What's in an Answer: A Theoretical Perspective on Deductive Question Answering*

L. K. Schubert
L. Watanabe

Department of Computing Science
The University of Alberta
Edmonton T6G 2H1

Abstract -- Questions and answers are examined with a view towards improving the quality of answers producible by a deductive question answering algorithm. Most past work assumes that question answering is reducible to theorem proving. This reduction is often accompanied by implicit assumptions about the form of the question and answer. The focus in this paper is on questions interpretable as requests for information. The form of such requests is analyzed, and desirable properties of responses defined. An answer is classified as categorial or selective, and characterized by properties of correctness, completeness, definiteness, nontriviality, comprehensibility, and informativeness. Extant deductive question answering procedures are evaluated with respect to these properties and shown to be deficient in their ability to generate certain classes of answers. The study leads to various interesting open problems, including that of devising a reasonably efficient procedure for categorial question answering.

1. Introduction

"What's in an answer?" is a question left formally unanswered by most researchers in question-answering. Usually, questions and answers are defined by informal elucidation and examples. Such definitions make it difficult to evaluate the performance of question answering systems.

Indeed, we believe that the lack of formal definitions has led to a general misapprehension about the status of deductive question answering: it is widely assumed that certain procedures for deductive question answering, to be found in any AI text that deals seriously with deduction, are logically adequate for yes-no and wh-questions (at least if the reasoning needed to prove that the answer follows from the premises can be carried out within a standard first-order framework). We will show that this procedure, due to Green, Luckham and Nilsson, is not in general adequate for wh-question answering.

One problem that arises in any discussion of question answering is how to separate question answering from question interpretation. We are not concerned here with natural language understanding, but rather with deductive information retrieval from a knowledge base. Therefore we will focus on certain kinds of formal ***requests for information***, taking for granted the process by which a request might have been derived from an English utterance. The utterance might have been a question ("Who in the department knows prolog?"), an imperative sentence ("Tell me who.."), an indicative sentence ("I would like to know who..."), an exclamation ("If only you would tell me who..."), or a fragment of any of these types of sentences. Nevertheless, we will continue to refer to requests for information informally as questions, since that is their paradigmatic form in English.

* We would like to thank Jeff Pelletier and the referees for their helpful comments on the manuscript. This research was supported in part by NSERC Operating Grant A8818.

Broadly speaking, there are two types of requests for information, corresponding respectively to yes-no questions and constituent (or wh-) questions. In the first type, the question-asker is requesting the truth value of a sentence; in the second, he is requesting information about the membership of a specified set, or for multi-constituent questions, of a specified relation (as in "Who in your class loves whom?").

Evidently, any complete proof procedure is formally adequate for answering yes-no questions: we interleave a proof attempt and a disproof attempt for the sentence in question, and answer T ("Yes") if and when the proof attempt succeeds, F ("No") if and when the disproof attempt succeeds, and U ("Unknown") if the sentence is found to be logically independent of the available information, or resource constraints are exceeded before a T or F answer is determined. Let us assume that the information available for question answering is self-consistent. Then this process has the property that for any question, there exists a finite (though in general unknown) resource bound such that if the process is run with that bound or a more generous one, the answer will be T iff the sentence is a logical consequence of the available information, F iff its negation is a logical consequence of the available information, and U iff both the sentence and its negation are consistent with the available information. In other words, the question is answered as correctly and completely as the available information permits, ***in the limit*** as resource constraints are relaxed. Given the undecidability of first-order logic, this correctness and completeness in the limit is the best one could hope for.

In view of the formal adequacy of standard deductive methods for yes-no questions, we will focus on (single-) constituent questions.

2. Related Work

Our immediate concern is with deductive question answering procedures of the type developed by Green [Green 69] and Luckham and Nilsson [Luckham & Nilsson 71], which are well-established in AI and will be discussed in some detail in section 7.

However, questions and answers have also been extensively studied in the field of philosophical logic, beginning with Adjukiewicz's analysis of questions and answers in terms of propositional functions. Our own proposals will lean on this tradition to some extent. Like Adjukiewicz (and more recently [Bennett 79]), we take single-constituent questions to be, in essence, questions about the truth set of a given predicate. (This is defined as the ***request predicate***, R, in the next section.)

