The Semantics of Questions

Working Note 36
Revised March 22th, 2010
Peter Clark, peter.e.clark@boeing.com

Abstract

AURA is a knowledge-based system designed to answer exam-style science questions posed in English. To date, we have only given a semi-formal account of the semantics of the questions that AURA attempts to answer. In particular, AURA's question interpreter does not output a full first-order logic interpretation of the question, but instead outputs a set of ground (Skolemized) binary predicates, and thus the placement of quantifiers and implications remains unclear. This document is step towards rectifying this, by specifying the full semantics for some of the question types that AURA handles, and discussing some of the issues involved. We do not address the natural language processing issues (how to generate those semantics from the English), nor reasoning issues (how to answer the question given its semantics), these are out of scope of this document. The analysis presented here is thorough but incomplete, and largely example-driven with both details and general frameworks missing in places. However, it offers some important insights into the topic of question semantics, in particular on the role of presuppositions in questions and the relationship between universal and hypothetical questions.

1. Introduction

AURA is a knowledge-based system designed to answer exam-style science questions posed in English (Barker et al., 2004; Clark et al., 2007; Gunning et al., 2010). AURA contains knowledge bases for (areas of) physics, chemistry, and biology, and for our current research biology is the primary domain of interest. To date, we have only given a semi-formal account of the semantics of the questions that AURA attempts to answer. In particular, AURA's question interpreter does not output a full first-order logic interpretation of the question, but instead outputs only a semi-formal representation, namely a set of ground (Skolemized) binary predicates, and thus the presence of quantifiers and implications remains unclear. This document is step towards rectifying this, by specifying the full semantics for some of the question types that AURA handles, and discussing some of the issues involved in the formalization process.

Questions vary substantially in both the way they are expressed in English and the answers that they are intended to elicit. We first attempt to characterize this variation by identifying several different question types, describing some of the ways questions of those types are posed in English, discussing what the intended semantics are, and characterizing those semantics in logic. Given such a characterization, a deductive reasoning engine could then answer those questions, although the deductive process of answering questions is outside the scope of this paper.

Questions can be categorized along various dimensions:

- By the syntactic structure of information they return (value, set, list, table, clauses, ...)
- By the semantic type of information returned (person, date, company, definition,...). This dimension has been particularly useful for information retrieval, e.g., Webclopedia's QA typology (Hovy et al., 2001).
- By the type of reasoning required to answer them (solving equations, constraint reasoning, subsumption, ...)
- As a set of general or domain-specific templates, e.g., What is a X? What is the cause of X? (e.g., Clark et al. 2003, Cucina et al., 2001)

For our purposes, as we wish to answer questions using a deductive reasoner, we organize them according to the type of expression that the reasoner is given (described later in this document). From that perspective, the following categorization into 12 different types is useful:

"Primitive" questions:

1. true/false - "Does a cell contain DNA?"
2. find a value (or values)[1] - "What does a cell contain?", "How heavy is a cell?"
3. subsumption – "Is a eukaryotic cell a cell with a nucleus?"
4. cardinality - "How many nuclei are in a cell?"
5. taxonomic - "What types of cell are there?", "What type of thing is a ribosome?"
6. possibility - "Can an adenine bond with cytocine?"
7. reflective – "Does a cell's size affect its speed of division?"

"Composite" questions:

8. definition - "What is a cell?"
9. description – "What happens during prophase?"
10. example - "What is an example of a bond?"
11. similarity - "What are the similarities/differences between a eukaryotic cell and a prokaryotic cell?"
12. relationship - "What is the relationship between ribosomes and protein synthesis?"

We have loosely divided these into two categories, "primitive" and "composite"[2]. The idea is that primitive questions are answered by a direct query to the KB, while composite questions require assembling an answer from potentially multiple queries to the KB. The primary issue for primitive questions is one of formal semantics, namely what is the formal query that should be issued to the KB to answer the question – this is the issue of concern in this paper. The primary issue for composite questions is more of human psychology, namely what constitutes an adequate answer to the question, perhaps assembled from answers to several primitive queries to the KB (e.g., Lester and Porter 1996, 1997). These two issues are quite different. For our purposes here, we only discuss the first of these issues, namely the semantics of primitive questions, and treat "composite" questions as out of scope.

