Preventing False Inferences[1]

Aravind Joshi and Bonnie Webber
Department of Computer and Information Science
Moore School/D2
University of Pennsylvania
Philadelphia PA 19104

Ralph M. Weischedel[2]
Department of Computer & Information Sciences
University of Delaware
Newark DE 19716

ABSTRACT

I Introduction

In cooperative man-machine interaction, it is taken as necessary that a system truthfully and informatively respond to a user's question. It is not, however, sufficient. In particular, if the system has reason to believe that its planned response might lead the user to draw an inference that it knows to be false, then it must block it by modifying or adding to its response. The problem is that a system neither can nor should explore all conclusions a user might possibly draw: its reasoning must be constrained in some systematic and well-motivated way.

Such cooperative behavior was investigated in [5], in which a modification of Grice's *Maxim of Quality* is proposed:

> Grice's *Maxim of Quality* -
> Do not say what you believe to be false or for which you lack adequate evidence.
>
> Joshi's *Revised Maxim of Quality* -
> If you, the speaker, plan to say anything which may imply for the hearer something that you believe to be false, then provide further information to block it.

This behavior was studied in the context of interpreting certain definite noun phrases. In this paper, we investigate this revised principle as applied to question answering. In particular the goals of the research described here are to:

1. characterize tractable cases in which the system as respondent (R) can anticipate the possibility of the user/questioner (Q) drawing false conclusions from its response and can hence alter or expand its response so as to prevent it happening;
2. develop a formal method for computing the projected inferences that Q may draw from a particular response, identifying those factors whose presence or absence catalyzes the inferences;
3. enable the system to generate modifications of its response that can defuse possible false inferences and that may provide additional useful information as well.

Before we begin, it is important to see how this work differs from our related work on responding when the system notices a discrepancy between its beliefs and those of its user [7, 8, 9, 18]. For example, if a user asks "How many French students failed CSE121 last term?", he shows that he believes inter alia that the set of French students is non-empty, that there is a course CSE121, and that it was given last term. If the system simply answers "None", he will assume the system concurs with these beliefs since the answer is consistent with them. Furthermore, he may conclude that French students do rather well in a difficult course. But this may be a false conclusion if the system doesn't hold to all of those beliefs (e.g., it doesn't know of any French students). Thus while the system's assertion "No French students failed CSE121 last term" is true, it has misled the user (1) into believing it concurs with the user's beliefs and (2) into drawing additional false conclusions from its response.[3] The differences between this related work and the current enterprise are that:

1. It is not assumed in the current enterprise that there is any overt indication that the domain beliefs of the user are in any way at odds with those of the system.
2. In our related work, the user draws a false conclusion from what is said because the presuppositions of the response are not in accord with the system's beliefs (following a nice analysis in [10]). In the current enterprise, the user draws a false conclusion from what is said because the system's response behavior is not in accord with the user's expectations. It may or may not also

[1]This work is partially supported by NSF Grants MCS 81-07290, MCS 83-05221, and IST 83-11400.

[2]At present visiting the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104.

[3]It is a feature of Kaplan's CO-OP system [7] that it points out the discrepancy by saying "I don't know of any French students"

involve false domain beliefs that the system attributes to the user.

In this paper, we describe two kinds of false conclusions we are attempting to block by modifying otherwise true response:

- false conclusions drawn by standard default reasoning - i.e., by the user/listener concluding (incorrectly) that there is nothing special about this case
- false conclusions drawn in a task-oriented context on the basis of the user's expectations about the way a cooperative expert will respond.

In Section II, we discuss examples of the first type, where the respondent (R) can reason that the questioner (Q) may inappropriately apply a default rule to the (true) information conveyed in R's response and hence draw a false conclusion. We characterize appropriate information for R to include in his response to block it. In Section III, we describe examples of the second type. Finally, in Section IV, we discuss our claim regarding the primary constraint posed here on limiting R's responsibilities with respect to anticipating false conclusions that Q may draw from its response: that is, it is only that part of R's knowledge base that is already in focus (given the interaction up to that point, including R's formulating a direct answer to Q's query) that will be involved in anticipating the conclusions that Q may draw from R's response.

