Appears in *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning,* pp. 499-508, Cambridge, MA, October, 1992

Abductive Plan Recognition and Diagnosis: A Comprehensive Empirical Evaluation

Hwee Tou Ng and **Raymond J. Mooney**
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712
htng@cs.utexas.edu, mooney@cs.utexas.edu

Abstract

While it has been realized for quite some time within AI that abduction is a general model of explanation for a variety of tasks, there have been no empirical investigations into the practical feasibility of a general, logic-based abductive approach to explanation. In this paper we present extensive empirical results on applying a general abductive system, ACCEL, to moderately complex problems in plan recognition and diagnosis. In plan recognition, ACCEL has been tested on 50 short narrative texts, inferring characters' plans from actions described in a text. In medical diagnosis, ACCEL has diagnosed 50 real-world patient cases involving brain damage due to stroke (previously addressed by set-covering methods). ACCEL also uses abduction to accomplish model-based diagnosis of logic circuits (a full adder) and continuous dynamic systems (a temperature controller and the water balance system of the human kidney). The results indicate that general purpose abduction is an effective and efficient mechanism for solving problems in plan recognition and diagnosis.

1 INTRODUCTION

Finding explanations for events and actions is an important aspect of general intelligent behavior. A diverse set of intelligent activities, including natural language understanding, diagnosis, scientific theory formation, and image interpretation, requires the ability to construct explanations for observed phenomena. For instance, in text understanding, a reader infers the high-level goals and plans of the characters in a text in order to explain the events and actions described in the text. Such inference is called *plan recognition*, which is known to be an important component of text understanding [Allen, 1987]. Similarly, in medical diagnosis, based on the observed symptoms of a patient, a physician infers the possible diseases that may explain the symptoms. In physical device diagnosis, based on the observed misbehavior of a physical device, a diagnostician infers the possible faults that may explain the misbehavior.

In this paper, we view explanation as *abduction*. The standard logical definition of abduction is: given a set of axioms T (the domain theory) and a conjunction of atoms O (the observations), find minimal sets of atoms A (the assumptions) such that $A \cup T \models O$ and $A \cup T$ is consistent [Charniak and McDermott, 1985; Levesque, 1989].[1]

While it has been realized for quite some time within AI that abduction is a general model for explanation [Charniak and McDermott, 1985], there have been no empirical investigations into the practical feasibility of such a general abductive approach to explanation. Many important questions remain unexplored. For example, is it possible to have a general-purpose yet efficient algorithm that can be used for making useful abductive inference and solving moderately complex problems in reasonable time in all the various domains? Do we need special-purpose control heuristics separately tailored for each domain to achieve efficient abduction? Do the criteria for selecting the best explanations vary according to the domain? How difficult is it to encode the knowledge necessary for constructing explanations in the various domains?

This paper attempts to address these important issues. Towards this end, we have built a domain-independent system called ACCEL (Abductive Construction of Causal Explanations in Logic). In our system, knowledge about a variety of domains is uniformly encoded in first-order Horn-clause axioms. A general-purpose abduction algorithm, AAA (ATMS-based Abduction Algorithm), efficiently constructs explanations in these

[1] In this paper, the terms "abduction" and "abductive" are used to refer to this specific logical formulation as opposed to the more general notion of any method for inferring cause from effect.

domains. We have applied our abductive system to two general tasks: plan recognition in text understanding, and diagnosis of medical diseases, logic circuits, and dynamic systems. In plan recognition, Accel has been tested on 50 narrative texts, where each text consists of 1–4 short sentences. The system infers the high-level plans of characters based on the actions described in a text. In medical diagnosis, Accel diagnoses 50 real-world patient cases using a sizable knowledge base with over six hundred symptom-disease rules. The disorders are damaged regions of a human brain due to stroke. These cases have been previously used to test set covering methods [Tuhrim *et al.*, 1991; Peng and Reggia, 1990]. Accel also achieves model-based diagnosis via abduction, and has successfully diagnosed a full adder (an example of discrete, combinational logic circuits), a temperature controller, and the water balance system of the human kidney (examples of continuous dynamic systems).

The rest of this paper is organized as follows. Section 2 gives a brief overview of Accel. Section 3, 4, and 5 present empirical results for plan recognition, set-covering-based diagnosis, and model-based diagnosis, respectively. Section 6 discusses the results and presents the conclusions.

2 OVERVIEW OF ACCEL

Given an existentially quantified conjunction of atoms that encodes the input observations, and a set of first-order Horn clauses that encodes the domain theory, the algorithm AAA computes abductive explanations by backward-chaining on the observations, much like Prolog. However, even when there is no fact or consequent of a rule that unifies with a subgoal in the current proof attempt, instead of failing, the algorithm has the choice of making the subgoal an assumption if it is consistent to do so. The requirement for consistency means that abduction is in general undecidable. In Accel, inconsistency is detected using a predetermined list of *nogoods*, and by procedural code (for efficiency reasons). A nogood is a set of assumptions that implies falsity, and consistency checking ensures that an assumed set of atoms is not subsumed by any nogoods.

Accel can be used to compute all minimal sets of abductive assumptions; however, minimality is generally too unrestrictive and even in the propositional case, the number of minimal explanations can grow exponentially [Selman and Levesque, 1990]. Consequently, beam-search is generally used to limit computation to a fixed-sized subset of the currently best partial explanations according to a user-defined evaluation function. In a further attempt to improve efficiency, Accel performs ATMS-like caching of partial explanations. Empirical results have shown that caching can achieve more than an order of magnitude speedup in run time. More details on Accel and the AAA algorithm are given in [Ng, 1992]. (A previous version of Accel is described in [Ng and Mooney, 1991].)