We treat the semantics of a question as a formal specification of the question's answer. We do not discuss how that specification might then be evaluated (proved), rather assume that we have

[1] True/false and find-a-value are intended to denote "normal" true/false and find-a-value, i.e., excluding the special cases listed in subsequent categories such as cardinality, possibility, etc. (strictly every question is some kind of true/false or find-a-value in their most general sense, but that is not a very helpful perspective)

[2] The names "primitive" and "composite" aren't the best, but they will do for now.

a purely deductive reasoning engine available that can perform this task. We also assume that question answering is a purely deductive process, and are agnostic about the reasoning engine that is used for that purpose.

The semantics of questions have been studied extensively in the linguistics literature, e.g., (Karttunen, 1977; Chierchia, 1993). However, from my (limited) reading to date, there is somewhat of a mismatch between the linguistic literature and our needs here. The linguistics literature delves into some very complex questions, for example involving complex quantification, coreference, and modal behavior (e.g., "Who killed whom?", "Can you guess which crimes the FBI doesn't know how to solve?"), typically more complex than those that are of concern in AURA. Conversely, some of the semantic issues that have presented themselves within AURA, in particular the role of presuppositions in questions, do not appear to have been given much attention in the linguistics literature on question semantics (at least to my knowledge). Further digging is needed to better align the prior work in linguistics and the analyses presented in this document.

In this document, we go into some depth analyzing true/false questions, and comparing formulations expressed as universals and expressed as hypotheticals as these are common in AURA. Analyses for other question types then largely follow the same pattern. We then offer some thoughts on the notion of "partial answers" – or more precisely full answers to specialized questions – that may be helpful to the user in the case when the answer to the original answer is "no" (for true/false questions) or "nothing" (for find-a-value questions). The analysis is largely example-driven, mainly looking at subtle variants of asking about the contents of a cell's nucleus, and some broader generalizations are still missing. However, even properly capturing the semantics of these simpler questions poses some challenges, which (we hope) are adequately described and answered by this document.

2. Foundations

2.1 Class Terminology

Before starting, we define some terminology. A class is an entity with an associated definition and members. A definition is a formula F(x) with free variable x. An instance x satisfies a definition F(x) if F(x) is true. The class's members are all those instances satisfying the class's definition F(x). An instance is in a class if it is a member of the class. A class may or may not be noted by a symbol C in the KB, i.e., the class may or may not be reified. For reified classes, we denote class membership in the class C defined by $F_C(x)$ using the predicate isa(x,C), i.e., $\forall x$ $F_C(x) \leftrightarrow$ isa(x,C). The definition $F_C(x)$ is the intension of the class C. The set $\{ x \mid F_C(x) \}$ is the extension of the class C. We use is-subclass-of(x,y) to denote the class/subclass relationship, i.e., if $\forall x$ isa(x,C) $\rightarrow$ isa(x,D) then is-subclass-of(C,D) is true, and vice versa.

2.2 Presuppositions of Uniqueness

An important concept we will need later is the notion of *presuppositions* in questions. A presupposition is an implicit statement about the world, make in the question. For example. consider the question:

(1) Does the nucleus of every eukaryotic cell contain DNA?

We might write its semantics[3,4] as asking:

"For all eukaryotic cells, does it have a nucleus containing DNA?"

(2) is-it-true[$\forall$x isa(x,EukCell) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z)]

where is-it-true[] is a valuation function mapping a formula to its truth value. However, as well as asking a question, a normal reading of (1) *presupposes* that every eukaryotic cell has exactly one nucleus (expressed by the "the" determiner of "the nucleus"). We can express this presupposition as:

"Every eukaryotic cell has exactly one nucleus"

$\dot{\forall}$x isa(x,EukCell) $\dot{\rightarrow}$ $\exists$!y isa(y,Nuc) $\wedge$ has-part(x,y) (UE)

where $\exists$! denotes unique existence, i.e., .

$\exists$!x F(x) means $\exists$x F(x) $\wedge$ ($\dot{\forall}$y F(y) $\rightarrow$ y = x)

where F(x) is a formula with free variable x. In this case, (UE) expands to

"Every eukaryotic cell has exactly one nucleus"

(3) is-it-true[$\dot{\forall}$x isa(x,EukCell) $\dot{\rightarrow}$ $\exists$y isa(y,Nuc) $\wedge$ has-part(x,y) $\wedge$ ($\dot{\forall}$z has-part(x,z) $\rightarrow$ y = z)]

In other words, (1) is not just asking a question but also implicitly implying something about the world (which may or may not be represented in the KB), namely that every eukaryotic cell has exactly one nucleus. Thus, a complete specification of the semantics of (1) would be:

"Given every cell has exactly one nucleus, does that nucleus contain DNA?"

