ON STRATIFIED AUTOEPISTEMIC THEORIES

Michael Gelfond
Computer Science Department
The University of Texas at El Paso
El Paso, TX 79968

Abstract

In this paper we investigate some properties of "autoepistemic logic" approach to the formalization of common sense reasoning suggested by R. Moore in [Moore, 1985]. In particular we present a class of autoepistemic theories (called stratified autoepistemic theories) and prove that theories from this class have unique stable autoepistemic expansions and hence a clear notion of "theoremhood". These results are used to establish the relationship of Autoepistemic Logic with other formalizations of non-monotonic reasoning, such as negation as failure rule and circumscription. It is also shown that "classical" SLDNF resolution of Prolog can be used as a deductive mechanism for a rather broad class of autoepistemic theories. Key words and phrases: common sense reasoning, autoepistemic logic, negation as failure rule, non-monotonic reasoning. (Science section).

1. Introduction

In this paper we will investigate some properties of "autoepistemic logic" approach to the formalization of common sense reasoning suggested by R. Moore in [Moore, 1985]. This approach is based on ideas from [McDermott and Doyle, 1980] and [McDermott, 1982] and is meant to capture "nonmonotonicity" of common sense reasoning; i.e., the ability of a reasoning agent to withdraw some of his conclusions when a new evidence is presented. Moore concentrates on the type of reasoning which can be interpreted as reasoning about agent's knowledge or belief and uses modal logic (namely the notion of autoepistemic theory) to formalize this type of reasoning. Let us review some of the basic notions of his approach.

By an autoepistemic theory T we mean a set of formulae in the language of propositional calculus augmented by a belief operator L where Lf is interpreted as "f is believed" for any formula f. A formula in this language is called irreducible if it is an atom or begins with L. It is easy to see that each formula can be represented in exactly one way as a propositional combination of irreducible subformulas. The language of an autoepistemic theory T is the set of all propositional combinations of the irreducible components of the formulas from T. By Cn(T) we will denote the set of all formulas in the language of T which follow from T by propositional calculus. Obj(T) will stand for the set of all objective formulae from Cn(T) (i.e., the formulae of Cn(T) which do not contain the belief operator L).

Definition 1. (Moore) A set of formulae E(T) is a stable autoepistemic expansion of T if it satisfies the following condition:

$$E(T) = Cn(T + \{Lp: p \text{ is in } E(T)\} + \{\neg Lp: p \text{ is not in } E(T)\})$$

Moore shows that stable expansions contain all and only those formulae which are true in every interpretation of formulae from the language of T which satisfies T and makes Lp true for every formula, p, in the extension.

The notion of a stable autoepistemic expansion of a theory T plays a major role in the Moore's formalization of autoepistemic logic: it describes a set of beliefs of a rational agent with a set of premises T. The agent is rational in a sense that he believes in all and only those facts which are based on evidence rooted in his premises or in the stability condition. If this expansion is unique then it can be viewed as the set of theorems which follow from T in autoepistemic logic.

EXAMPLE 1. Consider the autoepistemic theory $T = \{\neg Lp \rightarrow q\}$. Let us informally investigate the construction of E(T). An agent with the set of premises T does not have any evidence in favor of p and hence p does not belong to his set of believes E(T). Therefore $\neg Lp$ is in E(T) (due to the stability condition) and hence q

is in E(T) (due to agent's ability to reason) and the only objective formulae belonging to E(T) are those from Cn(g). To construct formulae which express the agent's beliefs about objective statements we have to add to E(T) all formulae of the form Lf where f is in Cn(q) and formulae of the form ~Lf if f is not in Cn(q). In a similar way we can construct agent's beliefs about his beliefs about objective formulae, etc. It is easy to see that the resulting E(T) is the only stable expansion of T. This construction as well as the proof of the uniqueness of E(T) will be discussed in detail in Section 2. Example 1 also illustrates nonmonotonic nature of autoepistemic logic. The agent's present state of knowledge forces him to conclude q. But if new information about p becomes available this conclusion can be withdrawn which reflects the nonmonotonocity of this form of reasoning.

Unfortunately, as was recognized by Moore, a theory T may have more than one stable expansion or even no consistent stable expansion at all. To see why let us look at the following examples from [Moore, 1985]:

EXAMPLE 2. Let T = {~Lp → p}. A theory T has no consistent stable expansion. Informally: we have no evidence for p, hence we conclude that ~Lp which leads us to p and therefore Lp. Contradiction.

EXAMPLE 3. Let T = {~Lp → q, ~Lq →p}. It is easy to see that T has two stable expansions: E1 with an objective part Cn(q) and E2 with an objective part Cn(p).