3 ABDUCTIVE PLAN RECOGNITION

Given a logical representation of the literal meaning of a narrative text in terms of an existentially quantified conjunction of input atoms, Accel infers an "embellished" interpretation by constructing an abductive proof in which a set of higher-level plans is assumed and the assumed plans logically entail the characters' observed actions. An abductive proof is considered an interpretation of the input sentences. We do not focus on the parsing aspect of natural language understanding, and Accel does not accept natural language input.

Examples of the 50 narrative texts processed by Accel include: "Bill went to the liquor-store. He pointed a gun at the owner."; "Bill took a bus to a restaurant. He drank a milkshake. He pointed a gun at the owner. He got some money from him."; "Fred got a gun. He went to the restaurant. He packed a suitcase."; etc. The knowledge base axioms are formulated such that higher-level plans (like shopping and robbing) together with appropriate role-filler assumptions (like someone is the shopper of a shopping plan or the robber of a robbing plan) imply the input atoms representing the observed actions (like going to a store and pointing a gun).

In the plan recognition domain, explanations are evaluated by a criterion called *explanatory coherence* [Ng and Mooney, 1990; Ng, 1992]. This criterion attempts to capture the notion that natural language text is coherently structured and therefore explanations that better "tie together" various parts of the input sentences are to be preferred. Hence, evaluating explanations based on explanatory coherence takes into account the well known "Grice's conversational maxims" [Grice, 1975], which are principles governing the production of natural language utterances, such as "be relevant", "be informative", etc.

Specifically, the coherence metric C is defined as follows:

$$\mathcal{C} = \frac{\displaystyle\sum_{1 \le i < j \le l} N_{i,j}}{l(l-1)/2}$$

where l = the total number of input atoms; and $N_{i,j}$ = 1 if there is some node n in the proof graph such that there is a (possibly empty) sequence of directed edges from n to n_i and a (possibly empty) sequence of directed edges from n to n_j, where n_i and n_j are input atoms. Otherwise, $N_{i,j} = 0$.[2] More details of

[2] The definition of coherence given in this paper is a

the coherence metric are described in [Ng and Mooney, 1990; Ng, 1992].

In the domain of plan recognition, we have tested Accel on 50 short narrative texts. To facilitate comparison between different approaches, the first 25 texts were taken from Goldman's PhD thesis [Goldman, 1990], where they were used to test a probabilistic approach to text understanding. An additional set of 25 similar narrative texts were created by Ray Mooney unbeknown to the system developer (Hwee Tou Ng). The intent is that the additional 25 examples will test for other novel combinations and sequences of actions that the knowledge base constructed for the initial 25 examples in principle should be able to handle. We will call the first set of 25 examples the *training* examples, and the second set of 25 examples the *test* examples.

The plans in the knowledge base include shopping, robbing, restaurant dining, traveling in a vehicle (bus, taxi, or plane), partying, and jogging. Each of these plans in turn has subplans, and some of the plans contain recursive subplans. For instance, traveling by plane includes the subplan of traveling (in some vehicle) to the airport to catch a plane. For each example, a set of input atoms representing the sentences is given to Accel. To give a sense of the size of our examples and the knowledge base used, there is a total of 107 KB rules and 70 taxonomy-sort symbols. Every taxonomy-sort symbol p will add an axiom (in addition to the 107 KB rules) of the form $inst(X, p) \rightarrow inst(X, supersort\text{-}of\text{-}p)$. The average number and maximum number of input atoms per example are 12.6 and 26 respectively. The knowledge base and the 50 examples are included in [Ng, 1992].

For each example, the correct explanation was determined based on the authors' intuition before running the example. To measure the quality of an explanation computed by Accel, we compared it to the correct explanation and recorded three error rates: the recall error rate R = the number of missing assumptions divided by the number of assumptions in the correct explanation, the precision error rate P = the number of excess assumptions divided by the number of assumptions in the computed explanation, and the overall error rate O = the average of the recall and precision error rates. (We used similar quality measures and terminology as in [Lehnert and Sundheim, 1991].) If more than one best explanations are computed for an example, we take the error rates for the example to be the average of the error rates over all the best explanations.

We ran Accel on the 50 examples using two different evaluation metrics: the coherence metric (breaking ties based on simplicity, defined as the number of assumptions made in an explanation) and the simplicity metric. The empirical results are summarized in Table 1, which shows the average recall (R), precision (P), and overall (O) error rates for the training examples, test examples, and all examples. The average run time per example is 1.83 minutes on a Sun Sparc 2 workstation.

Table 1: Empirical Results Comparing Coherence and Simplicity.

Example	Coherence			Simplicity		
type	R	P	O	R	P	O
Training	0.2%	0%	0.1%	26%	25%	25%
Test	2%	2%	2%	39%	38%	38%
All	1.1%	1%	1%	32%	31%	32%

The empirical results demonstrate that Accel can efficiently process these narrative texts, and it is sufficiently general to be able to handle similar plan recognition problems not known to the system developer in advance. Furthermore, coherence consistently performs better than simplicity on the examples tested.

Even though our knowledge base does not contain any probabilistic or likelihood information, the results on the training examples achieved by Accel are the same as those of Goldman' system which uses a probabilistic approach to language understanding. In the probabilistic approach, the primary purpose of *a priori* probabilities is to select a most likely explanation when there are otherwise multiple competing explanations. In Accel, we accomplish an analogous effect by writing axioms that only explain some input atom in terms of a high-level plan but not the other competing plans. More details are given in [Ng, 1992].

