NON-MONOTONIC REASONING USING DEMPSTER'S RULE

Matthew L. Ginsberg
Department of Computer Science
Stanford University
Stanford, California 94305

ABSTRACT

Rich's suggestion that the arcs of semantic nets be labelled so as to reflect confidence in the properties they represent is investigated in greater detail. If these confidences are thought of as ranges of acceptable probabilities, existing statistical methods can be used effectively to combine them. The framework developed also seems to be a natural one in which to describe higher levels of deduction, such as "reasoning about reasoning".

I SEMANTIC NETS

Rich [Rich, 1983] has suggested labelling semantic nets [Quillian, 1968] with "certainty factors" indicating the depth of conviction held in the properties they represent. The non-monotonic rule "birds fly" would thus be represented not as

$$\text{birds} \xrightarrow{\text{isa}} \text{flyers}$$

but as

$$\text{birds} \xrightarrow[.95]{\text{isa}} \text{flyers}, \tag{1}$$

where the certainty factor of .95 indicates that 95% of birds fly. Monotonic rules have certainty factors of 1, as in

$$\text{ostriches} \xrightarrow[1.]{\text{isa}} \text{non-flyers}, \tag{2}$$

which is also written by Rich as

$$\text{ostriches} \xrightarrow[-1.]{\text{isa}} \text{flyers}.$$

Shafer [Shafer, 1976] has argued that probabilities such as those above are better thought of not as specific values, but as ranges. It seems unreasonable to believe that *exactly* 95% of all birds fly—much better to believe that between 90% and 98% do. Instead of having the conditional probability $p(\text{flyer}(x)|\text{bird}(x)) = .95$, we take $p(\text{flyer}(x)|\text{bird}(x)) \in [.9, .98]$.

We will write such a probability range as a pair $(c\ \ d)$, where c is the extent to which we believe a given proposition to be confirmed by the available evidence (.9 in the above example), and d is the extent to which it is disconfirmed ($1 - .98 = .02$ above). We will also write $\bar{c}$ for $1 - c$, $\bar{d}$ for $1 - d$, etc., so that the probability interval referred to in the last paragraph is $[c, \bar{d}]$ in general.

The beliefs (1) and (2) now become

$$\text{birds} \xrightarrow[(.9\ \ .02)]{\text{isa}} \text{flyers}$$

and

$$\text{ostriches} \xrightarrow[(0\ \ 1)]{\text{isa}} \text{flyers}$$

respectively. Since we need $c \leq \bar{d}$, we will always have $c + d \leq 1$, with equality only if the probability interval $[c, \bar{d}]$ is in fact a single point. As $c + d = 1$ corresponds to complete knowledge of a probability, so $c + d = 0$ corresponds to the interval $[0, 1]$ and therefore to no knowledge at all.

To perform simple reasoning using this representation, suppose we have

$$x \xrightarrow[(a\ \ b)]{\text{isa}} y \qquad \text{and} \qquad y \xrightarrow[(c\ \ d)]{\text{isa}} z, \tag{3}$$

and want to evaluate $x \xrightarrow{\text{isa}} z$. From the fact that the minimum probability of an x being a y is a, it follows that the minimum probability of an x being a z is at least ac. The probability of an x *not* being a z is at least ad for similar reasons. Thus the value of the arc $x \xrightarrow{\text{isa}} z$ is $(ac\ \ ad)$ and we have, for example, that

$$\text{Tweety} \xrightarrow[(1\ \ 0)]{\text{isa}} \text{birds} \xrightarrow[(.9\ \ .02)]{\text{isa}} \text{flyers}$$

gives rise to the non-monotonic conclusion

$$\text{Tweety} \xrightarrow[(.9\ \ .02)]{\text{isa}} \text{flyers}. \tag{4}$$

II DEMPSTER'S RULE

The difficulties with this scheme arise when differing applications of the rule used in (3) lead to different conclusions. If we have

$$\text{Tweety} \xrightarrow[(1\ \ 0)]{\text{isa}} \text{ostriches} \xrightarrow[(0\ \ 1)]{\text{isa}} \text{flyers},$$

we obtain

$$\text{Tweety} \xrightarrow[(0\ \ 1)]{\text{isa}} \text{flyers}, \tag{5}$$

in contradiction with (4).