(4) is-it-true[{ KB $\cup$ UE } $\vdash$
$\forall$x isa(x,EukCell) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z)]

What happens if the question's presupposition (UE) is wrong, e.g., some eukaryotic cells have zero nuclei, or two nuclei? Formally, the presupposition would introduce a contradiction into the KB (if the KB represented the contrary facts). For such questions, we will say the question is nonsensical and the answer is undefined.

2.3 Presuppositions of Existence

A second aspect of questions that we need to discuss concerns the implied/suggested existence of (at least one) member of a class. Consider these two question variants:

(5a) Do all the nuclei of eukaryotic cells contain DNA?
(5b) Do all the nuclei of prokaryotic cells contain DNA?

A straightforward encoding of the semantics might be as follows:

(6a) is-it-true[$\forall$xy isa(x,EukCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$ $\exists$z isa(z,DNA) $\wedge$ encloses(y,z)]
(6b) is-it-true[$\forall$xy isa(x,ProkCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$ $\exists$z isa(z,DNA) $\wedge$ encloses(y,z)]

This might seem okay, but there is something not quite right still: Biologically speaking, there are no prokaryotic cells with nuclei, so what should the answer be to:

[3] We assume the linguistic processing (word sense disambiguation, semantic role labeling, etc.) is performed correctly, e.g., assume the NLP engine correctly interprets "of" as has-part(x,y). The NLP processing itself is outside the scope of this paper.
[4] For simplicity we treat DNA here as an object, although biologically speaking it is more like a substance.

(5b) Do all the nuclei of prokaryotic cells contain DNA?

Formally (6b) evaluates to true as the antecedent of the implication is always false. However, it seems a little unintuitive that the answer to (5b) is "yes". Or consider the variant:

(5c) Do all the nuclei of prokaryotic cells contain chocolate?

(6c) is-it-true[$\forall$xy isa(x,ProkCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$
$\exists$z isa(z,Chocolate) $\wedge$ encloses(y,z)]

It is even more unintuitive that the answer to (5c) is "yes", although this is logically implied by (6c). This suggests that (6c) does not fully capture a natural reading of (5c), and likewise (6b) does not fully capture a natural reading of (5b).

It might seem that we need a modal operator of necessity $\Box$, i.e., (5b) should be read as asking something like:

(5b') Do or would the nuclei of prokaryotic cells, if there were any, necessarily contain DNA?

In other words, although the KB may not know of any prokaryotic cells with nuclei, if there were such cells (in some other possible world), would their nuclei contain DNA? This seems okay if the KB is agnostic about the existence of such cells. However, suppose the KB explicitly represents that prokaryotic cells have no nuclei? In this case, there would not be any such possible worlds, i.e., (5b) and (5c) would still be true for all possible worlds, and thus a modal treatment would still answer "yes" to them, hence still not capturing our intuition.

A perhaps better treatment of (5a) - (5c) is to say that they, too, come with a presupposition about the world, namely that such cells can exist. We could represent this presupposition in two alternative ways, namely using a modal statement of possibility ($\Diamond$):

"It is possible that there is at least one prokaryotic cell with a nucleus"

$\Diamond$ $\exists$xy isa(x,ProkCell)$\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) (PE)

or (our preference here, as it avoids modal logic) simply a presupposition of existence:

"There is at least one prokaryotic cell with a nucleus"

$\exists$xy isa(x,ProkCell)$\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) (E)

(PE) states that at least one such cell exists in some possible world, while (E) simply states that at least one such cell exists (or, in a modal formalism, exists in all possible worlds). If we take E as the appropriate presupposition, then we can express the complete semantics of (5b) as:

"Given there are at least some prokaryotic cells with nuclei, do all their nuclei contain DNA?"

(6b') is-it-true[{ KB $\cup$ E } $\vdash$
$\forall$xy isa(x,ProkCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$ $\exists$z isa(z,DNA) $\wedge$ encloses(y,z)]

Again, what happens if the question's presupposition (E) is wrong, and the KB states that prokaryotic cells do not have a nucleus? Again, formally, the presupposition would introduce a contradiction into the KB. For such questions, we again say the question is nonsensical and the answer is not defined. Intuitively, this behavior seems the same as the behavior a human expert would exhibit when answering (5b) and (5c), namely to report that the presupposition is invalid, and thus that the question cannot be answered.