II Blocking Potential Misapplication of Default Rules

Default reasoning is usually studied in the context of a logical system in its own right or an agent who reasons about the world from partial information and hence may draw conclusions unsupported by traditional logic. However, one can also look at it in the context of interacting agents. An agent's reasoning depends not only on his perceptions of the world but also on the information he receives in interacting with other agents. This information is partial, in that another agent neither will nor can make everything explicit. Knowing this, the first agent (Q) will seek to derive information implicit in the interaction, in part by contrasting what the other agent (R) has made explicit with what Q assumes would have been made explicit, were something else the case. Because of this, R must be careful to forestall inappropriate derivations that Q might draw. The question is on what basis R should reason that Q may assume some piece of information (P) would have been made explicit in the interaction, were it the case.

One basis, we contend, is the likelihood that Q will apply some *standard default rule* of the type discussed by Reiter [15] if R doesn't make it explicite that the rule is not applicable. Reiter introduced the idea of default rules in the stand-alone context of an agent or logical system filling in its own partial information. Most standard default rules embody the sense that "given no reason to suspect otherwise, *there's nothing special about the current case*". For example, for a bird what would be special is that it can't fly - i.e., "Most birds fly". Knowing only that Tweety is a bird and no reason to suspect otherwise, an agent may conclude by default that there's nothing special about Tweety and so he can fly.

This kind of default reasoning can lead to false conclusions in a stand-along situation, but also in an interaction. That is, in a question-answer interaction, if the respondent (R) has reason for knowing or suspecting that the situation goes counter to the standard default, it seems to be common practice to convey this information to the questioner (Q), to block his potentially assuming the default. To see this, consider the following two examples. (The first is very much like the "Tweety" case above, while the second seems more general.)

A. Example 1

Suppose it's the case that most associate professors are tenured and most of them have Ph.Ds. Consider the following interchange

Q: Is Sam an associate professor?

R: Yes, but he doesn't have tenure.

There are two things to account for here: (1) Given the information was not requested, why did R include the "but" clause, and (2) why this clause and not another one? We claim that the answer to the second question has to do with that part of R's knowledge base that is currently in focus. This we discuss more in Section IV. In the meantime, we will just refer to this subset as "RBc".

Assume RBc contains at least the following information:

(a) Sam is an associate professor.

(b) Most associate professors are tenured.

(c) Sam is not tenured.

(b) may be in RBc because the question of tenure may be in context. Based on RBc, R's direct response is clearly "Yes". This direct response however could lead Q to conclude falsely, by default reasoning, that Sam is tenured. That is, R can reason that, given just (b) and his planned response "Yes" (i.e., if (c) is not in Q's knowledge base), Q could infer by default reasoning that *Sam is tenured*, which R knows with respect to RBc is false. Hence, R will modify that planned response to block this false inference, as in the response above.

In general, we can represent R's reasoning about Q's reaction to a simple direct response "Yes, B(a)", given Q believes "Most Bs F", in terms of the following default schema, using the notation introduced in [15].

told(R,Q,B(c)) & (Most x)[B(x) ⇒ F(x)]
& ¬told(R,Q,¬F(c)): M(F(c))

F(c)

As in Reiter's discussion, "M(P)" means it is consistent to assume that P. In the associate professor example, B corresponds to the predicate "is an associate professor", F, to the predicate "has tenure", and c, to Sam. Using such an instantiated rule schema, R will recognize that Q is likely to conclude F(c) - "Sam has tenure" - which is false with respect to RBc (and hence, with respect to all of R's knowledge base). Thus R will modify his direct response so as to block this false conclusion.

B. Example 2

Consider a user one of the mail systems on the DEC-20. To exit from this system, a user who has finished reading all the messages he earlier specified can just type a carriage return. To exit under other circumstances, the user must type QUIT. Consider the following interchange between a new user who has finished reading all his messages and either a mail system expert or the mail system itself.

Q: How do I get out of mail?
R: Since you have read all your specified messages, you can just type a carriage return. In all cases, you can get out by typing QUIT.

Here the problem is to account for all that part of R's response beyond the simple truthful statement "You can type a carriage return."

A general statement of this problem is as follows: Agent Q is in one situation (S1) and wants to be in another (S2). There is a general procedure P for achieving S2 from any of several situations including S1. There is a special prodecure P* (i.e., shorter, faster, simpler, etc.) for achieving S2 from S1. Q doesn't know how to achieve S2, but R does (including procedures P and P*). Q asks R how to achieve S2.