Aqvist [Aqvist 65] and Hintikka [Hintikka 76] treat questions as requests for information and attempt to analyze them in terms of the knowledge states of the questioner and answerer. For example, according to Hintikka a question of form "Which x is (has property) F?" has associated with it a desired state (or ***desideratum***)

$$(\exists x)K_iF(x)$$

where "K_i" means "I know that". This desired epistemic state is attained if I come to know that $F(b)$, and can truthfully assert

I know what (or who) b is.

Our own criteria for questions of this type will be somewhat weaker, requiring only that answers be as specific as the question answerer's knowledge permits, and that they avoid "private symbols" known only to the question answerer.

Belnap and Steel's concepts of category conditions, selection, completeness-claim, and distinctness-claim have analogues in our concepts of answer types, completeness, and definiteness, but require the introduction of new types of quantifiers and operators into the logic [Belnap & Steel 76].

In general, the philosophical logic literature has only been moderately useful for our endeavour, since it has little to say about the *inference* of answers to questions. A focus on deductive question answering leads to problems quite different from those considered in the philosophical literature, ones for which even an adequate formal vocabulary is lacking. Thus one of our main objectives is to begin development of such a vocabulary. As far as possible, we avoid modal notions (especially in the object language of the question asker and answerer), since our initial aim is to analyze deductive question answering based on first-order logic. However, some of our definitions make use of a computational notion of "knowing" similar to that of Konolige [Konolige 85].

Finally we should mention relevant work in computational linguistics, particularly that of Cohen and Levesque [Cohen & Levesque 85] who analyze questions as speech acts involving the goals and intentions of the questioner. We assume that the reasoning processes they describe would sometimes lead to requests for information of the type which provide the starting point for our own investigation. Moreover, goal-directed reasoning might be resumed if the initial request cannot be satisfactorily met (according to criteria such as those we shall propose).

3. The Form of Questions and Answers

The form of a single-constituent question (more accurately, request for information) expressible in first-order logic may be taken to be

$$(\text{wh } x\text{: } P(x))Q(x)$$

where P is the *presupposed category*,
Q is the *question predicate*,
$R =_{df} \lambda x[P(x) \land Q(x)]$ is the *request predicate*

Formally, $P(x)$, $Q(x)$ and $R(x)$ are open sentences containing x as their only free variable. The question is intuitively to be read as "Which P is a Q?" or "Which P's are Q's?". For our purposes, the wh operator is considered part of the *interaction* language but not part of the *representation* language of the question answering system. (The source of this distinction is [Levesque 84].) We will not attempt to explicate the semantics of wh directly, but only indirectly through our formalization of *answers* to wh-questions.

Our terminology is motivated by the following considerations: a question of the above form presupposes the existence of entities of the presupposed category P. For example, if we ask, "Which CMPUT 552/85 students obtained 9's?", we presuppose that there *are* CMPUT 552/85 students. If there are none, the response should be corrective, i.e., it should not simply be "None", but rather, "There are no CMPUT 552/85 students" (cf. [Webber & Mays 83]).

Q is called the question predicate since it can be thought of as deriving from the verb phrase (i.e., the predicate) of a question expressing the request for information.

R, the request predicate which combines P and Q, turns out to be useful in the formulation of desirable constraints on answers.

Note that our formulation precludes "How...", "Why...", and "How many..." questions, among others, unless our logic treats "ways, manners, and means", reasons, and numbers as ordinary individuals. Questions such as "Which two people..." are also excluded, unless our logic treats collections (such as pairs, triples, etc., of people) as ordinary individuals. Nevertheless, the range of questions we are allowing is quite sufficient for a preliminary study of answers and their properties, and for an assessment of deductive question answering methods.

In answering a request for information of the above type, we assume that we first check whether any presuppositions of the request (including the presupposition $(\exists x)P(x)$) are violated. If we prove a presupposition to be violated, we generate a corrective response (i.e., its denial). Besides attempting to disprove presuppositions, we also attempt to prove a negative answer to the question itself, i.e., $\sim(\exists x\text{: } P(x))Q(x)$, and respond appropriately if the attempt succeeds. After this "negative phase" of the question-answering process, we attempt to provide a positive answer. (We do not discuss the negative phase further since, as in the case of yes-no questions, standard deductive methods are adequate for it; the negative phase would in practice be interleaved with the positive phase.)

The form of a positive answer may be taken to be

$$(\alpha x\text{: } A(x))Q(x)$$

where α is a quantifier and A is the *answer predicate*. (Again, $A(x)$ is an open sentence whose only free variable is x.)