4 DIAGNOSIS BASED ON SET COVERING

Over the past decade, Reggia and his colleagues have developed an increasingly sophisticated theory of diagnosis based on set covering and applied the theory primarily to medical disease diagnosis [Peng and Reggia, 1990]. The basic diagnostic problem in the Generalized Set Covering (GSC) model is defined by four sets, (D, M, C, M^+), where D is a finite set of potential disorders, M is a finite set of potential manifestations (symptoms), $C \subseteq D \times M$ is a causation relation where $(d, m) \in C$ means "d may cause m", and $M^+ \subseteq M$ is the set of observed manifestations for the current case. $E \subseteq D$ is called a *cover* of M^+ iff for each $m \in M^+$ there exists $d \in E$ such that $(d, m) \in C$. A cover is said to be *minimum* if its cardinality is the smallest among all covers and *irredundant* (*minimal*) if none of its proper subsets is also a cover. Depending on the

slight modification of the one given in [Ng and Mooney, 1990]. The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths.

domain, one may consider all minimum or all minimal covers of the observed symptoms as the best diagnoses.

We can map a GSC diagnostic problem into an abduction problem in Accel as follows: Let the domain theory T be the set of axioms $\{d \rightarrow m | (d, m) \in C\}$, and let the input atoms $O = \bigwedge_{m \in M^+} m$. Suppose only atoms $d \in D$ are assumable (i.e., we use predicate specific abduction [Stickel, 1988] in which only atoms with certain predicates are assumable). It can be easily proved that the set of covers of GSC is the same as the set of explanations in Accel [Ng, 1992].[3]

The logical abduction approach, being based on a more expressive representation language, can accommodate more naturally "causal chaining" [Peng and Reggia, 1990], incompatible disorders, and symptoms caused by combinations of disorders. As GSC diagnostic problems can be nicely represented as abduction problems, the remaining question is whether a general logic-based abductive system can solve such problems efficiently. Further, because the GSC diagnostic problem is NP-hard [Reggia *et al.*, 1985], the issue then becomes whether a logical abductive system can solve real problems in reasonable time and is competitive with existing set-covering algorithms. To address this issue, we tested Accel on the medical problem studied in [Tuhrim *et al.*, 1991], which specifies 25 brain areas (e.g. right frontal lobe) whose damage can explain 37 basic symptom types (e.g. impaired gag reflex). The knowledge base is quite large, consisting of 648 rules of the form $d \rightarrow m$. We were only able to obtain 50 of the original 100 cases from the authors of the initial study, each consisting of an average of 8.56 symptoms.

Accel efficiently computed all of the minimal (w.r.t. subset) explanations in an average of 2.4 seconds per case on a Sun Sparc 2 workstation. Unfortunately, we could not compare this result to that obtained in the original study, since no information on run time was provided. However, the empirical results strongly suggest that a general abductive system can solve real diagnostic problems in reasonable time.

Since abduction computes the same explanations as set covering when given the same evaluation criteria, Accel should replicate the accuracy results of the original study. As discussed in the original study, minimality is too unrestrictive to produce useful results (Accel returned an average of 26.6 minimal diagnoses per case). With minimum cardinality, Accel produced an average of only 4.6 diagnoses per case. In 44% of the cases, one of these diagnoses matches the expert's exactly; and in another 46% of the cases, one of the system's diagnoses was a subset or superset of the expert's (called a "close match" in [Tuhrim *et al.*, 1991]). The remaining 10% of the cases have a diagnosis that either partially matches the expert's (2%) or all of the diagnoses are totally wrong (8%). These results are slightly better than those reported in the original study: 6.5 diagnoses/case with 40% exact, 38% close, 5% partial, 17% wrong. This is presumably due to the fact that our results are based on only 50 of the original 100 cases. Two other evaluation metrics reported in the original study, most-probable and minimum-collapsed, performed even better. In [Tuhrim *et al.*, 1991], it is claimed that, although there have been no direct comparisons, the results from any of the covering metrics appear more promising than those obtained from standard rule-based approaches to this problem.

5 MODEL-BASED DIAGNOSIS VIA ABDUCTION

Accel also performs model-based diagnosis, which concerns inferring faults from first principles given knowledge about the correct structure and behavior of a system. Much research in model-based diagnosis has taken the *consistency-based* approach and has been applied primarily to devices with static, persistent states such as combinational logic circuits [Davis, 1984; de Kleer and Williams, 1987; Reiter, 1987; de Kleer and Williams, 1989]. In the consistency-based approach, a diagnosis is a set of normality and abnormality assumptions about device components that are *consistent* with the observations and the system description. This is in contrast to the *abductive* approach of diagnosis used in Accel, where normality and abnormality assumptions about device components together with the system description must *imply* or *explain* the observations.

Poole has proved that the consistency-based and abductive approaches are equivalent for propositional theories [Poole, 1988], and Konolige has extended the conditions under which equivalence holds to general first-order causal theories allowing for correlations, uncertainty, and acyclicity in the causal structure [Konolige, 1992].[4] In view of such formal equivalence results, issues such as ease of representation and computational efficiency are most important. Our empirical results suggest that a number of diagnostic problems, ranging from combinational logic circuits to continuous dynamic systems such as a proportional temperature controller and the water balance system of the human kidney, can be effectively represented and efficiently diagnosed using an abductive approach.

Research in model-based diagnosis can also be clas-

[3] Actually, that all minimal covers of GSC are all minimal explanations in abduction also follows as a corollary of two published theorems, Theorem 7.1 in [Reiter, 1987] and Theorem 4.2 in [Poole, 1988].