If R knows that Q is in situation S1 and truthfully responds to Q's request by simply telling him P*, Q may falsely conclude that P* is a general procedure for achieving S2. That is, as in the Tweety and Sam examples, if Q has no reason to suspect anything special about S1 (such that P* only applies to it), then there is nothing special about it. Therefore P* is adequate for achieving S2, whatever situation Q is in.[4] Later when Q tries to apply P* in a different situation to achieve S2, he may find that it doesn't work. As a particular example of this, consider the mail case again. In this case,

S1 = Q has read all his messages
S2 = Q is out of the mail system
P = typing QUIT
P* = typing a carriage return

Assume RBc contains at least the following information:

(a) S1
(b) want(Q,S2)
(c) ∀s∈Σ . P(s) = S2
(d) P*(S1) = S2
(e) S1∈Σ
(f) simpler(P*,P)
(g) ∀s∈Σ . ¬(s = S1) ⇒ ¬(P*(s) = S2)

where Σ is some set of states which includes S1 and P(s) indicates action P applied to state S.

Based on RBc, R's direct response would be "You can exit the mail system by typing carriage return". (It is assumed that an expert will always respond with the "best" procedure according to some metric, unless he explicitly indicates otherwise - cf. Section III, case 2). However, this could lead Q to conclude falsely, by default, something along the lines of ∀s . P*(s) = S2.[5] Thus R will modify his planned response to call attention to S1 (in particular, how to recognize it) and the limited applicability of P* to S1 alone. The other modification to R's response ("In all cases, you can get out by typing QUIT"), we would ascribe simply to R's adhering to Grice's *Maxim of Quantity* - "Make your contribution as informative as is required for the current purposes of the exchange" - given R's assumption of what is required of him in his role as expert/teacher.

III Blocking False Conclusions in Expert Interactions

The situations we are concerned with here are ones in which the system is explicitly tasked with providing help and expertise to the user. In such circumstances, the user has a strong expectation that the system has both the experience and motivation to provide the most appropriate help towards achieving the user's goals. The user does not expect behavior like:

Q: How can I get to Camden?
R: You can't.

As many studies have shown [1], what an advice seeker (Q) expects is that an expert (R) will attempt to recognize what plan Q is attempting to follow in pursuit of what goal and respond to Q's question accordingly. Further studies [11, 12, 13] show that Q may also expect that R will respond in terms of a better plan if the recognized one is either sub-optimal or unsuitable for attaining Q's perceived goal. Thus because of this principle of "expert cooperative behavior", Q may expect a response to a more general question than the one he has actually asked. That is, in asking an expert "How do I do X?" or "Can I do X?", Q is anticipating a response to "How can I achieve my goal?"

---

[4] Moreover if Q (falsely) believes that R doesn't know Q is in S1, Q will certainly assume that P* is a general procedure. However, this isn't necessary to the default reasoning behavior we are investigating.

[5] Clearly , this is only for some subset of states, ones corresponding to being in the mail system.

Consider a student (Q) asking the following question, near the end of the term.

Q: Can I drop CIS577?

Since it is already too late to drop a course, the only direct answer the expert (R) can give is "No". Of course, part of an expert's knowledge concerns the typical states users get into and the possible actions that permit transitions between them. Moreover it is also part of this expertise to infer such states from the current state of the interaction, Q's query, some shared knowledge of Q's goals and expectations and the shared assumption that an expert is expected to attend to these higher goals. How the system should go about inferring these states is a difficult task that others are examining [2, 12, 13]. We assume that such an inference has been made. We also assume for simplicity that the states are uniquely determined. For example, we assume that the system has inferred that Q is in state Sb (student is doing badly in the course) and wants to be in a state Sg (student is in a position to do better in this course or another one later), and that the action a (dropping the course) will take him from Sb to Sg.

Given this, the response in (2) may lead Q to draw some conclusions that R knows to be false. For example, R can reason that since a principle of cooperative behavior for an expert is to tell Q the best way to go from Sb to Sg, Q is likely to conclude from R's response that there is no way to go from Sb to Sg. This conclusion however would be false if R knows some other ways of going from Sb to Sg. To avoid potentially misleading Q, R must provide additional information, such as

R: No, but you can take an incomplete and ask for more time to finish the work.

As we noted earlier, an important question is how much reasoning R should do to block false conclusions on Q's part. Again, we assume that R should only concern itself with those false conclusions that Q is likely to draw that involve that part of R's knowledge base currently in focus (RBc), including of course that subset R needs in order to answer the query in the first place.