3. The Semantics of "Primitive" Questions

Armed with these foundations, we now discuss the semantics of the seven primitive question types, mainly discussing questions of the first two types (true/false and find-a-value).

3.1 True/False Questions

For our analysis, it is useful to consider two types of question that occur frequently in AURA:

Universals: Questions about properties that hold for all individuals with a particular description.
Hypotheticals: Questions about properties that hold for some individual in a hypothetical situation.

For example, consider the two questions below:

(7) Does every eukaryotic cell have a nucleus containing DNA?
(8) A eukaryotic cell has a nucleus. Does that nucleus contain DNA?

(7) is most naturally read as a ***universal question***, i.e., about all members of a class, here eukaryotic cell. (8) is most naturally read as a ***hypothetical question***, as it posits a world in which some eukaryotic cell with a nucleus exists.

The relationship between the two is important in AURA because users often reformulate questions like (7) as questions like (8). If they do so, are they still asking the same question? In addition, some questions are ambiguous about whether a universal or hypothetical reading is most appropriate, for example:

(9) Does the nucleus of a eukaryotic cell contain DNA?

(9) can be interpreted as a universal question, if "a eukaryotic cell" is interpreted as a generic (i.e., no specific eukarotic cell). Alternatively, it could be interpreted as a hypothetical, if it is read a positing some specific eukaryotic cell. Does the choice matter? The following analysis provides a semantics for these types of questions, helping to answer these issues of equivalence.

3.1.1 Universals

Building on our earlier analysis in Section 2, we will represent English questions presupposing uniqueness (e.g., "the nucleus") by using a uniqueness presupposition formula in the semantics, and English statements of universal quantification ("all nuclei") that presuppose existence using an existence presupposition formula in the semantics. If a presupposition is false then we say the answer to the question is undefined.

Thus equipped, let's consider three similar universal questions:

(10a) Does every eukaryotic cell have a nucleus containing DNA?
(10b) Does the nucleus of every eukaryotic cell contain DNA?
(10c) Do all the nuclei of eukaryotic cells contain DNA?

Using our analysis in Section 2, we can express the semantics of each as:

"Does every eukaryotic cell have a nucleus containing DNA?" (10a)

(11a) is-it-true[∀x isa(x,EukCell) → ∃yz isa(y,Nuc) ∧ isa(z,DNA) ∧ has-part(x,y) ∧ encloses(y,z)]

"Does the nucleus of every eukaryotic cell contain DNA?" (10b)
or literally: "Given every euk. cell has one nucleus, does that nucleus contain DNA?"

(11b) is-it-true[{ KB ∪ UE } ⊢
∀xy isa(x,EukCell) ∧ isa(y,Nuc) ∧ has-part(x,y) → ∃z isa(z,DNA) ∧ encloses(y,z)]

"Do all the nuclei of eukaryotic cells contain DNA?" (10c)
or literally: "Given some euk. cells have a nucleus, do all those nuclei all contain DNA?"

(11c) is-it-true[{ KB ∪ E } ⊢
∀xy isa(x,EukCell) ∧ isa(y,Nuc) ∧ has-part(x,y) → ∃z isa(z,DNA) ∧ encloses(y,z)]

Where UE and E are defined:

∀x isa(x,EukCell) →. ∃!y isa(y,Nuc) ∧ has-part(x,y) (UE) *"exactly 1 nucleus"*
∃xy isa(x,EukCell) ∧ isa(y,Nuc) ∧ has-part(x,y) (E) *"some cell has a nucleus"*

Note the different stances the questions take on whether eukaryotic cells have a nucleus: (11a) asks if it is always true; (11b) presupposes that it is always true; and (11c) presupposes that it is sometimes true.

With a bit of formula manipulation, we[5] can show that (11b) is equivalent to:

(11b') is-it-true[{ KB ∪ UE } ⊢
∀x isa(x,EukCell) → ∃yz isa(y,Nuc) ∧ isa(z,DNA) ∧ has-part(x,y) ∧ encloses(y,z)]

i.e., (11b) ↔ (11b'), i.e., it is sufficient to prove either of these to answer "yes" to the question.