[4] Abduction appears to be better in some cases, as Konolige has reported that "the utility of the consistency based method is open to question", since in explanatory diagnostic tasks, "the answers it produces may have elements that are not relevant to a causal explanation" [Konolige, 1992, page 257].

sified according to whether information about fault models is utilized in diagnosis. The *normality-based* approach of [Reiter, 1987; de Kleer and Williams, 1987] does not utilize fault models and any misbehavior differing from the correct functioning of a device can be diagnosed. However, the lack of fault models may result in hypothesizing implausible faults [de Kleer and Williams, 1989; Struss and Dressler, 1989]. On the other hand, the work of [Dvorak and Kuipers, 1989] is *fault-based* in that the fault models are *a priori* determined and given to the diagnostic system. Hence, unanticipated faults are not detected. ACCEL combines both normality-based and fault-based diagnosis in that information about fault models is used in diagnosis and any deviation from the correct behavior can be diagnosed. The diagnostic systems Sherlock [de Kleer and Williams, 1989] and GDE+ [Struss and Dressler, 1989] have similar capability.

In the model-based diagnosis domain, ACCEL uses predicate specific abduction, where the assumable atoms include component behavioral mode assumptions of three types: (1) a component is normal; (2) a component is in some known fault mode; or (3) a component is abnormal (but not necessarily in any known fault mode). Other assumable atoms are "auxiliary" assumptions including assumptions that the input values of a device are as given, and in dynamic system diagnosis, that some qualitative magnitude is positive/negative, that two qualitative values obey some corresponding value constraint, etc. (More details about these auxiliary assumptions will be provided later.) Explanations in this domain are evaluated based on simplicity, where the best explanation is one with the least number of components that are not normal, which include components that are in some known fault mode and those that are not. Normality assumptions and auxiliary assumptions are "free" and do not affect the simplicity metric of an explanation. If two explanations have the same number of components that are not normal, then the one with the most number of components that are in some known fault mode is preferred.

5.1 DIAGNOSING LOGIC CIRCUITS

In this section, we describe how the abductive approach of ACCEL is used to diagnose a full adder which is representative of standard, combinational logic circuits. Figure 1 shows a full adder which consists of 2 exclusive-or gates (x1, x2), two and gates (a1, a2), and one or gate (o1). We assume that each gate has 4 behavioral modes: normal (the output bit reflects the correct gate behavior at all times), stuck-at-0 (the output bit is stuck at 0 regardless of the input bits), stuck-at-1 (the output bit is stuck at 1 regardless of the input bits), and abnormal (the behavior of the gate is unconstrained).

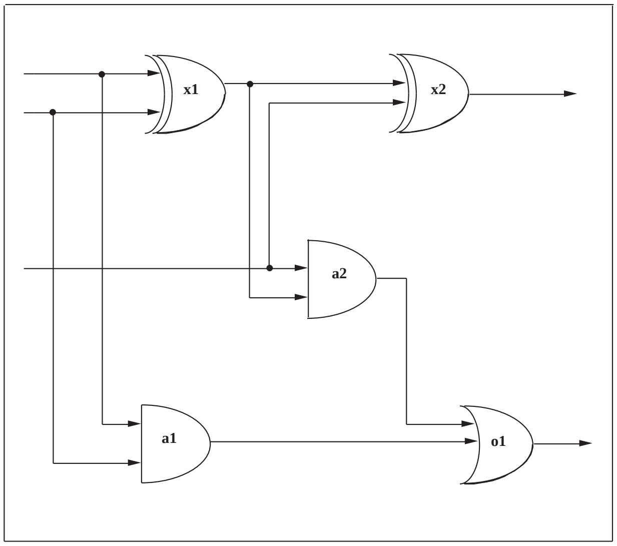

Figure 1: Full Adder

The knowledge base axiom that describes the correct behavior of an exclusive-or gate is:

$$\begin{aligned} out(X,W,T) \quad \leftarrow \quad & xorg(X) \wedge in1(X,U,T) \wedge \\ & in2(X,V,T) \wedge norm(X) \wedge \\ & xor(U,V,W) \end{aligned}$$

The axiom asserts that if X is an exclusive-or gate ($xorg(X)$), the first input of X is U at time T ($in1(X,U,T)$), the second input of X is V at time T ($in2(X,V,T)$), X is normal ($norm(X)$), and the exclusive-or of U and V is W ($xor(U,V,W)$), then the output of X is W at time T ($out(X,W,T)$). In addition we have the facts $xor(0,0,0)$, $xor(0,1,1)$, $xor(1,0,1)$, and $xor(1,1,0)$. The axioms for and gates and or gates are similar.

The following axiom describes the fault mode *stuck-at-0* for all gates:

$$\begin{aligned} out(X,0,T) \quad \leftarrow \quad & in1(X,U,T) \wedge in2(X,V,T) \wedge \\ & stuck\text{-}at\text{-}0(X) \end{aligned}$$

The axiom for the fault mode *stuck-at-1* is similar. Note that when a gate is assumed to be abnormal, no prediction can be made about its output bit. However, abduction requires that the observations be proved from the component behavioral mode assumptions (including the abnormality assumptions). To overcome this problem, we employ a technique used by Poole to "parameterize" the abnormality assumption as follows [Poole, 1989b]:

$$\begin{aligned} out(X,W,T) \quad \leftarrow \quad & in1(X,U,T) \wedge in2(X,V,T) \wedge \\ & ab(X,U,V,W,T) \end{aligned}$$

The antecedent $ab(X,U,V,W,T)$ in the rule is to be interpreted as "X is abnormal in such a way that at time T, given input bits U and V, its output bit is W". Note that for any input bits U and V, and any output bit W, the above axiom always allows us to assume that the component is abnormal by making the assumption $ab(X, U, V, W, T)$. This axiom achieves our objective of being able to prove the output observations from the parameterized abnormality assumption $ab(X, U, V, W, T)$.