This raises an important question of characterization of autoepistemic theories with unique stable expansions (i.e., clear notion of "theoremhood"). This question was first addressed in [Marek, 1986]. His results immediately imply the following theorem:

THEOREM 1. (Marek) Any consistent objective theory T (i.e., consistent theory without the belief operator) has a unique stable expansion E(T).

In the first part of this paper we will generalize this result and give sufficient conditions which guarantee the existence of a unique stable expansion for a much broader class of theories T. Theories from this class will be called stratifiable autoepistemic theories. Informally the notion is based on requiring the presence of certain hierarchy of predicates defined by a theory T which allows the use of formulae of the form Lf on the level k of this hierarchy only if f itself is fully defined on the lower levels.

The second part is devoted to the investigation of the relationship between "autoepistemic logic" formalization of common sense reasoning and the alternative formalization based on the "negation as failure rule" used in logic programming. We start with the review of the definition of stratified logic programs and their semantics [Apt et al, 1986] [Van Gelder, 1986] and then show that stratified logic programs can (in some precise sense) be interpreted in terms of belief. This, together with results from [Lifshitz, 1986], [Gelfond, Przymusinska, 1986] establishing the relationship between circumscription, autoepistemic logic and stratified logic programs shows that in the presence of a suitable hierarchy of definitions in a knowledge base different formalizations of nonmonotonicity in common sense reasoning essentially coincide. Another important consequence of this result is that it gives us a feasible deductive procedure we can use to characterize theorems of a broad class of autoepistemic theories.

To give a flavor of the techniques used to prove these results we include the complete proof of theorem 2. Complete proofs of other results will be published elsewhere.

2. Stratified Autoepistemic Theories

By literals we mean formulae of the forms p, ~p, Lf, ~Lf where p is a propositional letter and f is an objective formula. Literals which contain the belief operator L will be called autoepistemic while those without L will be called objective. From now on we will restrict our attention to autoepistemic theories consisting of clauses of the form S-->V where S is a list of literals and V is a list of atoms (both S and V can be empty).

DEFINITION 1. An autoepistemic theory T is called stratified if there is a partition $T = T_0 + \ldots + T_n$ such that:

(i) T_0 is objective (possibly empty)

(ii) clauses with the empty conclusions do not belong to Tk where $k > 0$.

(iii) if a propositional letter p belongs to the conclusion of a clause in T_k then literals p and ~p do not belong to $T_0, \ldots, T_{k-1}$ and literals Lf and ~Lf where f contains p do not belong to $T_0, \ldots, T_k$.

We will say that the degree of a propositional letter p is k and write D(p) = k if p belongs to the conclusion of a clause in T_k. If there is no such clause then the degree of p is 0. (It is obvious that if an autoepistemic theory is stratified then every propositional letter p has exactly one degree). The degree of an objective formula f is the maximum degree

of its propositional letters.

It is easy to see that theories from Examples 2 and 3 are not stratified while the theory from Example 1 is stratified with T_0 = { } and T1 = {~Lp -> q}.

We will start with a construction of the stable expansion of T. The idea is to first build the objective core of such an expansion and then to apply Marek's construction to it. Such an objective core is build gradually by expanding the corresponding layers of the stratified theory T. More precisely:

$K_0 = Cn(T_0)$.

$K_{m+1} = Cn(K_m + \{Lp:D(p) = m \ \& \ p \text{ in } K_m\} + \{\sim Lp:D(p) = m \ \& \ p \text{ not in } K_m\} + T_{m+1})$.

The following simple lemmas capture important properties of this construction.

LEMMA 1. Any model M_m of K_m can be expanded to a model M_{m+1} of K_{m+1}.

Proof. Let $M_{m+1} = M_m + \{Lp : D(p) = m \ \& \ p \text{ in } K_m\} + \{q : D(q) = m+1\}$. It can be easily seen from the definition of stratified autoepistemic theories that M_{m+1} is indeed a model of K_{m+1}.

LEMMA 2. (a) A theory K_m is consistent if T_m is consistent.
(b) K_{m+1} is a conservative extension of K_m.

Proof. Follows immediately from Lemma 1.

Now we can construct a stable expansion E of T.

DEFINITION 2. Let $K = K_n$ where T_n is the last layer of the partition of T. K is consistent and hence, in virtue of Theorem 1, there is a unique stable expansion of obj(K). Let us denote it by E.

To show that E is indeed a stable expansion of T we need the following Lemma.

LEMMA 3. For any objective formula f of degree m, f in obj(K_m) iff f in E.