This situation is typical of non-monotonic reasoning. Default rules by their very nature admit exceptions; what we need is some way to combine conflicting conclusions such as (4) and (5).

Dempster [Dempster 1968 or Shafer 1976] has discussed this situation in depth, and our problem is in fact a special case of his investigations. If we denote by $(a\ \ b) + (c\ \ d)$ the inference obtained by combining the two inferences $(a\ \ b)$ and $(c\ \ d)$, Dempster's rule gives us

$$(a\ \ b) + (c\ \ d) = \left(1 - \frac{\bar{a}\bar{c}}{1-(ad+bc)} \quad 1 - \frac{\bar{b}\bar{d}}{1-(ad+bc)}\right). \tag{6}$$

This formulation has the following attractive properties:

a. It is commutative and associative. In many non-monotonic systems, the order in which non-monotonic (or other) inferences are drawn is critical, since the application of one rule may invalidate another. The commutativity and associativity of (6) guarantees that we will be able to overcome this difficulty.

b. $(a\ \ b) + (0\ \ 0) = (a\ \ b)$. The probability range $(0\ \ 0)$ corresponds to no knowledge at all and will result from any attempt to apply an inapplicable rule. We might, for example, generate the arc $\text{Tweety} \xrightarrow[(0\ \ 0)]{\text{isa}} \text{flyers}$ from the pair

$$\text{Tweety} \xrightarrow[(0\ \ 1)]{\text{isa}} \text{elephants} \xrightarrow[(0\ \ 1)]{\text{isa}} \text{flyers}.$$

The point here is that such an inference (should we draw it) will have no effect on our eventual conclusions.

c. $(a\ \ 0) + (c\ \ 0) = (a + c - ac\ \ 0)$. The probability ranges $(a\ \ 0)$ and $(c\ \ 0)$ each indicate no *disbelief* in the corresponding arcs; in this case, the (independent) probabilities combine in the usual fashion.

d. For $(c\ \ d) \neq (0\ \ 1)$, $(1\ \ 0) + (c\ \ d) = (1\ \ 0)$; for $(c\ \ d) \neq (1\ \ 0)$, $(0\ \ 1) + (c\ \ d) = (0\ \ 1)$. This result implies that no application of a non-monotonic rule can ever outweigh a logical certainty. There is no danger when applying a non-monotonic rule to obtain (4) that an eventual conclusion such as (5) will be invalidated; the result of combining the two results is simply (5) again. This allows us to avoid the most computationally difficult aspect of non-monotonic reasoning—that of determining when it is legitimate to apply a non-monotonic rule of inference.

e. $(0\ \ 1) + (1\ \ 0)$ is undefined. Such a combination indicates that an arc has been proven both valid and invalid and as such represents a conflict in the database.

f. $+$ is (nearly) invertible. If we denote the inverse by $-$, we have, for $(c\ \ d) \neq (0\ \ 1)$ or $(1\ \ 0)$,

$$(a\ \ b) - (c\ \ d) = \left(\frac{\bar{c}(a\bar{d} - \bar{b}c)}{\bar{c}\bar{d} - \bar{b}c\bar{c} - \bar{a}d\bar{d}} \quad \frac{\bar{d}(b\bar{c} - \bar{a}d)}{\bar{c}\bar{d} - \bar{b}c\bar{c} - \bar{a}d\bar{d}}\right). \tag{7}$$

This enables us to easily retract the conclusion of an earlier inference without influencing conclusions drawn using other means.