So far, the axioms given are not specific to the full adder; they are used to model the behavior of exclusive-or gates, and gates, and or gates. We also need axioms that specify the connections among the gates in the given adder, such as

$$in1(a1, X, T) \leftarrow in1(x1, X, T)$$

as well as facts that identify the five components: $xorg(x1), xorg(x2)$, etc. Furthermore, in order to allow backward-chaining to terminate at the terminal input values of the full adder (these terminal input values cannot be further explained in terms of the other gate values), we need the axiom

$$in1(x1, X, T) \leftarrow given\text{-}in1(x1, X, T)$$

and two other similar axioms for the second input of $x1$ and the first input of $a2$. We let $given\text{-}in1(\ldots)$ (and $given\text{-}in2(\ldots)$) be assumable. They are the auxiliary assumptions, and do not affect the simplicity metric of an explanation.

To assess the performance of Accel, we randomly generated 10 scenarios by assuming that the various behavioral modes of each gate occur with the following probabilities: *norm* 65%, *stuck-at*-0 15%, *stuck-at*-1 15%, and *ab* 5%. Each of the 10 scenarios that was actually generated had one or two gates that were faulty, and the scenarios included some where a gate was abnormal (*ab*). For each scenario, we gave Accel I/O tuples where the input-output bits of the adder differed from those of a correctly functioning adder. (By an I/O tuple, we mean a particular combination of input and output values of the full adder.) For each I/O tuple, we first gave the three input bits and the two output bits of the adder, and then the output bits of the three gates x1, a1, and a2, in that order. For each scenario, we stopped as soon as the best diagnosis found by Accel is the correct diagnosis. We recorded the number of I/O tuples needed to converge on the correct diagnosis for each scenario. On a Sun Sparc 2 workstation, Accel took an average of 17 seconds to identify the correct diagnosis for each of the 10 scenarios tested. The average number of I/O tuples needed before the correct diagnosis was found is 2.1.

5.2 DIAGNOSING DYNAMIC SYSTEMS

Much research in model-based diagnosis has focused on diagnosing static, discrete devices like logic circuits. However, many devices and biological systems are continuous and dynamic and require reasoning about changes in behavior over time. Although there has been a great deal of research on modeling and simulating such systems [Kuipers, 1986; Forbus, 1984], there have been few attempts to apply general, model-based diagnostic methods to them. The work of [Ng, 1991; Ng, 1990] attempts to address this deficiency by diagnosing dynamic systems using the consistency-based approach. In this section, we present an abductive approach to diagnosing continuous, dynamic systems.

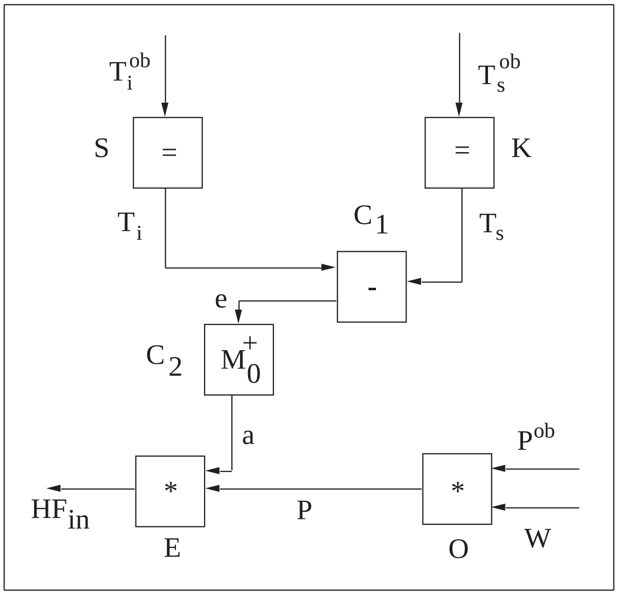

Figure 2: Temperature Controller

We adopt the representation of continuous dynamic systems used in the work of Kuipers' qualitative simulation (QSIM) [Kuipers, 1986]. The continuously changing behavior of a dynamic system over time is represented as a sequence of qualitative states, where a qualitative state consists of the qualitative values of the variables of the system. A qualitative value has two components: a qualitative magnitude (*qmag*) and a qualitative direction (*qdir*). A qualitative magnitude can either be a landmark value or an open interval between two landmark values, where a landmark value is a value of special significance that a variable takes on at some point in time. A qualitative direction can be one of increasing (*inc*), decreasing (*dec*), or steady (*std*).

The behavior of each dynamic system is governed by a set of qualitative constraints. The qualitative constraints on the temperature controller (Figure 2) are as follows (each constraint is preceded by a name identifying that constraint):

1. $S : T_i^{ob} = T_i$;
2. $K : T_s^{ob} = T_s$;
3. $C_1 : T_s - T_i = e$;
4. $C_2 : m_0^+(e) = a$;

5. $O : P^{ob} \cdot W = P$; and
6. $E : a \cdot P = HF_{in}$

The $m_0^+(e) = a$ constraint asserts that there is a strictly monotonically increasing function between e and a. However, the exact form of this monotonic function is unspecified. This accounts for the qualitative nature of the constraint. The purpose of this device is to control the temperature T_i^{ob} in the room, so that if the device is connected to a power source with power P^{ob}, the power switch is turned on (represented as $W = on$), and the temperature T_s^{ob} set by the temperature control knob differs from the temperature T_i^{ob} in the room, heat flow HF_{in} (in the form of hot air or cold air, depending on the direction of temperature difference) will be generated. Furthermore, the amount of heat flow generated is proportional to the temperature difference $T_s^{ob} - T_i^{ob}$.

