Failure-Driven Line of Reasoning Explanations

William J. Clancey

Institute for Research on Learning
3333 Coyote Hill Road
Palo Alto, CA 94304

Extended Abstract
Fourth International Workshop on Natural Language Generation
July 18-20, Catalina, CA 1988

We are continuing to investigate in NEOMYCIN possible methods for summarizing a line of reasoning in an expert system. In our current research, we are attempting to generate the least text possible to satisfy the questioner. In this paper, we relate this approach to an explanation-based learning (EBL) method that attempts to unify a theory of knowledge acquisition (machine learning) with explanation (human learning).

In MYCIN, an inquiry about why the program was requesting a patient finding was answered by indicating how the information would be used to make a specific domain inference (e.g., "if the patient is an alcoholic, and if ..., then I will be able to conclude that e.coli...").

In NEOMYCIN, the program indicated more generally what it was trying to do in the course of diagnosing the patient, which we termed a "strategic explanation" (e.g., "I am testing the hypothesis that the patient has e.coli meningitis. Alcoholism predisposes for e.coli meningitis."). Originally, we generated such text by printing the task (strategic goal) and its focus (e.coli in this example). We used a template associated with each metarule to indicate the connection between the current finding request and the task. The entire line of reasoning was described up to the first task without a specific finding or disease focus (Hasling et al., 1983).

We desired a more principled method for translating metarules, avoiding use of ad hoc templates, and this motivated the re-representation of the original Lisp-coded metarules in the predicate calculus. We then experimented with fairly primitive paragraph summaries, omitting computational clauses (such as bookkeeping information about whether a rule had been applied). This text was informative, but often verbose. Many rephrasings seemed possible. With many alternatives seeming equally good, we didn't yet have a theory of what constituted an adequate explanation.

We are now trying a radically different explanation strategy. Rather than printing a complete summary of what the program did, we are attempting to minimize what the program says. We are testing the hypothesis that the minimum explanation indicates how each focus (finding or hypothesis) relates to the previous focus. Specifically, the program inspects the metarules to determine which propositions establish a binding between the previous and new focus. For example, the program shifts focus from "headache" (a new finding for which it is doing data-directed reasoning) to "headache-duration" by the relation PROCESS-SUBTYPE--the duration specifies the process of headache in more detail. Similarly, the program shifts from "diplopia" (again, a new finding the program is reasoning about) to "increased-intracranial-pressure" by the CAUSES relation--the former causes the latter. These are the relations in the premises of metarules that establish a new focus for the called task (e.g., ask about headache-duration, look for evidence of increased-pressure). The tasks themselves, we hypothesize, need not be mentioned because the listener will be able to infer that we are "clarifying new information" and "testing a hypothesis." To make the line of reasoning clear, we proceed from the most recent focus, up the stack, to earlier contexts.

With this design, it is much easier to combine sentences as well as to make analogies between explanations (e.g, "the program is doing the same thing it was doing at question 12, only here we are focusing on the cause of diplopia").

A methodological advantage of this minimalist approach to explanation is that it provides a failure-driven constraint to the research itself. Explanations in the past always seemed merely "okay." We believe now that they provided too much information, and we were really investigating forms of paraphrasing, rather than determining what information the listener needed to hear and what he could infer. Under the current approach, failures to understand will be more easily tracked to awkward phrasing of the relations themselves, leaving out a computational or filtering relation (see below) or leaving out a task. That is, we will build up our theory of explanation by identifying specific reasons why something needs to be said.

In related research, we are attempting to unify an explanation-based (EBL) approach to knowledge acquisition (machine learning) with our theory of explanation for a person trying to understand NEOMYCIN's consultation behavior.

The EBL program, called GUIDON-DEBUG, critiques a diagnostic solution in terms of the syntactic constraints a diagnostic explanation should satisfy (the form of a diagnostic model) and then invokes a "why not" explanation program to determine why the constraints weren't satisfied. A constraint, such as "every abnormal finding should be explained," specifies when particular kinds of links should be added to a graph representing the diagnostic model of the patient. Each terminal task in NEOMYCIN's diagnostic procedure (e.g, test-hypothesis, process-hypothesis, refine-hypothesis, process-finding) places particular links in the graph, so an unsatisfied constraint corresponds to a missing link, which in turn corresponds to a failed task or a task that wasn't invoked.

This suggests a unified theory for line of reasoning explanations, justifying the minimal approach given above. In the case of GUIDON-DEBUG, explaining why the program *didn't* do something, we give t*he most general reason* why the particular metarule failed. That is, if several propositions establish a binding relation between a previous and new focus, we mention the most general proposition. For example, it makes more sense to say "I didn't ask about X as a result of this rule because X isn't a laboratory test," rather than to give the more specific reason for failure that "... X isn't a cause of process Y." In the context being questioned, the program only asks about laboratory tests; it is a more general focus-binding relation.

More specifically, we find that relations in metarule premises are of four types:

1) unary relation (e.g., X is a laboratory test);
2) focus-binding relations (e.g., Y causes X);
3) domain specialization relations (e.g., Y necessarily causes X);
4) context applicability relations (e.g., X has not been requested yet).

When explaining why a metarule wasn't applied in a particular context, we want to determine the most general (weakest) failed precondition. In the case of an observer's inability to understand why NEOMYCIN requested a finding (or more generally, why it did any given task), we start by assuming that the observer knows the context of what the program is doing (possibly a wrong assumption. We indicate the *most-specific* domain relation (1, 2 or 3) (e.g., "because Y necessarily causes X") leaving him to infer the more general relations (1 and 2) from the task (e.g., X is being considered in general because we are considering laboratory tests).

One justification for this approach is that the observer would have questioned an earlier action in the same general context if he didn't know what the program is doing *in general* at this time. This argues for the importance of giving explanations about a reasoning process within the context of an ongoing series of finding requests and observer inquires, making assumptions on the basis of what the observer apparently understood about earlier interactions.

More specifically, it is apparent that this research must be placed in some context of *what the user is trying to do.* After the initial text generation capability has been debugged, we expect to continue development within the context of GUIDON-MANAGE. In this program, the student solves a diagnostic problem by issuing abstract commands to NEOMYCIN, directing its reasoning. The student might request an explanation in order to understand the program's advice of what to do next or to understand how a particular finding request generated by NEOMYCIN followed from the strategic command supplied by the student. At the same time we are are attempting to formally derive the diagnostic metarules from the constraints mentioned above and a set of assumptions about the world (covering the communication involved in acquiring findings and assumptions about the case population). This additional information could then serve to justify the metarules and provide another form of explanation.