If $\alpha = \forall$ (every), we have a *categorial answer*.

If $\alpha = \exists$ (some), we have a *selective answer*.

In standard first-order logic, categorial and selective answers assume the respective forms

$$(\forall x)[A(x) \to R(x)], \text{ and } (\exists x)[A(x) \land R(x)].$$

Categorial answers and selective answers can be thought of as corresponding respectively to plural and singular wh-constituents in the question. Thus the question "Which department members know prolog?" might prompt the categorial answer "All department members other than Jones (know prolog)", i.e.,

$$A(x) = \text{dept-member}(x) \land \sim(x = \text{Jones})$$

while the question "Which department member knows prolog?" might prompt the selective answer "Smith", i.e.,

$$A(x) = (x = \text{Smith})$$

or perhaps "Smith or Jones (I'm not sure which)", i.e.,

$$A(x) = (x = \text{Smith}) \lor (x = \text{Jones})$$

However, it is in general a matter of pragmatics, requiring reasoning about the goals of the questioner, whether a categorial or selective answer is desired; thus we will not further discuss this issue here.

We think that categorial and selective answers are two of the most important types (cf. [Hintikka 76]), though not the only ones. For example, a reasonable and useful response to the first of the above questions might be "Most of the department members (know prolog)". However, we will not consider responses involving nonstandard quantifiers like "most", which cannot be translated into ordinary first-order logic. It is worth remarking here that answers involving the negative quantifier "no", such as "No department members who know Fortran (know prolog)" need not be separately considered, since they can be regarded as categorial answers to other questions, such as "Which department members do not know prolog?". (The issue of when an answer to the *negative* form of the surface question is conversationally appropriate is again a matter of pragmatics.)

We now proceed to our characterization of answers.

4. Correctness and Completeness (or Definiteness)

Certainly correctness is an essential property of any cooperative response to a request for information.

We define an answer $(\alpha x\colon A(x))Q(x)$ to be *correct* iff

$$S_a \models (\alpha x\colon A(x))R(x),$$

where S_a is the set of facts available for question answering and "$\models$" denotes "entails" (has as logical consequence).[1] Thus, to be correct not only must the answer hold, but as well, the instances of Q provided by it must be of the presupposed category P. Note that for categorial answers and selective answers, the correctness criterion is equivalent respectively to:

$$S_a \models (\forall x)[A(x) \rightarrow R(x)]$$

$$S_a \models (\exists x)[A(x) \wedge R(x)]$$

Thus, for example, if the question is "Which countries have part of the Rocky Mountains in them?" the following answers are considered correct (where P = "country", Q = "has part of the Rocky Mountains in it", R = "country which has part of the Rocky Mountains in it"):

(a) "Canada and the US": $(\forall x\colon x=\text{Canada} \vee x=\text{US})Q(x)$

(b) "Canada": $(\exists x\colon x=\text{Canada})Q(x)$

(c) "Canada or the US": $(\exists x\colon x=\text{Canada} \vee x=\text{US})Q(x)$

(d) "Canada or Alberta":

$$(\exists x\colon x=\text{Canada} \vee x=\text{Alberta})Q(x)$$

Note that (a) is categorial, while (b) - (d) are selective. (We are taking for granted that the decision to answer categorially or selectively has been made on suitable pragmatic grounds.)

The following answers are incorrect even though they are true:

(e) "Canada or Alberta":

$$(\forall x\colon x=\text{Canada} \vee x=\text{Alberta})Q(x)$$

(f) "Alberta": $(\exists x\colon x=\text{Alberta})Q(x)$

(g) "Alberta or Ontario":

$$(\exists x\colon x=\text{Alberta} \vee x=\text{Ontario})Q(x)$$

It is worth noting that we could have imposed correctness criterion

$$S_a \models (\alpha x\colon A(x))Q(x) \text{ and } S_a \models (\forall x\colon A(x))P(x)$$

instead of the one proposed, without essential change. This requires the answer predicate to specify a subcategory of P. For categorial answers, this is equivalent to the previous criterion, but for selective answers, it is somewhat stronger: it judges (d) above to be incorrect. However, any correct answer according to the weaker criterion is easily converted to a correct answer according to the stronger criterion, by conjoining $P(x)$ to $A(x)$. For example, the following is correct according to the stronger criterion:

(d') "A country identical with Canada or Alberta":

$$(\exists x\colon P(x) \wedge (x=\text{Canada} \vee x=\text{Alberta}))Q(x)$$[2]

Another desirable property of answers is that they be as *complete* as possible. However, one has to be careful in formalizing this idea. In the case of categorial answers, it is clear that "completeness" can be identified with "maximum coverage" by A (of the positive instances of R) but in the case of ***selective answers*** this notion of maximum coverage is inappropriate, since selective answers ***by definition*** need to supply only one instance of R. What ***is*** desirable, however, is that this instance be as fully specified as possible. The best we can hope for is that A ***uniquely identifies an instance***; in that case the existential quantifier in the selective answer can be replaced by the ***definite quantifier***, "the". These considerations motivate the following definitions.