If we assume that the user does not ask questions with false presuppositions, then we can write the semantics as follows:

"Does every eukaryotic cell have a nucleus containing DNA?" (10a)

(12a) is-it-true[∀x isa(x,EukCell) → ∃yz isa(y,Nuc) ∧ isa(z,DNA) ∧ has-part(x,y) ∧ encloses(y,z)]

"Does the nucleus of every eukaryotic cell contain DNA?" (10b)

(12b) is-it-true[∀x isa(x,EukCell) → ∃yz isa(y,Nuc) ∧ isa(z,DNA) ∧ has-part(x,y) ∧ encloses(y,z)]

"Do all the nuclei of eukaryotic cells contain DNA?" (10c)

(12c) is-it-true[∀xy isa(x,EukCell) ∧ isa(y,Nuc) ∧ has-part(x,y) → ∃z isa(z,DNA) ∧ encloses(y,z)]

Note that although (12a) and (12b) are the same, this equivalence in semantics (between (10a) and (10b)) only holds if UE (presupposed by (10b), but not (10a)) is true. If it is not, then (10b) is nonsensical and its answer is undefined.

(12a) and (12b) universally quantify over a single class, while (12c) universally quantifies over multiple classes. In practice, questions requiring quantification over multiple classes are rare in the AP biology exams.

Finally let us consider four KBs, that represent different information as follows (some is biologically incorrect, but ignore that for now):

[5] or more precisely Michael Kifer (thank you!)

KB1: Every eukaryotic cell has exactly one nucleus containing DNA.
KB2: Every eukaryotic cell has exactly two nuclei, one containing DNA and one not.
KB3: Some eukaryotic cells have exactly one nucleus containing DNA, and the rest have no nucleus.
KB4: No eukaryotic cells have a nucleus.

We can summarize the answers that these KBs will give for the three questions in a table based on our semantics:

	(10a)	**(10b)**	**(10c)**
KB1	yes	yes	yes
KB2	yes	undefined	yes
KB3	no	undefined	yes
KB4	no	undefined	undefined

3.1.2 Hypotheticals

The Semantics of Hypotheticals
Hypotheticals posit a world in which certain objects exist, and then ask about the properties of those objects. Let us consider the hypothetical question introduced earlier:

(8) A eukaryotic cell has a nucleus. Does that nucleus contain DNA?

If we denote the hypothesized objects by constants X0 and Y0 (constants not in the KB), then we can express the semantics of (8) by considering the KB augmented with the hypothesized facts:

"A eukaryotic cell has a nucleus. Does that nucleus contain DNA?" (8)
or literally "Given there's a euk. cell with a nucleus, does that nucleus contain DNA?"
(13) is-it-true[KB $\cup$ { isa(X0,EukCell) $\wedge$ isa(Y0,Nuc) $\wedge$ has-part(X0,Y0) } $\vdash$
$\exists$z isa(z,DNA) $\wedge$ encloses(Y0,z)]

which can be rewritten (using universal generalization):

is-it-true[KB $\vdash$ $\forall$xy isa(x,EukCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$
$\exists$z isa(z,DNA) $\wedge$ encloses(y,z)]

or simply (treating the "KB $\vdash$" as implicit):

"A eukaryotic cell has a nucleus. Does that nucleus contain DNA?" (8)
equivalent to "Do all eukaryotic cells' nuclei contain DNA?"
(14) is-it-true[$\forall$xy isa(x,EukCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$ $\exists$z isa(z,DNA) $\wedge$ encloses(y,z)]

In other words, only if the nuclei of all eukaryotic cells contain DNA will some hypothetical eukaryotic cell's nucleus also contain DNA.

Note that (14) is the same as (12c), and thus question (8) is equivalent to (10c):

(10c) Do all the nuclei of eukaryotic cells contain DNA?
(8) A eukaryotic cell has a nucleus. Does that nucleus contain DNA?

This is important, as it establishes an equivalence between universals and hypotheticals -- it is valid for the user to reformulate (10c) as (8).

