battery-terminals) --> (connect battery battery-terminals)
& (corroded battery-terminals) --> (not (connect battery battery-terminals))
```

Alternatively, a chain might be possible from known information, such that a condition indirectly causes a contradiction of the condition.

Contradictions may not be detected immediately, though, because the causal information is distributed throughout the causal model, and because an arbitrary amount of causal chaining may be necessary to detect the contradiction.

In cases where the new input is found to be inconsistent, the source of the information can be used to decide what to believe. Knowledge from the expert is given precedence over older information. Information that contradicts the expert is either removed or its strength is decreased, depending upon implementation.

IMPLEMENTATIONS

Implementation of the general model described above has followed two parallel paths. This decision was made because the research is exploratory, and therefore should generate several alternative approaches to the problem, and highlight different inconsistencies and difficulties with the model.

EDSEL-1 is based upon a simple active semantic net, similar to local connectionist models such as McClelland and Rumelhart's [1981] interactive activation model. EDSEL-2 uses an enhanced version of the causal model described in Allison [1987] . Although the process in both implementations closely follows the general model presented above, there are two significant differences. First, when a *causal gap* is present but no explanation fills that gap, EDSEL-1 uses generic knowledge about what affects what, whereas EDSEL-2 uses a less general but far simpler notion of filling gaps between recently proposed forward chains from hypotheses and backward chains from the symptom. The first method is a more flexible metric for evaluating whether a given *causal gap* should be filled, and therefore should lead to more reliable causal relationships. The second difference involves where causal information is stored and how it can be accessed. EDSEL-1 does not address the issue of limited availability of causal relationships, whereas EDSEL-2 allows the more realistic situation in which causes are not maximally indexed when they enter the system. That is, they may not necessarily be retrieved when needed unless the proper cues are present. This is a more realistic and efficient approach for a system with a very large memory.

RELATED WORK

Rajamoney and DeJong [1987] has specifically addressed the problem of inconsistencies or missing information in a causal model for simulation. If more than one simulation is possible, his system will experimentally search for disambiguating features in the environment. Although this approach is clearly useful, it does not allow for modification of the general causal information in the model. It concentrates on quantitative values for the current situation, and does not learn any general knowledge.

As has already been noted, in order to update causal models, the current effort uses an explanation based technique that is in some ways similar to those of DeJong and Mooney [1986] and Mitchell et al. [1986] . Specifically, in diagnosis, a causal chain must be discovered in a potentially very large network of causal information. EBL can be profitably used to permit instruction to produce "short cuts" in that network.

Classical EBL, however, does not produce enough learning when the causal network is incomplete. This may be remedied by the learning by failing to explain (LBFE) [Hall, 1986] technique of isolating the information that is present in the input but is not understood, and subsequently adding it to the existing EBL system. Although the current model has not yet been described in exactly these terms, it is in fact an example of LBFE. It differs from Hall's work by proposing that the information that must be added in the absence of an explanation is not necessarily explicitly represented in the input. Also, the current effort presents a domain independent notion of LBFE that describes how potentially any diagnostic causal net might grow, whereas Hall's effort was, in his own view, domain specific.

One of the methods that is used to augment incomplete networks in the current approach is to use relationships that are more general than causes in order to infer causation. For example, an action and a state change that relate to the same object tend to be causally related. This technique was originally used by Pazzani [1987] and a similar approach was suggested by Russell [1987] .

CONCLUSIONS AND FUTURE DIRECTIONS

This paper has discussed the LBUE paradigm for dealing with an incomplete or inconsistent causal model in the training of car mechanics. The main contribution of the current effort is in the ability to accept new knowledge and incorporate it into the causal model, while using it to understand an explanation and form a new generalization.

There are several directions for future research. First, in diagnosis, causal chaining is not the only strategy used, though it was the most common in the protocols. The explanations used by the instructor reflect several different strategies. For this reason, and to allow strategies to be learned and improved, diagnostic strategies must be explicitly represented. Some initial work has been done on this representation.

Second, better representation of the causal knowledge is needed to take full advantage of the inferencing possible from qualitative models. The aim is to use more levels of abstraction to allow reasoning at whatever level may be appropriate. Third, more learning may be possible in this paradigm if Case-Based Reasoning [Kolodner & Simpson, 1984], a method of using previous episodes and evaluation of their results to suggest solutions to new problems, could be integrated.

REFERENCES

Allison, K. R. (1987). Use of a working model in fault diagnosis. In *Proceedings of the 25th Annual Conference of the Southeast Region ACM.*

DeJong, G. & Mooney, R. (1986). Explanation based learning: an alternative view. *Machine Learning, 1,* 145–176.

de Kleer, J. & Brown, J. S. (1981). Mental models of physical mechanisms and their acquisition. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

Hall, R. (1986). Learning by failing to explain. In *Proceedings of the National Conference on Artificial Intelligence.*

Kolodner, J. & Simpson, R. (1984). A case for case-based reasoning. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society.*

Kuipers, J. (1984). Commonsense reasoning about causality: deriving behavior from structure. *Artificial Intelligence, 24,* 169–203.

Lancaster, J. & Kolodner, J. (1987). Problem solving in a natural task as a function of experience. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.*

Lancaster, J. (personal communication). August, 1987.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part 1. an account of basic findings. *Psychological Review, 88,* 375–407.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision.* New York: McGraw-Hill.

Mitchell, T. M., Kellar, R. M., & Kedar-Cabelli, S. T. (1986). Explanation based learning: an unifying view. *Machine Learning, 1,* 47–80.

Pazzani, M. (1987). Inducing causal and social theories: a prerequisite for explanation-based learning. In *Proceedings of the Fourth Annual International Workshop on Machine Learning.*

Rajamoney, S. A. & DeJong, G. F. (1987). *Active ambiguity reduction: An experimental design approach to tractable qualitative reasoning.* Technical Report UILU-ENG-87-2225, University of Illinois at Urbana-Champaign.

Redmond, M. & Martin, J. (1988). Learning by understanding explanations. In *Proceedings of the 26th Annual Conference of the Southeast Region ACM.*

Russell, S. J. (1987). Analogy and single-instance generalization. In *Proceedings of the Fourth Annual International Workshop on Machine Learning.*