A categorial answer is *complete* iff

$$S_a \models (\forall x)[R(x) \rightarrow A(x)]$$

Note that completeness and correctness of categorial answers amounts to equivalence of the answer predicate to the request predicate, i.e.,

$$S_a \models (\forall x)[A(x) \equiv R(x)]$$

An ideal categorial question answerer should always try to prove or disprove that its answer is complete and, unless $A \equiv P$, inform the questioner as to the outcome of the proof/disproof attempt. (E.g., "Canada and the US, and no others"; or "Canada, among others"; or "Canada and possibly others".)

A selective answer is *definite* iff

$$S_a \models (\exists y)(\forall x)[A(x) \rightarrow x=y]$$

Note that whenenever $A(x)$ is of form $x=c$, where c is a constant (or more generally, a variable-free term), the answer is definite (but not necessarily informative, as we shall see).

While it is clear that these properties of answers are desirable, it will also become clear, first, that they may not be obtainable in general and second, that they are insufficient to ensure that answers will be informative. We address each of these points in turn.

5. Relative Completeness (or Definiteness)

In general, the knowledge of the question answerer may not be sufficient to derive a categorial answer which is complete (and also nontrivial -- a point we take up in the next section). We therefore require a formal criterion for determining whether the answer is as complete as the available knowledge permits.

Accordingly, we say that a categorial answer is *relatively complete* iff for all formulas $B(x)$ (of the object language of the question answerer) containing x as the only free variable,

$$\text{if } S_a \models (\forall x)[B(x) \rightarrow R(x)]$$

[1] A false answer which follows from mistaken beliefs on the part of the question answerer may thus be deemed correct; but this "defect" is unavoidable in any *useful* notion of correctness.

[2] Any reader who feels that this answer is still peculiar is reminded that there can be no guarantee that the question-answerer's knowledge will include the fact that Alberta is not a country.

$$\text{then } S_a \models (\forall x)[B(x) \rightarrow A(x)],$$

i.e., A is at least as general as any derivable correct answer.

Analogously, we say that a selective answer is *relatively definite* iff it is as specific as the question answerer's knowledge permits (i.e., any further constraint the question answerer is able to add is already implicit in the meaning of $A(x)$ and $R(x)$); or formally, iff for all formulas $B(x)$ containing x as the only free variable,

$$\text{if } S_a \models (\exists x)[A(x) \wedge B(x) \wedge R(x)]$$

$$\text{then } MP_a \models (\forall x)[(A(x) \wedge R(x)) \rightarrow B(x)],$$

where MP_a (a subset of S_a) is the set of *meaning postulates* available to the question answerer. (These are axioms expressing such facts as "All bachelors are unmarried", "The territory lying within the borders of any country belongs to that country", i.e., facts which hold solely in virtue of the conventional meanings of the terms they involve.)[3]

6. Nontriviality and Comprehensibility

The insufficiency of the correctness and completeness conditions for categorial answers is evident from our earlier observation that they amount to *equivalence* of R and A. Thus an answer of form

$$(\forall x: R(x))Q(x)$$

is always correct and complete, but is clearly useless (e.g., "All countries that have part of the Rocky Mountains in them have part of the Rocky Mountains in them".)