Proof. The only if part is obvious. To prove the if part it suffices to notice that f in E implies f in obj(K) (see Theorem 2 from [Marek, 1986] and hence, by clause (b) of Lemma 2, we have that f is in obj(K_m).

THEOREM 2. Any consistent and stratified autoepistemic theory T has a stable expansion E(T).

Proof. To show that E is a stable expansion of T we have to prove that E satisfies the following condition:

(1) $E = Cn(T+\{Lf: f \text{ in } E\} + \{\sim Lf: f \text{ not in } E\})$.

Let us denote the set on the right side of this equation by R. From the definition of E we have that

(2) $E = Cn(obj(K) + \{Lf: f \text{ in } E\} + \{\sim Lf: f \text{ not in } E\})$.

and hence it remains to show that R = E.

(a) To show that E is in R let us prove first that for any m, K_m in R. We will use induction on m. The base is obvious and the inductive step follows immediately from Lemma 3. Now it suffices to notice that, by the definition of K, obj(K) is in R.

(b) To show that R is a subset of E we will prove that every model of E is a model of R. Suppose it is not the case and there is a model M of E which is not a model of R. Let U = (S --> V) be a clause from T of the lower degree m such that V is not empty and M(S) = True and M(V) = False. It is easy to see that such clause always exists and its premise S must contain autoepistemic literals (otherwise U would be in obj(K) and false in M which is impossible). Suppose that the first such a literal is Lq. Since E is complete w.r.t. autoepistemic literals (i.e., E |- Lq or E |- ~Lq) and M(Lq) = True we have that q is in E. The theory T is autoepistemic, therefore the degree of q is less than m, by Lemma 3 we have that q is in Km-1 and hence Lq is in Km. Now we can eliminate Lq from S and obtain a clause U1 which belongs to Km and fails in M. If the first autoepistemic literal in S is ~Lq it can be eliminated in exactly the same manner. By repeating this process we will eventually obtain a clause Ur which is objective, belongs to Km and fails in M which contradicts our assumption. Hence R is a subset of E. Q.E.P.

THEOREM 3. E is the only stable expansion of T.

3. Stratified Logic Programs

We will start with recalling the notion of stratified logic program (for the propositional case) and its semantics (see [Apt. et al., 1986], [Van Gelder, 1986]). By a logic program we mean a collection of clauses of the form S --> p where S is a (possibly empty) list of literals and p is a propositional letter. Logic programs are used to answer queries of the form l_1 V.. Vl_n where l_1, ..., ln are literals. In this process the 'negation as failure rule' is used which makes the precise definition of the notion of an answer to a query Q somewhat difficult to come up with. Recently several researchers independently suggested the characterization of a class of logic programs for which this notion allows elegant and clear semantics.

DEFINITION 3. A logic program LP is called stratified if there is a partition $LP = T_0 + \ldots + T_n$ such that if a propositional letter p belongs to the conclusion of a clause in T_k then p does not belong to $T_0, \ldots, T_{k-1}$ and ~p does not belong to $T_0, \ldots, T_k$.

Stratifiability is a condition on the use of negation in a logic program. Intuitively it forbids use of negation on formulas which are not completely defined.

EXAMPLE 4. It is easy to see that a program p, p --> q, ~q --> r is stratified with T_0 = {p, p --> q} and T1 = {~q --> r}.

The notion of an answer to a query for a stratified program LP is based on the following definition:

DEFINITION 4. Consider a sequence of theories ELPk (where ELP stands for an 'extension of logic program') such that
$ELP_0 = CWA(T_0) + \{\sim p:$ there is no clause in LP with a conclusion $p\}$;
$ELP_{k+1} = CWA(ELP_k + T_{k+1})$;
$ELP = ELP_k$;
where CWA(T) is Reiter's Closed Work Assumption of a theory T (see [R])i.e., CWA(T) = T + {~p: T |-/- p}.

PROPOSITION 1. For any stratified theory LP = T_0 + ... + T_n, ELP is consistent and has the unique model.

This model is intended to represent the universe described by LP. (It can be shown that model is exactly the "canonical" model of LP defined in [Apt, et al.]. The notion of "an answer to a query Q" is defined as follows:

DEFINITION 5. We will say that the query Q in LP has a positive answer and write LP | = Q iff Q is true in M. Otherwise the answer to Q is negative.

4. The Relationship

To investigate the relationship between logic programs and autoepistemic theories we will need a suitable mapping I from the propositional language in which logic programs are written into the corresponding language with the belief operator L.