III RULES AND METARULES

A more efficient approach to non-monotonic deduction is implied by McCarthy's formulation [McCarthy, 1984]:

$$\begin{aligned}
&\text{bird}(x) \wedge \neg\text{abnormal}_1(x) \rightarrow \text{flies}(x)\\
&\text{ostrich}(x) \rightarrow \text{abnormal}_1(x)\\
&\text{ostrich}(x) \wedge \neg\text{abnormal}_2(x) \rightarrow \neg\text{flies}(x).
\end{aligned}$$

The effect of these rules is to have the fact that Tweety is an ostrich invalidate not the conclusion that Tweety can fly, but the rule which led to that conclusion. In our formulation we want to deactivate not the arc

$$\text{Tweety} \xrightarrow[(.9\ \ .02)]{\text{isa}} \text{flyers}$$

but the rule corresponding to

$$\text{birds} \xrightarrow[(.9\ \ .02)]{\text{isa}} \text{flyers} \tag{8}$$

itself. In order to see how to do this, we need first to describe the rule (8) in greater detail.

We will think of a rule as a triple $(a \quad c \quad p)$ where a is a list of antecedents, c is a consequent, and p is a probability interval. The intention is that if all of the antecedents are satisfied, then the consequent holds with probability range p. An example will probably provide the best clarification: The rule "if x is a bird, then x can fly" will be represented as

$$\begin{array}{l}(((\text{isa } x \text{ birds}))\\ \ (\text{isa } x \text{ flyers})(.9 \ \ .02)).\end{array} \tag{9}$$

The antecedent list consists of the single arc (isa x birds). The consequent is (isa x flyers), with confidence (.9 .02).

The rule itself is activated at the same level as the antecedent (for multiple antecedents, a product should be used). Thus if the value of (isa x birds) is $(a \quad b)$, the ensuing increment to (isa x flyers) will be $(.9a \quad .02a)$.

Returning to the ostrich case, we have the rules

$$\begin{array}{l}(((\text{isa } x \text{ ostriches}))\\ \ (\text{isa } x \text{ flyers})(0 \ \ 1))\end{array} \tag{10a}$$

and

$$\begin{array}{l}(((\text{isa } x \text{ ostriches}))\\ \ (\text{rule } ((\text{isa } x \text{ birds})) \ (\text{isa } x \text{ flyers})(.9 \ \ .02))(0 \ \ 1)).\end{array} \tag{10b}$$

The first of these simply repeats the rule that ostriches cannot fly. The second, however, deactivates the rule (9) itself. If the rule has been applied, the reversability of Dempster's rule ensures that the conclusions will remain accurate; if the rule has not been applied, we will be saved the work of doing so.

In the example we have been considering, the certainty of the rule that ostriches do not fly guarantees that we will reach the same conclusion whether or not we apply (9) to an ostrich. But consider the following example:

Newspaper articles are true. (.9 .05) (11a)

Articles in the *National Enquirer* are true. (.5 .4) (11b)

If I read something in the National Enquirer, both rules can be applied and I will believe the story to be true with probablity interval (.92 .07). Here we really do need a rule such as (10b) that ensures that (11a) will not be applied when (11b) can be.

Better still would be the metarule, "Never apply a rule to a set when there is a corresponding rule which can be applied to a subset." We could write this as

$$\begin{array}{l}(((\text{isa } x \ y)\\ \ \ (\text{rule } (a \ (\text{isa } z \ x) \ b) \ c \ d))\\ \ (\text{rule } (a \ (\text{isa } z \ y) \ b) \ c \ e)(0 \ \ 1)).\end{array} \tag{12}$$

As a special case, we have

$$\begin{array}{l}(((\text{isa Enquirer-article newspaper-article})\\ \ \ (\text{rule } ((\text{isa } x \text{ Enquirer-article})) \ (\text{accurate } x) \ (.5 \ \ .4)))\\ \ (\text{rule } ((\text{isa } x \text{ newspaper-article})) \ (\text{accurate } x) \ (.9 \ \ .05))\\ \qquad (0 \ \ 1)).\end{array} \tag{12'}$$

Now suppose we read an article in the National Enquirer. Rules (11a) and (11b) are activated, with (11b) activating the metarule (12) and therefore deactivating (11a). The article is now believed to be true with confidence (.5 .4).