For an answer to be ***nontrivial,*** we would like to make sure that the questioner does not yet ***know*** that answer, i.e.,

$$\sim K_q(\alpha x: A(x))Q(x)$$

where we interpret "knowing a fact" as being able to verify it effectively (and presumably, quickly) by application of a "deductive recall" algorithm K to a base of facts S_q ***explicitly*** known to the questioner (cf. [Konolige 85]). This might be written as follows. K is a (quickly) computable function from pairs S, ϕ, where S is a set of sentences and ϕ is a sentence, to {T,F,U} (meaning "true", "false", and "unknown" respectively), where

$$K(S,\phi) = \begin{cases} T \text{ only if } S \models \phi \\ F \text{ only if } S \models \sim\phi \\ U \text{ only if neither } \phi \text{ nor } \sim\phi \text{ is in } S.^4 \end{cases}$$

In terms of K, K_q is defined as

$$K_q(\phi) \quad \text{iff} \quad K(S_q,\phi) = T.$$

The preceding criterion is still too weak, however. For example, the "Rocky Mountain" question might be answered with "Canada, and all other countries that have part of the Rocky Mountains in them." This is nontrivial if the questioner knows nothing about the Rocky Mountains, yet intuitively seems "partially trivial"; this is because there is a potentially less comprehensive answer than the one given, from which the questioner could easily infer the one given.

We therefore define a categorial answer to be *entirely* nontrivial iff there does not exist a correct categorial answer with answer predicate A' such that

$$S_a \not\models (\forall x)[A(x) \rightarrow A'(x)] \text{ and}$$

$$K_q[(\forall x: A'(x))Q(x) \rightarrow (\forall x: A(x))Q(x)]$$

This criterion is violated in the above example, with

$$A'(x) =_{df} [x = \text{Canada}],$$

$$A(x) =_{df} [x = \text{Canada}] \vee [\sim [x = \text{Canada}] \wedge Q(x)].$$

In a closely related sense, the notion of partial triviality also applies to selective answers, as in "A North or South American country with part of the Rocky Mountains in it". We therefore say that $(\exists x: A(x))Q(x)$ is an ***entirely nontrivial selective answer*** iff there does not exist a potentially less definite correct answer from which the questioner could easily infer the answer given; or formally, iff there does not exist a correct selective answer with answer predicate A' such that

$$S_a \not\models (\forall x)[A'(x) \rightarrow A(x)] \text{ and}$$

$$K_q[(\exists x: A'(x))Q(x) \rightarrow (\exists x: A(x))Q(x)].$$

This criterion is violated in the above example, with

$$A'(x) =_{df} [NAC(x) \vee SAC(x)],$$

$$A(x) =_{df} [NAC(x) \vee SAC(x)] \wedge Q(x),$$

where NAC and SAC mean "is a North American country" and "is a South American country" respectively.

Our definition of "Knows" involved a set S_q of axioms known to the questioner and an algorithm K which tries to determine whether a sentence lies in the logical closure of S_q. In a question-answering system, a good starting point for S_q might be the meaning postulates of the language, especially when no question context or user model is available. For this case, we can define an *analytic* answer to be one easily deducible from meaning postulates.

For example, in response to "Who is a bachelor?", the answer "An unmarried man " would be considered analytic if "A bachelor is an unmarried man " was assumed to be a meaning postulate.

However, this example also points out the fallibility of this approach: if the questioner did not know this meaning postulate and wanted to know "What is a bachelor?" then the response "An unmarried man" would be nontrivial.

Conversely, a non-analytic answer may be trivial. Consider the categorial version of the "Rocky Mountain" question. The response "All countries which maintain parks in the Rocky Mountains" is correct, non-analytic and complete but will be trivial if, for example, the questioner already knows that a country maintains parks in a mountain range only if the mountain range lies partially within its borders. (Note that this is contingent knowledge, rather than knowledge of meaning postulates.)

As Hintikka noted [Hintikka 76] it is desirable for the term designated by the answerer to be "known" to the questioner for the answer to be comprehensible. We will impose the less stringent requirement that an answer should be *comprehensible* in the sense that it contains no "private symbols". A "private symbol" may be defined as any constant symbol, function symbol, or predicate symbol of the answer formalism which has no predefined English equivalent. For a predicate logic question answering system, we can assume

[3] Our condition of relative definiteness is perhaps too strong. For example, if a system possesses the axioms $(\exists x: A(x))R(x)$ and $(\exists y: A(y)\ \&\ B(y))R(y)$, where A and B are unrelated predicates, a selective answer based on the first of these axioms will be deemed non-relatively definite. Yet, it is in a sense the most definite answer the system can give "for the x referenced by the first axiom", since "the y referenced by the second axiom" may be an entirely distinct entity. However, weakening the criterion of relative definiteness to allow for this observation appears to be impossible without adding modal operators to the object language.