Additional Transformations if Unique Existence is True

Now suppose that eukaryotic cells have exactly one nucleus, i.e., (UE) is true:

$\forall$x isa(x,EukCell) $\rightarrow$. $\exists$!y isa(y,Nuc) $\wedge$ has-part(x,y) (UE) *"exactly 1 nucleus"*

then with some formula manipulation we can rewrite (14) as:

"A eukaryotic cell has a nucleus. Does that nucleus contain DNA?" (8)
equivalent to "Do all eukaryotic cells have a nucleus with DNA?" ***if*** *UE is true*
(15) is-it-true[$\forall$x isa(x,EukCell) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z)]

Note that we have been able to move clauses from the antecedent to the consequent in (15) because (UE) is true. Also note that (15) is the same as (10a) and (10b). Thus this analysis shows that, providing (UE) is true, then (8) is also a valid reformulation of (10a) and (10b):

(10a) Does every eukaryotic cell have a nucleus containing DNA?
(10b) Does the nucleus of every eukaryotic cell contain DNA?
(8) A eukaryotic cell has a nucleus. Does that nucleus contain DNA?

This is an important result regarding AURA's question formulation, as it means that users can also rewrite (10a) and (10b) as (8), providing (UE) is true.

Symmetrically, suppose that, biologically speaking, every nucleus is part of exactly one eukaryotic cell, i.e., (UE2) below is true:

$\forall$y isa(y,Nuc) $\rightarrow$. $\exists$!x isa(x,EukCell) $\wedge$ has-part(x,y) (UE2) *"exactly 1 cell"*

Under this assumption, we can rewrite (14) as:

(16) is-it-true[$\forall$y isa(y,Nuc) $\rightarrow$ $\exists$xz isa(x,EukCell) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z)]

Note that (16) universally quantifies over the nucleus, not the cell.

Overall Semantics

Overall, then, there are three ways of proving (8):

"A eukaryotic cell has a nucleus. Does that nucleus contain DNA?" (8)
is-it-true[$\forall$xy isa(x,EukCell) $\wedge$ isa(y,Nuc) $\wedge$ has-part(x,y) $\rightarrow$ $\exists$z isa(z,DNA) $\wedge$ encloses(y,z) (14)
$\vee$ (UE $\wedge$ $\forall$x isa(x,EukCell) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z)) (15)
$\vee$ (UE2 $\wedge$ $\forall$y isa(y,Nuc) $\rightarrow$ $\exists$xz isa(x,EukCell) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z))] (16)

Although the topic of answering these queries is out of scope for this document, we will mention that in AURA at present, NewQF's question answering tries both (15) assuming (not proving) UE, and (16) assuming (not proving) UE2, to answer such questions.

"Novel" Hypotheticals

Let us now consider questions where the antecedent ("setup") of the question's semantics is not repeating facts already known to be true in the KB. For example:

(17) A eukaryotic cell has a ruptured nucleus. Does the nucleus contain DNA?

Or in physics:

(18) A ball falls for 10 seconds. What is the distance of the fall?

Again a direct formalization of (17) would look:

(18) is-it-true[KB ∪ { isa(X0,EukCell) ∧ isa(Y0,Nuc) ∧ has-part(X0,Y0) ∧ ruptured(Y0) }
⊢ ∃z isa(z,DNA) ∧ encloses(Y0,z)]

which can be rewritten (again using universal generalization):

(19) is-it-true[∀xy isa(x,EukCell) ∧ isa(y,Nuc) ∧ has-part(x,y) ∧ ruptured(y)
→ ∃z isa(z,DNA) ∧ encloses(y,z)]

This is okay as a final semantics. In (19), we cannot move clauses from the antecedent to the consequent (as we did to transform (14) to (15)) as there is no uniqueness assumption (that every cell has a single ruptured nucleus). In general, only to the extent that unique existence axioms can be either proved or assumed can formulas be rewritten to move clauses from the antecedent to the consequent of the implication.

3.2 Find-A-Value Questions

3.2.1 Describing Individuals

Find-a-value questions involve finding the object (or objects) that the question is asking about; or more precisely, finding the *description* of those objects, where the description is the identifying information to report to the user. We use the predicate desc(instance,description) to relate an instance to its reportable description (a string).

In AURA, for desc(x,d), d is normally (a string containing the name of) the class of x, e.g., desc(nucleus01,"nucleus"). However, in the special case of x being a property value, d is a string representation of x's value, e.g., desc(length-value01,"10 meters"). In cases where d is an unknown property value, d might describe constraints on the value, e.g., desc(length-value02, “less than 10mm”). Additionally, in some cases it might be desirable for d to be more than just the class name, for example for an event, d might include mention of the participants such as desc(move01, “the nucleus moves to the center of the cell”). For our purposes here we remain agnostic about what the string should contain, and instead merely assume that desc(x,d) has an appropriate definition.