We have successfully represented QSIM's knowledge about the various qualitative constraints ($=, -, \cdot, /, d/dt, m_0^+$) in Horn-clause axioms in a way suitable for logic-based abductive diagnosis. Since these Horn-clause axioms encode general knowledge about QSIM constraints, they are needed in the diagnosis of every dynamic system. These axioms encode the various qualitative constraints by defining a "holds.constraint-type" predicate for each type of qualitative constraint. For example, one of the 9 axioms that encode the m_0^+ constraint is:

$$\begin{array}{l} holds.m_0^+(F, G, M1, inc, M2, inc) \\ \leftarrow \\ pos(M1) \wedge pos(M2) \wedge corr\text{-}mag.m_0^+(F, G, M1, M2) \end{array}$$

The predicate $holds.m_0^+(F, G, M1, D1, M2, D2)$ asserts that $m_0^+(F) = G$ holds with the qualitative value of the variable $F = \langle M1, D1\rangle$ and the qualitative value of the variable $G = \langle M2, D2\rangle$. The predicate $pos(M1)$ ($neg(M1)$) asserts that the qualitative magnitude $M1$ is positive (negative). The predicate $corr\text{-}mag.m_0^+(F, G, M1, M2)$ asserts that $m_0^+(F) = G$ holds with the qualitative magnitude of $F = M1$ and the qualitative magnitude of $G = M2$. In QSIM, $(M1, M2)$ are referred to as corresponding values. The 9 axioms for the m_0^+ constraint cover all the distinct possibilities in which $m_0^+(F) = G$ holds since the qualitative magnitude of F can be positive, negative, or zero, and its qualitative direction can be *inc*, *std*, or *dec*. The other "holds.constraint-type" predicates, *holds.−*, *holds.∗*, *holds./*, and *holds.d/dt*, are defined by 39, 97, 70, and 9 axioms, respectively. The axioms for $holds. * (F, G, H, M1, D1, M2, D2, M3, D3)$ ensure that, among other things, the first-order derivative constraint $F \cdot G' + F' \cdot G = H'$ is obeyed. The exact axioms for all the qualitative constraints are listed in [Ng, 1992].

Besides the axioms that encode general QSIM constraints, there are also Horn-clause axioms that encode knowledge about a specific dynamic system. We assume in this paper that a dynamic system malfunctions because of one or more violated constraints, and that the task of mapping from violated constraints to the affected components is done by some other module external to Accel. The following axioms describe the normal behavior:

$$\begin{array}{l} qval(ti, M1, D1, T) \\ \leftarrow \\ norm(s) \wedge qval(ti\text{-}ob, M1, D1, T) \end{array}$$

$$\begin{array}{l} qval(e, M3, D3, T) \\ \leftarrow \\ norm(c1) \wedge qval(ts, M1, D1, T) \wedge qval(ti, M2, D2, T) \wedge \\ holds. - (ts, ti, e, M1, D1, M2, D2, M3, D3) \end{array}$$

The predicate $qval(ti, M1, D1, T)$ asserts that the qualitative value of the variable ti is $\langle M1, D1\rangle$ at time (qualitative state) T. The first axiom asserts that if constraint s is normal, and the qualitative value of *ti-ob* is $\langle M1, D1\rangle$ at time T, then the qualitative value of ti is also $\langle M1, D1\rangle$ at time T. This encodes the equality constraint between the variables *ti-ob* and ti. The second axiom asserts that if constraint $c1$ is normal, the qualitative value of ts is $\langle M1, D1\rangle$ at time T, the qualitative value of ti is $\langle M2, D2\rangle$ at time T, and $ts - ti = e$ holds with $ts = \langle M1, D1\rangle, ti = \langle M2, D2\rangle, e = \langle M3, D3\rangle$, then the qualitative value of e is $\langle M3, D3\rangle$ at time T. Similar axioms encode the other constraints.

Note that atoms with the predicate *qval* are not assumable. As such, in order to allow backward-chaining to terminate at the terminal input values of a dynamic device (these terminal input values cannot be further explained), we also need the axiom

$$qval(ti\text{-}ob, M1, D1, T) \leftarrow given\text{-}qval(ti\text{-}ob, M1, D1, T)$$

and three other similar axioms for *ts-ob*, *p-ob*, and *w*. We let *given-qval* be assumable. They are part of the "auxiliary" assumptions in an abductive explanation.

Note the directionality in which one qualitative value is explained in terms of other qualitative values. Since abductive diagnosis requires that the input observations (which consists of the qualitative values of the variables of a dynamic system) be *proved*, the axioms are formulated in such a way that the output values (e.g., $qval(hfin, \ldots)$) of a dynamic system can be proved from normality assumptions (e.g., $norm(s)$), fault mode assumptions, and auxiliary assumptions about the input values (e.g., $given\text{-}qval(ti\text{-}ob, \ldots)$) and the qualitative magnitudes and corresponding values of the variables (these are introduced when Accel attempts to prove the holds.constraint-type atoms).