[4] This is only a partial specification of K, of course. Various other reasonable constraints could be imposed on K, such as monotonicity: if $K(S,\phi) \neq U$ then for all ψ, $K(S \cup \{\psi\},\phi) \neq U$.

that all predicate symbols are "public" (non-private). However, skolem constants and functions created when translating natural language sentences into quantifier-free form will often be private.[5]

Finally, we define an *informative* answer to be one that is nontrivial and comprehensible.

We do not claim that correctness, completeness (or definiteness) and informativeness in the sense in which we have defined these properties provide either *necessary* or *sufficient* conditions for an answer to an information request to be helpful or appropriate. What we do claim is that these properties are usually *desirable* in an answer, regardless of the context in which the information is requested.

We now attempt to evaluate the theoretical adequacy of some standard deductive question answering methods with respect to our characterization of answers.

7. A Reexamination of Deductive Answer Extraction

The problem of modifying a theorem-prover to generate answers to wh-questions was considered by Cordell Green [Green 69] in his QA3 system. Green's method consisted of formulating an existential presupposition of the question in quantifier-free form, denying the presupposition, and then using a resolution theorem prover to refute the denial. An ANS predicate was then appended to the denial of the theorem to collect the composition of substitutions performed on the variables corresponding to the existentially quantified variables of the theorem. When the proof was repeated, the ANS predicate would then have substituted for its variable a term which could in some sense be considered as an answer to the original question.

Luckham and Nilsson [Luckham & Nilsson 71] proposed an extension to Green's method that produced a sentential answer, rather than just a substitution. The answer formula was either similar to the entire question formula or to one of its disjuncts (see below). They also provided a theoretical justification for both methods, showing that the answer was a logical consequence of the axioms, and that the theorem was a logical consequence of the answer.

To what extent do answers provided by the Green-Luckham-Nilsson procedure conform with the desiderata we have proposed? The first and perhaps most important point to note is that the procedure is generally unsuitable for deriving categorial answers. To take a very simple example, a system which knows nothing except the fact that all creatures are mortal should be able to supply a complete categorial answer to the question, "What creatures are mortal?". But the Green-Luckham-Nilsson procedure, which would attempt to prove that some creature is mortal, would fail to obtain an answer.

Of course, the answer predicate of any correct, definite, nontrivial selective answer also provides an answer predicate for a correct, nontrivial categorial answer. However, even in cases where both selective answers and categorial answers are entailed by the knowledge base, no *definite* selective answer may be derivable from it. Besides, a categorial answer based on a selective answer will not in general be complete or even relatively complete. Within the prolog community, there appears to be a general assumption that complete answers to questions, though not obtainable in general with any one proof, can be obtained in principle by enumerating proofs and collecting (selective) answers obtained by different proofs. However, it is clear from the preceding remarks that this assumption is mistaken. Also, a categorial answer which enumerates instances may be unnecessarily verbose. (For some ideas on compressing enumerative answers, see [Kalita et al. 84].) To the best of our knowledge, the problem of computing categorial answers is wide open at this time.

It is worth noting that a procedure is easily formulated for computing complete, informative answers which is as effective as we can hope -- though it is hopelessly inefficient. The idea is to syntactically enumerate possible comprehensible answers, and to interleave attempts to prove these possible answers and their completeness and nontriviality with the enumeration itself. (To establish nontriviality, we must of course have a model of the questioner that allows us to simulate K operating on S_q.) If a complete, informative answer is implicit in the knowledge base, this procedure will eventually find one.[6] It is an open problem whether there is also a semi-decision procedure of this type for relatively complete (and/or entirely nontrivial), comprehensible answers.

Let us now look more closely at the Green-Luckham-Nilsson procedure, and try to assess its performance with respect to selective answers. Certainly the answers are correct (in view of the theorems mentioned above), but are they relatively definite and informative?

To expect that *every* proof would yield a selective answer with these properties would be to expect too much. For example, if a knowledge base contains separate assertions to the effect that "Someone knows prolog" and "John knows prolog", either assertion can be used to prove that someone (or something) knows prolog; but the former does not produce a relatively definite, informative answer while the latter does. The best we can expect, therefore, is that *some* proof or combination of proofs will allow extraction of a relatively definite, informative answer.