In general, an object might be describable in several ways, e.g., at different levels of abstraction. In other words, desc(x,d) is not necessarily a “functional” predicate[6], and may have multiple values of d for some given value x. For example, chromosome01 might be describable as “a human chromosome”, “a chromosome”, and “genetic material”.

3.2.2 Semantics: Universal Questions

We can generalize the earlier analysis to find-a-value questions. Consider the following example, similar to (7):

[6] In fact in AURA’s current implementation it is functional, i.e., there is exactly one description per object x. However, being able to return multiple descriptions of an object turns out to be important in the analysis here, so the specifics of AURA’s current implementation should be ignored.

(20) What does the nucleus of every eukaryotic cell contain?

(20) is another universal question. We can characterize its semantics as follows:

"The set of d, where d is a description of something that every euk cell's nucleus contains"

(21) { d | ∀x isa(x,EukCell) → ∃yz isa(y,Nuc) ∧ has-part(x,y) ∧ encloses(y,z) ∧ desc(z,d) }

In other words, the answer is the set of descriptions (of a content of a eukaryotic cell's nucleus) that are common to *every* eukaryotic cell, e.g., "DNA", "chromosome". Note that the answer is not the contents themselves (dna01, dna02, etc.) as each eukaryotic cell has different contents; rather the answer is the content *descriptions* that are shared by all eukaryotic cells, an abstraction of the contents themselves.

As mentioned earlier, an object might be described in several ways, e.g., at different levels of abstraction. For example, chromosome01 might be describable as "a human chromosome", "a chromosome", and "genetic material". (21) collects only those description(s) that are common to all eukaryotic cells. For example, it might return "a chromosome" as all eukaryotic cell's nuclei contain an object described as "a chromosome", but not "a human chromosome", as "human chromosome" is only a description of the contents of some nuclei. Thus if desc(x,d) finds descriptions at multiple levels of abstraction, (21) will automatically collect descriptions at the appropriate level of generality, i.e., at a level that applies to all the cells' contents[7]. An additional post-processing filtering of the results to remove redundant[8] descriptions would be a cosmetic nicety for presentation to a user.

So far, we have assumed encloses(y,z) will find individual entities z. If that is the case, then the enclosure of multiple entities with the same description will be hidden from the user. For example, suppose every eukaryotic cell's nucleus contains exactly two nucleoli[9], each with a description "a nucleolus". In this case, (21) will return simply "a nucleolus", masking the fact that there are in fact two per cell. To handle plurality, encloses(y,z) would need to also find aggregates as well as individual entities, each with an aggregate description, for example encloses(nucleus01,aggregate01), where aggregate01 has the description desc(aggregate01,"two nucleoli"), in order that an answer such as "two nucleoli" would be found to be common for all eukaryotic cells.

3.2.3 Semantics: Hypothetical Questions

Now let us consider a hypothetical question asking for values:

(22) A eukaryotic cell has a nucleus. What does that nucleus contain?

We can express its semantics as:

(23) { d | isa(X0,EukCell) ∧ isa(Y0,Nuc) ∧ has-part(X0,Y0) } ⊢ ∃z encloses(Y0,z) ∧ desc(z,d) }

which can be re-expresssed using universal generalization:

[7] AURA does not currently do this as desc(x,d) is functional, thus only if that (unique) description d is shared by some object in all eukaryotic cells' nuclei will an answer be found.

[8] A description D2 is redundant to D1 if ∀x desc(x,D1) → desc(x,D2). D1 can be thought of as a more detailed/precise description.

[9] This is biologically false, but we imagine it is true here for illustrative purposes.

"A eukaryotic cell has a nucleus. What does that nucleus contain?" (22)
or literally: What description(s) hold for some content of all euk cells' nuclei?"

(24) $\{ d \mid \forall xy\ isa(x,EukCell) \wedge isa(y,Nuc) \wedge has\text{-}part(x,y) \rightarrow \exists z\ encloses(y,z) \wedge desc(z,d) \}$

As for true/false questions, if UE is true then we can also rewrite the formula as:

"A eukaryotic cell has a nucleus. What does that nucleus contain?" (15)
or literally: What description(s) hold for some content of the nucleus of all euk cells?"