Equally important is what happens if we later read the same article in the New York Times. Now rule (11a) alone is applied and the article is believed to be true, corroborated to some extent by the Enquirer appearance.

IV PROBABILITIES FOR RULES

The power of the methods we have described potentially extends well beyond the examples we have given thus far. The best interpretation of a metarule such as (12), for example, is probably as a way to assign a probability range to a rule itself. Thus in applying a rule with probability range $(a \quad b)$, we should weight its conclusion by $\bar{b}$ before updating any other probabilities, since $\bar{b}$ is the maximum extent to which the rule may be applicable.

Implementation of this idea will require us to maintain a list of rules which have been either used or activated by other rules. There are three advantages to this. Firstly, it enables us to avoid reapplying a single rule without obtaining new information. Since Dempster's rule assumes independence of the probability estimates being combined, multiple used of a single rule need to be avoided.

Secondly, this approach enables us to *partially* deactivate a rule. Returning to our newspaper example, perhaps all we should say is that the rule (11a) is not as useful for Enquirer articles:

(((isa x Enquirer-article))
(rule ((isa x newspaper-article)) (accurate x)(.9 .05))
(0 .4)).
(13)

If we use this rule instead of ($11b$)—note that ($12'$) now disappears—an article in the National Enquirer will be believed to be true with probability range (.54 .03).

Note also that the commutativity and invertability of Dempster's rule mean that we need not apply (13) before ($11a$) in order to obtain this result—provided that we store the information that we *have* used ($11a$), we will have no difficulty reversing the inference after a subsequent invocation of (13).

A final (but currently unexplored) advantage of this approach is that it may allow us to focus the attention of the system. For a rule which has probability range (a b), we can think of a as the extent to which the rule is likely to be useful. To focus attention on the fact that birds fly, we might have

(((isa x birds))
(rule ((isa x birds)) (isa x flyers) (.9 .02))(.5 0)).
(14)

(Such a rule will itself need a high level of activation to be of any use). If we are maintaining a list of rules and the levels with which they are expected to be useful, a rule such as (14) can be used to ensure, for any forward-chaining system, that the inference that any given bird can probably fly will be drawn early.

More generally, we can translate, "When considering an element of a group, think about properties which are unique to that group," into the metarule

(((isa x y)
(isa z y)) (15)
(rule ((isa x y)) (isa x z) a) (.5 0)).

If birds were the only flyers, so that flyers $\xrightarrow{\text{isa}}$ birds had truth value (1 0), this would reproduce (14), with y = birds and z = flyers. As it stands, the result of applying (15) to birds and flyers will be somewhat weaker.

V CONCLUSION

The assignment of confidence ranges to arcs in semantic nets seems to solve some of the problems which would otherwise be encountered in dealing with them. Non-monotonic inferences can be described easily, and mesh neatly with their monotonic counterparts.

Further power can be obtained by allowing the rules themselves to be treated as arcs, both by including them within other rules and by assigning them probabilistic weights of their own. Reasoning about reasoning can be discussed, and attention can be focussed. This framework seems to be a promising one in which to describe general knowledge of the type we have been examining.

ACKNOWLEDGEMENTS

The author would like to thank John McCarthy, Ben Grosof, Mike Genesereth and Sally Greenhalgh for many stimulating and useful discussions.

REFERENCES

[1] Dempster, A.P. "A Generalization of Bayesian Inference." J. Roy. Stat. Soc. *B*, **30** (1968), 205–247

[2] McCarthy, J. "Applications of Circumscription to Formalizing Common Sense Knowledge," to appear (1984)

[3] Quillian, M.R. "Semantic Memory," in *Semantic Information Processing,* ed. M. Minsky. Cambridge: MIT Press, 1968, pp. 216–270.

[4] Rich, E. "Default Reasoning as Likelihood Reasoning." In *Proc. AAAI-83*. Los Altos, California: William Kaufmann, Inc., 1983, pp. 348–351

[5] Shafer, G. *A Mathematical Theory of Evidence.* Princeton: Princeton University Press, 1976