DEFINITION 6. For any propositional formula f, I(f) is the formula obtained from f by replacing every occurrence of every negative literal ~p in f by the negative autoepistemic literal ~Lp. For any logic program LP, I(LP) = {I(S):S in LP}.

THEOREM 4. For any stratified program LP there is a unique autoepistemic expansion E of I(LP).

THEOREM 5. For any stratified program LP and for any query Q, LP | = Q iff E |-I(Q)

REMARK

The following corollary of Theorem 5 establishes the relationship between autoepistemic theories and prioritized circumscription (see [McCarthy, 1986]).

COROLLARY. For any stratified program LP and any query Q we have CIRC(LP,P1 > ... > Pn) |= Q iff E(LP) |- I(Q). Proof follows immediately from Theorem 3.1 of [L1], definition of ELP and Theorem 5. (On the relationship of circumscription and logic programs see also [L86 and P86]).

5. Conclusion

The Moore's formalization of autoepistemic logic is based on the notion of stable autoepistemic expansion of a theory T which represents a possible set of beliefs of a rational agent with a set of premises T. If such an expansion of a theory T is unique then it can be viewed as the set of theorems derivable from T in autoepistemic logic. In this paper we introduced a notion of stratified autoepistemic theory and showed that such theories have unique stable expansions. We believe that this result is of some importance not only because it clarifies the notion of "theoremhood" in autoepistemic logic but also because of the following reasons:
(a) Like many other nonmonotonic reasoning systems, autoepistemic logic was presented non-constructively. Neither the semantic basis nor the syntactic realization of this semantic provided a mechanism for arriving at the theorems of a given autoepistemic theory. We use the notion of stratification to show that "classical" SLDNF resolution of Prolog can be used as such a mechanism for a rather broad class of autoepistemic theories. This result also allows us to interpret the behavior of systems based on (propositional) Prolog in terms of belief and suggests possible directions in which Prolog can be extended to autoepistemic theories.
(b) The results of this paper suggest that a designer of a knowledge system based on autoepistemic logic may find it rewarding, both conceptually and computationally, to restrict yourself to stratified autoepistemic theories (very much as a designer of a traditional software system may find it rewarding to restrict yourself to traditional data structures such as stacks, trees, etc.). It is possible that other syntacticly described classes of autoepistemic theories more suitable for some types of applications will be discovered. But, in my judgement to make autoepistemic approach really practical we have to first extend it to allow quantification.

(c) Autoepistemic logic is based on an intuition rather different from those used for the development of other formalisms of this sort such as negation as failure rule or circumscription. We believe that the better understanding of the relationship between different forms of non-monotonic logic is essential for further development. It may help to single out areas of applicability of these methods, find their limitations and even eventually lead to the discovery of deeper underlying principals of non-monotonic reasoning.

6. Acknowledgements

I am grateful to Vladimir Lifshitz and Halina Przymusinska for numerous discussions on the subject of this paper.

7. References

[Apt, Blair, and Walker, 1986] Toward a Theory of Declarative Knowledge, In: Preprints of Workshop on Foundations of Deductive Databases and Logic Programs, 1986.

[Gelfond and Przymunsinska, 1986] On the Relationship between Circumscription and Autoepistemic Logic. Proceedings of International Symposium on Methodologies for Intelligent Systems, 1986.

[Lifschitz, 1985] Closed-World Databases and Circumscription. Artificial Intelligence 27, 1985.

[Lifschitz, 1986] On the Declarative Semantics of Logic Programs with Negation. In: Preprints of Workshop on Foundations of Deductive Databases and Logic Programs, 1986.

[Moore, 1985] Semantical Considerations on Nonmonotonic Logic, Artificial Intelligence 25 (1), 1985.

[Marek, 1986] Stable Theories in Autoepistemic Logic, (Preprint), 1986.

[McCarthy, 1986] Applications of Circumscription to Formalizing Common Sense Reasoning, Artificial Intelligence 28, 1986.

[McDermott and Doyle, 1980] Nonmonotonic Logic 1, Artificial Intelligence 13, 1980.

[McDermott, 1982] Nonmonotonic Logic 2: Nonmonotonic Modal Theories, J. ACM 29, 1982.

[Przymusinski, 1986] On the Semantics of Stratified Deductive Databases. In: Preprints of Workshop on Foundations of Deductive Databases and Logic Programs, 1986.

[Reiter, 1978] On Closed-World Databases In: Logic and Databases (H. Gallaire and J. Minker, Eds), 1978.

[Van Gelder, 1986] Negation as Failure Using Tight Derivations for General Logic Programs, Third IEEE Symposium on Logic Programming, 1986.