We also assume that the components corresponding to the various constraints exhibit the following fault modes: stuck-at-0-std (S, K, C_1, C_2, O, E), stuck-at-roomtemp-std (S), stuck-at-1st-in (C_1, O), and stuck-at-2nd-in (C_1). Under the fault mode stuck-at-0-std (stuck-at-roomtemp-std), the output of a component is $\langle 0, std\rangle$ ($\langle room\text{-}temp, std\rangle$) regardless of the input values. Under the fault mode stuck-at-1st-in (stuck-at-2nd-in), the output of a component is stuck at its first (second) input. One Horn-clause axiom is used to encode one fault mode, as follows:

$$\begin{aligned} qval(ti, 0, std, T) &\leftarrow stuck\text{-}at\text{-}0\text{-}std(s) \wedge \\ & qval(ti\text{-}ob, M1, D1, T) \\ qval(e, M1, D1, T) &\leftarrow stuck\text{-}at\text{-}1st\text{-}in(c1) \wedge \\ & qval(ts, M1, D1, T) \wedge \\ & qval(ti, M2, D2, T) \end{aligned}$$

The Horn-clause axioms in Accel that represent the qualitative constraints capture the knowledge that QSIM uses to propagate qualitative values across constraints in order to complete the qualitative values of variables in a qualitative state. In Accel, such knowledge is used for the purpose of diagnosis. However, since the knowledge is now encoded declaratively, it can also be used for simulation purpose by a forward-chaining inference procedure. In fact, QSIM can be viewed as a special-purpose theorem prover for predicting the behavior of dynamic systems described by qualitative constraints. However, not all of QSIM's knowledge in simulation has been captured in Accel. Specifically, knowledge of state transition that QSIM uses to generate the next qualitative state(s) from an initial qualitative state is not encoded in Accel, since such knowledge is not needed in diagnosis.

We randomly generated 10 scenarios for the temperature controller where each scenario contains one to two faults and in which no heat flow was generated into the room. For each scenario, we gave the input atoms representing the qualitative values of the variables in the following order: $T_s^{ob}, T_i^{ob}, P^{ob}, W, HF_{in}$ at the initial qualitative state (t_1); $T_s^{ob}, T_i^{ob}, P^{ob}, W, HF_{in}$ at the next distinguished time-point qualitative state (t_2); and the intermediate variables T_s, T_i, e, a, P at state t_2.

In 9 out of the 10 scenarios, Accel found the correct diagnosis as its best diagnosis. The one scenario that Accel failed to find the best diagnosis has two faults $\{stuck\text{-}at\text{-}0\text{-}std(c1), stuck\text{-}at\text{-}0\text{-}std(c2)\}$. In this case, the best diagnosis that Accel found after processing all the intermediate variables is $\{stuck\text{-}at\text{-}0\text{-}std(c1)\}$. This is as it should be, since when $c1$ is stuck at $\langle 0, std\rangle$, the correct behavior of $c2$ if it is normal is to output $a = \langle 0, std\rangle$ at all times, which is indistinguishable from the behavior of $c2$ if it is in the fault mode *stuck-at-0-std*. That $c2$ is in fact faulty would be detected when $c1$ is replaced by a normal, working component and the controller is still found to be malfunctioning. Overall, the average run time per scenario is 4.24 minutes, and the average number of measurements of intermediate variables needed to arrive at the correct diagnosis is 4.4.

We also tested Accel on 10 faulty scenarios of the kidney water balance system, a QSIM model of which is given in [Kuipers, 1985; Kuipers, 1991]. The system has 7 qualitative constraints and 10 qualitative variables. Two of the scenarios tested correspond to the disorders Diabetes Insipidus and the Syndrome of Inappropriate Secretion of Anti-Diuretic Hormone (SIADH), which are disorders found in real patients. Accel found the correct diagnosis as its best diagnosis in all the 10 scenarios. The average run time per scenario is 6.98 minutes, and the average number of measurements of intermediate variables needed to arrive at the correct diagnosis is 3.7.

6 DISCUSSIONS AND CONCLUSIONS

In this paper, we have presented a general-purpose yet efficient system for making useful abductive inference and solving moderately complex problems in plan recognition and diagnosis. The comprehensive empirical results presented span the tasks of abductive plan recognition, set-covering-based diagnosis, and model-based diagnosis of both discrete and continuous dynamic systems. We believe our approach represents a good trade-off between generality and efficiency — Accel is a general-purpose system capable of performing all of the above tasks, yet efficient enough to be of practical utility.

Although Accel provides a declarative approach to the generation of explanatory hypotheses, it is often necessary that axioms be formulated carefully so that the system will perform the desired task correctly and efficiently. As in traditional logic programming, it is frequently insufficient to just "state the correct knowledge" and expect the desired answers to be inferred. Appropriate programming methodologies must be developed so that a user knows how to axiomatize a problem to correctly and efficiently compute the desired answers [Poole, 1989a]. This is also true in "abductive logic programming". By successfully applying Accel to the tasks of plan recognition and diagnosis, we have demonstrated via many examples *how* a general abductive system can be used to achieve these tasks.

Our empirical results suggest that the best criterion for evaluating explanations varies according to the domain. In diagnosis, the simplicity metric (defined as making the least number of assumptions in an abductive explanation) suffices, whereas in plan recognition for text understanding, because of the need to take into account the importance of text coherence, our coherence metric is a better criterion than simplicity.

Our results in set-covering-based diagnosis show that a general-purpose, logic-based abductive system can effectively represent and efficiently solve large realistic problems previously solved by the set covering method.

Consequently, the desirability of the existing special-purpose approach for such problems is lessened. The logic-based approach is more general and flexible, yet capable of efficiently solving problems in this more restrictive class.

Our empirical results in model-based diagnosis indicate that ACCEL's abductive approach can diagnose devices and systems previously solved by consistency-based methods. Although the work of [Poole, 1988; Poole, 1989b; Konolige, 1992] has revealed some interesting relationships between consistency-based and abductive diagnosis, the extent to which the two approaches coincide and differ, especially in practical terms such as ease of representation and diagnostic efficiency, requires further investigation. In addition, our research does not focus on intelligently gathering additional measurements to further differentiate and narrow the diagnostic candidates. Future work needs to extend ACCEL to incorporate intelligent experimentation.