We will make this a little more precise by considering several cases corresponding to the different states of R's knowledge base with respect to Sb, Sg, and transitions between them. For convenience, we will give an appropriate response in terms of Sb, Sg and the actions. Clearly, it should be given in terms of descriptions of states and actions understandable to Q. (Moreover, by making further assumptions about Q's beliefs, R may be able to validly trim some of its response.)

1. Suppose that it is possible to go from Sb to Sg by dropping the course and that this is the only action that will take one from Sb to Sg.

   Sb Sg

   In this case, the response is

   R: Yes, α is the only action that will take you from Sb to Sg.

2. Suppose that in addition to going from Sb to Sg by dropping the course, there is a better way, say β, of doing so.[6]

   Sb Sg

   In this case, the response is

   R: Yes, but there is a better action β that will take you from Sb to Sg.

3. Suppose that dropping the course does not take you from Sb to Sg, but another action β will. This is the situation we considered in our earlier discussion.

   Sb Sg

   In this case the response is

   R: No, but there is an action β that will take you from Sb to Sg.

4. Suppose that there is no action that will take one from Sb to Sg.

   Sb Sg

   In this the response is

   R: No. There is no action that will take you from Sb to Sg.

Of course, other situations are possible. The point, however, is that the additional information that R provides to prevent Q from drawing false conclusions is limited to just that part of R's knowledge base that R is focussed on in answering Q's query.

IV Constraining the Respondent's Obligations

As many people have observed - from studies across a range of linguistic phenomena, including co-referring expressions [3, 4, 16], left dislocations [14], epitomization [17], etc. – a speaker (R) normally focuses on a particular part of its knowledge base. What he focuses on depends in part on (1) context, (2) R's partial knowledge of Q's overall goals, as well as what Q knows already as a result of the interaction up to that point, and (3) Q's particular query, etc. The precise nature of how these various factors affect focusing is complex and is receiving much attention [3, 4, 16]. However, no matter how these various factors contribute to focusing, we can certainly assume that R comes to focus on a subset of its knowledge base in order to provide a direct answer to Q's query (at some level of interpretation). Let us call this subset RBc for "R's current beliefs". Our claim is that one important constraint on cooperative behavior is that it is determined by RBc only. Clearly the information needed for a direct response is contained in RBc, as is the information needed for many types of helpful responses. In other words, RBc – that part of R's knowledge base that R decides to focus on in order to give a direct response to Q's query – also has the information needed to generate several classes of helpful responses. The simplest case is presupposition failure [7], as in the following

Q: How many A's were given in CIS 500 ?

where Q presumes that CIS 500 was offered. In trying to formulate a direct response, R will have to ascertain that CIS 500 was offered. If it was (Q's presumption is true), then R can go ahead and give a direct response. If not, then R can indicate that CIS 500 was not offered and thereby avoid misleading Q. All of this is straightforward. The point here is that the information needed to provide this extra response is already there in that part of R's knowledge base which R had to look up anyway in order to try to give the direct response.

In the above example, it is clear how the response can be localized to RBc. We would like to claim that this approach has a wider applicability: that RBc alone is the basis for responses that anticipate and attempt to block interactional defaults as well. Since RBc contains the information for a direct response, R can plan one (r). From r, R can reason whether it is possible for Q to infer some conclusion (g) which R knows to be false because $\neg$g is in RBc. If so, then R should modify r so as to eliminate this possibility. The point is that the only false inferences that R will attempt to block are those whose falsity can be checked in RBc.

[6] "Betterness" is yet another area for future research.

There may be other false inferences that Q may draw, whose falsity cannot be determined solely with respect to RBc (although it might be possible with respect to R's entire knowledge base). While intuitively this may not seen enough of a constraint on the amount of anticipatory reasoning that Joshi's revised maxim imposes on R, it does constrain things a lot by only considering a (relatively small) subset of knowledge base. Factors such as context may further delimit S's responses, but they will all be relative to RBc.

V Conclusion

There are many gaps in the current work and several aspects not discussed here. In particular,

1. We are developing a formalism for accommodating the system's reasoning based on a type of HOLDS predicate whose two arguments are a proposition and a state; see [6].
2. We are working on more examples, especially more problematic cases in which, for example, a direct answer to Q's query would be "yes" (or the requested procedure) BUT a response to Q's higher goals would be "no" or "no" plus a warning - e.g.,

Q: Can I buy a 50K savings bond?
S: Yes, but you could get the same security
on other investments with higher returns.