The answers produced by Luckham and Nilsson's procedure are restricted to certain syntactic forms bearing a specific relation to the question posed as a theorem. If the theorem is written in quantifier-free disjunctive normal form

$$\sim C_1(X(1), G(1)) \vee \cdots \vee \sim C_n(X(n), G(n))$$

where the $X(i)$ correspond to the universally quantified variables and the $G(i)$ correspond to the skolem functions representing the existentially quantified variables in the ith clause (conjunction of literals), then the answer would be of form

$$\sim C_1(X(1), Y(1)\theta(1,r)) \vee \cdots \vee \sim C_n(X(n), Y(n)\theta(n,r))$$

where $\theta(i,r)$ denoted the composition of substitutions in the resolution proof tree on paths from the root node to the leaf node corresponding to C_i (where the clause at that leaf node is the dual of C_i obtained when the theorem is denied and converted to conjunctive normal form.)

Because of this syntactic restriction, the only types of answers that can be generated are sentences "echoing"

[5] Though they need not be. E.g., consider "There is a girl and her name is Mary". The skolem constant created in the first clause is assigned the English translation "Mary" by the second clause.

[6] This procedure is impractical in the same way that syntactic enumeration procedures for *hypothesis generation* are impractical. Indeed, categorial question answering can be viewed as a kind of hypothesis generation process: the goal is to discover a property A whose possession (nontrivially) entails a given property R. It appears that methods like Morgan's [Morgan 71] or Pietrzykowski's [Pietrzykowski 78], which treat hypothesis formation as the dual of deduction, can be adapted to our task. In the "mortal creatures" problem, for example, to determine what properties entail being mortal, we would determine what properties are *entailed by not* being mortal, and use the negation of such properties as possible categorial answers. In particular, if we predicate "not mortal" of c (a skolem constant), we immediately obtain "not a creature" for c, from the premise that all creatures are mortal, so that the negation of this property, "creature", becomes a possible categorial answer, as required. Formalization of this method may well yield an effective method for categorial question answering.

portions of the question and containing terms in the Herbrand universe satisfying the question predicate. The answer cannot be phrased in terms of other predicates, though some additional information can be derived by tracing the skolem functions back to the clauses from which they arose. This can be seen in the following example:

Example

Assume that the axioms are "All creatures are mortal", "There is a creature", and the question is "Who (or what) is mortal?".

To answer this question, let $M(x)$ denote that x is mortal, and $C(x)$ denote that x is a creature. The denial of the presupposition of the question is "Nothing is mortal". Then the axioms can be written in clause form as

$$\sim C(x) \lor M(x),\ C(c),\ \text{and} \sim M(x)$$

where c is a skolem constant. The Luckham-Nilsson method would then generate the answer $M(c)$ which means "Something is mortal". This can be viewed as a selective answer, with answer predicate $\lambda x[x=c]$. The answer is correct and definite but, if c is not a public constant, is not comprehensible. Here a "deskolemization" procedure due to Cox and Pietrzykowski [Cox & Pietrzykowski 84] could be used to eliminate c in favour of an existentially quantified variable, but then the result will no longer be definite. (In fact, in this instance there can be no definite, comprehensible answer, since no uniquely identifying characteristics for c are available.)

Green was aware of this problem and attempted to repair the difficulty by supplementing the "formal" answer with additional information. For the previous example, the answer predicate $ANS(c)$ would be supplemented by the clause in which the constant c was created, $C(c)$, or perhaps the English language version of the clause, "There is a creature."

Another example from [Chang & Lee 73] further illustrates the sorts of answers obtained by the Green-Luckham-Nilsson method, and the problem of ensuring that answers will be comprehensible.

Example

Assume that the answerer knows:

"Everyone who entered this country and was not a VIP was searched by a customs official."
"William was a drug pusher."
"William entered this country."
"William was searched by drug pushers only."
"No drug pusher was a VIP."

Suppose the answerer is asked "Who was both a drug pusher and a customs official?" To answer this question, let $S(x,y)$ denote x was searched by y, $V(x)$ denote x is a VIP, $E(x)$ denote x entered the country, $C(x)$ denote x is a customs official, and a denote William. The answer produced by the Luckham-Nilsson method would be

$$D(f(a)) \land C(f(a))$$

which means "$f(a)$ is a drug pusher and a customs official" Unfortunately, the questioner might have no idea who is $f(a)$.