In summary, this paper has demonstrated via an implemented system that general and efficient abduction for the tasks of plan recognition and diagnosis is indeed possible, and the future holds much promise for such a general abductive approach to explanation.

Acknowledgements

The first author was supported by an IBM Graduate Fellowship. This research was also supported by the NASA Ames Research Center under grant NCC 2-629. We are indebted to Dr. Stanley Tuhrim of Mount Sinai School of Medicine, New York, and Dr. James Reggia of the University of Maryland at College Park who kindly provided us the knowledge base on brain damage and the 50 patient test cases.

References

[Allen, 1987] James F. Allen. *Natural Language Understanding.* Benjamin/Cummings, Menlo Park, CA, 1987.

[Charniak and McDermott, 1985] Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence.* Addison Wesley, Reading, MA, 1985.

[Davis, 1984] Randall Davis. Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, 24:347–410, 1984.

[de Kleer and Williams, 1987] Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[de Kleer and Williams, 1989] Johan de Kleer and Brian C. Williams. Diagnosis with behavioral modes. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1324–1330, Detroit, MI, 1989.

[Dvorak and Kuipers, 1989] Daniel Dvorak and Benjamin J. Kuipers. Model-based monitoring of dynamic systems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1238–1243, Detroit, MI, 1989.

[Forbus, 1984] Kenneth D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

[Goldman, 1990] Robert P. Goldman. *A Probabilistic Approach to Language Understanding.* PhD thesis, Department of Computer Science, Brown University, Providence, RI, December 1990. Technical Report CS-90-34.

[Grice, 1975] H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[Konolige, 1992] Kurt Konolige. Abduction versus closure in causal theories. *Artificial Intelligence*, 53:255–272, 1992.

[Kuipers, 1985] Benjamin J. Kuipers. Qualitative simulation in medical physiology: A progress report. Technical Report MIT/LCS/TM-280, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, June 1985.

[Kuipers, 1986] Benjamin J. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29:289–338, 1986.

[Kuipers, 1991] Benjamin J. Kuipers. Qualitative reasoning: Modeling and simulation with incomplete knowledge. Book draft, May 1991.

[Lehnert and Sundheim, 1991] Wendy Lehnert and Beth Sundheim. A performance evaluation of text-analysis technologies. *AI Magazine*, 12(3):81–94, 1991.

[Levesque, 1989] Hector J. Levesque. A knowledge-level account of abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1061–1067, Detroit, MI, 1989.

[Ng and Mooney, 1990] Hwee Tou Ng and Raymond J. Mooney. On the role of coherence in abductive explanation. In *Proceedings of the National Conference on Artificial Intelligence*, pages 337–342, Boston, MA, 1990.

[Ng and Mooney, 1991] Hwee Tou Ng and Raymond J. Mooney. An efficient first-order Horn-clause abduction system based on the ATMS. In *Proceedings of the National Conference on Artificial Intelligence*, pages 494–499, Anaheim, CA, 1991.

[Ng, 1990] Hwee Tou Ng. Model-based, multiple fault diagnosis of time-varying, continuous physical devices. In *Proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications*, pages 9–15, Santa Barbara, CA, 1990. To appear in *Readings in Model-based Diagnosis*, edited by Walter Hamscher, Luca Console, and Johan de Kleer.

[Ng, 1991] Hwee Tou Ng. Model-based, multiple-fault diagnosis of dynamic, continuous physical devices. *IEEE Expert*, 6(6):38–43, 1991.

[Ng, 1992] Hwee Tou Ng. *A General Abductive System with Application to Plan Recognition and Diagnosis.* PhD thesis, Department of Computer Sciences, University of Texas at Austin, Austin, TX, May 1992.

[Norvig and Wilensky, 1990] Peter Norvig and Robert Wilensky. A critical evaluation of commensurable abduction models for semantic interpretation. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland, August 1990.

[Peng and Reggia, 1990] Yun Peng and James A. Reggia. *Abductive Inference Models for Diagnostic Problem-Solving.* Springer-Verlag, New York, 1990.

[Poole, 1988] David Poole. Representing knowledge for logic-based diagnosis. In *Proceedings of the International Conference on Fifth Generation Computing Systems*, pages 1282–1290, Tokyo, Japan, 1988.

[Poole, 1989a] David Poole. A methodology for using a default and abductive reasoning system. Technical Report 89-20, Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada, September 1989.

[Poole, 1989b] David Poole. Normality and faults in logic-based diagnosis. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1304–1310, Detroit, MI, 1989.

[Reggia *et al.*, 1985] James A. Reggia, Dana S. Nau, and Pearl Y. Wang. A formal model of diagnostic inference. I. problem formulation and decomposition. *Information Sciences*, 37:227–256, 1985.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[Selman and Levesque, 1990] Bart Selman and Hector J. Levesque. Abductive and default reasoning: A computational core. In *Proceedings of the National Conference on Artificial Intelligence*, pages 343–348, Boston, MA, 1990.

[Stickel, 1988] Mark E. Stickel. A Prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. Technical Note 451, SRI International, September 1988.

[Struss and Dressler, 1989] Peter Struss and Oskar Dressler. Physical negation — integrating fault models into the general diagnostic engine. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1318–1323, Detroit, MI, 1989.

[Tuhrim *et al.*, 1991] Stanley Tuhrim, James Reggia, and Sharon Goodall. An experimental study of criteria for hypothesis plausibility. *Journal of Experimental and Theoretical Artificial Intelligence*, 3:129–144, 1991.