3. We need to be more precise in specifying RBc, if we are to assume that all the information needed to account for R's cooperative behevior is contained there. This may in turn reflect on how the user's knowledge base must be structured.
4. We need to be more precise in specifying how default rules play a role in causing R to modify his direct response, in recognition of Q's likelihood of drawing what seems like a generalized "script" default - if there is no reason to assume that there is anything special about the current case, don't.

REFERENCES

[1] Allen, J.
Recognizing Intentions from Natural Language Utterances.
In M. Brady (editor), *Computational Models of Discourse*, . MIT Press, Cambridge MA, 1982.

[2] Carberry, S.
Tracking User Goals in an Information-Seeking Environment.
In *Proceedings of the National Conference on Artificial Intelligence*, pages 59-63. AAAI, 1983.

[3] Grosz, B.
The Representation and Use of Focus in Dialogue Understanding.
Technical Report 151, SRI International, Menlo Park CA, 1977.

[4] Grosz, B., Joshi, A.K. & Weinstein, S.
Providing a Unified Account of Definite Noun Phrases in Discourse.
In *Proc. 21st Annual Meeting*, pages 44-50. Assoc. for Computational Ling., Cambridge MA, June, 1983.

[5] Joshi, A.K.
Mutual Beliefs in Question Answering Systems.
In N. Smith (editor), *Mutual Belief*, . Academic Press, New York, 1982.

[6] Joshi, A., Webber, B. & Weischedel, R.
Living Up to Expectations: Computing Expert Responses.
In *Proceedings of AAAI-84.* Austin TX, August, 1984.

[7] Kaplan, J.
Cooperative Responses from a Portable Natural Language Database Query System.
In M. Brady (editor), *Computational Models of Discourse*, . MIT Press, Cambridge MA, 1982.

[8] Mays, E.
Failures in natural language systems: application to data base query systems.
In *Proc. First National Conference on Artificial Intelligence (AAAI).* Stanford CA, August, 1980.

[9] McCoy, K.
Correcting Misconceptions: What to Say.
In *CHI'83 Conference Human Factors in Computing Systems.* Cambridge MA, December, 1983.

[10] Mercer, R. & Rosenberg, R.
Generating Corrective Answers by Computing Presuppositions of Answers, not of Questions.
In *Proceedings of the 1984 Conference*, pages 16-19. Canadian Society for Computational Studies of Intelligence, University of Western Ontario, London, Ontario, May, 1984.

[11] Pollack, M., Hirschberg, J. and Webber, B.
User Participation in the Reasoning Processes of Expert Systems.
In *Proc. AAAI-82.* CMU, Pittsburgh PA, August, 1982.
A longer version appears as Technical Report CIS-82-9, Dept. of Computer and Information Science, University of Pennsylvania, July 1982.

[12] Pollack, Martha E.
Goal Inference in Expert Systems.
Technical Report MS-CIS-84-07, University of Pennsylvania, 1984.
Doctoral dissertaion proposal.

[13] Pollack, M.
Good Answers to Bad Questions.
In *Proc. Canadian Society for Computational Studies of Intelligence (CSCSI).* Univ. of Western Ontario, Waterloo, Canada, May, 1984.

[14] Prince, E.
Topicalization, Focus Movement and Yiddish Movement: A pragmatic differentiation.
In D. Alford et al. (editor), *Proceedings of the 7th Annual Meeting*, pages 249-64. Berkeley Linguistics Society, February, 1981.

[15] Reiter, R.
A Logic for Default Reasoning.
Artificial Intelligence 13:81-132, 1980.

[16] Sidner, C. L.
Focusing in the Comprehension of Definite Anaphora.
In M. Brady (editor), *Computational Models of Discourse*, . MIT Press, Cambridge MA, 1982.

[17] Ward, G.
A Pragmatic Analysis of Epitomization: Topicalization it's not.
In *Proceedings of the Summer Meeting 1982.* LSA, College Park MD, August, 1982.
Also in Papers in Linguistics 17.

[18] Webber, B. & Mays, E.
Varieties of User Misconceptions: Detection and Correction.
In *Proc. IJCAI-8.* Karlsruhe, Germany, August, 1983.