If Green's method is used, the answer $ANS(f(a))$ would be generated. In response to a request for further information, the answerer might provide "Everyone who entered this country and was not a VIP was searched by a customs official." This is insufficient to determine a relatively definite, comprehensible answer. The answer we desire is

$$(\exists x\colon S(a,x) \land \sim V(x))D(x) \land C(x),$$

i.e., someone who searched William and is not a VIP is a drug pusher and a customs official (where we have assumed that none of the axioms are meaning postulates). Thus the clauses that ought to have been added to the answer extracted from the proof are $S(a,f(a))$ and $\sim V(f(a))$; from these the formula $A(x)$ $(=S(a,x) \land \sim V(x))$ needed for a relatively definite answer could have been derived, with the aid of a de-skolemization procedure. (A more general algorithm would be required than that in [Cox & Pietrzykowski 84] which processes individual literals only.)

Does Green's method of supplementing answers at least guarantee that *some* proof of any given question will yield a relatively definite, comprehensible answer? Unfortunately, the answer is negative, as the following very simple counterexample shows. Suppose that the question answerer possesses just two facts, namely, that Agatha was murdered by someone and that only a person who drives a dumptruck could have murdered Agatha:

$$M(c,a),\ \sim M(x,a) \lor D(x)$$

where c is a skolem constant. Then the *only* proof that answers the question "Who (or what) murdered Agatha?" is one that resolves the denial clause $\sim M(x,a)$ against the first of the two axioms. The fact that is needed, however, to flesh out the answer $M(c,a)$ to one that is relatively definite is $D(c)$ (or, more redundantly, $\sim M(c,a) \lor D(c)$). This fact is not a substitution instance of any clause in any proof, or of a base clause involving a private symbol that occurs in the answer.

8. Conclusion

We have introduced a set of formal concepts for assessing the quality of answers returned by deductive question answering procedures. In particular, we have formalized some desirable properties of two natural classes of answers, which we termed "categorial" and "selective". We have shown that standard question answering methods are in general unable to produce categorial answers, and that they also fall short in the kinds of selective answers they produce.

We have mentioned some open problems, the most important of which is the formulation of reasonably efficient procedures for categorial question answering. In general, our work indicates that there are unsuspected depths, and many unexplored issues, in the area of deductive question answering.

References

Aqvist 65.
Lennart Aqvist, *A new approach to the logical theory of interrogatives*, University of Uppsala, 1965.

Belnap & Steel 76.
Nuel D. Belnap and Thomas B. Steel, *The logic of questions and answers*, Yale University Press, New Haven, 1976.

Bennett 79.
Michael Bennett, *Questions in Montague Grammar*, Indiana University Linguistics Club, Bloomington, Indiana, 1979.

Chang & Lee 73.
Chin-liang Chang and Richard Char-tung Lee, *Symbolic Logic and Mechanical Theorem Proving*, Academic Press, New York, 1973.

Cohen & Levesque 85.
Philip R. Cohen and Hector J. Levesque, Speech acts and rationality, *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, 1985,

49-60.

Cox & Pietrzykowski 84.
P. T. Cox and T. Pietrzykowski, A complete, nonredundant algorithm for reversed skolemization, *Journal of Theoretical Computer Science 28*, (1984), 239-261, North-Holland.

Green 69.
Cordell Green, Theorem-proving by resolution as a basis for question-answering systems, in *Machine Intelligence*, vol. 4, Bernard Meltzer, Donald Michie and Michael Swann (ed.), Elsevier, New York, 1969, 183-205.

Hintikka 76.
Jaakko Hintikka, *The Semantics of Questions and the Questions of Semantics*, North-Holland, Amsterdam, 1976.

Kalita et al. 84.
J. K. Kalita, M. J. Colbourn and G. I. McCalla, A response to the need for summary responses, *Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics*, 1984, 432-436.

Konolige 85.
Kurt Konolige, A computational theory of belief introspection, *IJCAI-85 1*, (1985), 502-508.

Levesque 84.
Hector J. Levesque, Foundations of a functional approach to knowledge representation, *Artificial Intelligence 23*, (1984), 155-212.

Luckham & Nilsson 71.
David Luckham and Nils J. Nilsson, Extracting information from resolution proof trees, *Artificial Intelligence 2*, (1971), 27-54.

Morgan 71.
Charles G. Morgan, Hypothesis generation by machine, *Artificial Intelligence 2*, (1971), 179-187.

Pietrzykowski 78.
T. Pietrzykowski, Mechanical hypothesis formation, Res. Rept. CS-78-33, Department of Computer Science, University of Waterloo, Canada, 1978.

Webber & Mays 83.
Bonnie Lynn Webber and Eric Mays, Varieties of user misconceptions: detection and correction, *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, 1983, 650-652.