(25) $\{ d \mid \forall x\ isa(x,EukCell) \rightarrow \exists yz\ isa(y,Nuc) \wedge has\text{-}part(x,y) \wedge encloses(y,z) \wedge desc(z,d) \}$

3.3 Subsumption ("is-it-true-isa") Questions

A third class of questions is "isa" tests, which have a different quantification pattern. Consider:

(26) Is it true that every cell with a nucleus is a eukaryotic cell?
(27) Is it true that every cell with a nucleus is a cell with a cell call?

We can represent these as:

(28) $is\text{-}it\text{-}true[\forall x\ isa(x,Cell) \wedge (\exists y\ isa(y,Nuc) \wedge has\text{-}part(x,y)) \rightarrow isa(x,EukCell)]$
(29) $is\text{-}it\text{-}true[\forall x\ isa(x,Cell) \wedge (\exists y\ isa(y,Nuc) \wedge has\text{-}part(x,y)) \rightarrow (isa(x,Cell) \wedge \exists z\ isa(z,CellWall) \wedge has\text{-}part(x,z))]$

Or in general,

$is\text{-}it\text{-}true[\forall x\ F(x) \rightarrow G(x)]$

where F(x) and G(x) is a formula with free variable x. we can consider F(x) and G(x) to be definitions of (reified or unreified) classes, and thus the test is thus one of normal subsumption.

3.4 Cardinality Questions

Cardinality ("how many") questions can be formalized in a similar way to find-a-value questions, except the cardinality of the set of values, rather than the value descriptions, is returned. For example, the semantics of:

(30) How many chromosomes does every human cell's nucleus contain?

Can be expressed as:

(31) $\{ n \mid \forall x\ isa(x,HumanCell) \rightarrow \exists y\ isa(y,Nuc) \wedge has\text{-}part(x,y) \wedge n = ||\{ z \mid isa(z,Chromosome) \wedge encloses(y,z) \}|| \}$

We assume "How many" questions ask for the total number, rather than any number between 0 and that total. Thus this set of answers will necessarily be either a singleton or empty.

We also note that some cardinality questions might be answerable by reasoning about constraints on the set size, rather than enumerating the set explicitly. We currently do not know how to express this formally.

3.5 Taxonomic Questions

Taxonomic questions make straightforward calls to the explicitly asserted taxonomic links in the KB. We do not discuss these here as their semantics are simply queries about the subclass/superclass predicates in the KB. Examples of such questions are:

(32) What types of cells are there?
(33) What type of thing is a cell?

whose semantics can be written:

(34) { x | is-subclass-of(x,Cell) }
(35) { x | is-subclass-of(Cell,x) }

3.6 Possibility Questions

Statements of possibility ("Can..." rather than "Does...") are not supported by AURA, and require further analysis (still to be done). They are relatively unusual in biology, but an example from the Final Evaluation is:

(36) Can a DNA guanine bond to a RNA cytocine?

Note that it is not the case that DNA guanines are always bonded to RNA cytocines, nor is there even an example of a DNA guanine bonding with a RNA cytocine in the Biology KBs. This makes such questions challenging to answer.

3.7 Reflective Questions

By a reflective question, we mean questions that require some kind of meta-reasoning, looking at the axiom structure within the KB and/or of a proof. These questions are rare in biology exams, but an example (made up, i.e., not from a real biology exam) is:

(37) Does a cell's size affect its speed of division?

and similarly in chemistry (this one is from a real exam, and was part of the Halo Pilot project):

(38) Lithium, a very reactive metal with water, is above zinc, a metal used in galvanizing, in the activity series. This means that:
a. Zinc will react with lithium ions to produce lithium metal.
b. Lithium metal will react with zinc ions to produce zinc metal.
c. Zinc is oxidized.
d. One mole of lithium reacts with one mole of zinc ions.
e. No reaction will take place between lithium metal and zinc ions.

Note that (38) does not ask which statement is true (several might be true), but which statement follows from the premise that lithium is above zinc in the activity series. Again, we do not know how to formally specify such answers at present.

4 Partial Answers

4.1 True/False Questions

For true/false questions that are answered "no", although the answer may be correct it is not particularly revealing to the user about the contents of the KB. In this final section, we consider the notion of a "partial answer" to true/false questions that universally quantify over a single class, e.g., (12a) and (12b), to be offered if the answer to the original question is "no". The idea of a partial answer is, if a question answers "no", that the system also provides information about additional conditions under which the answer would be "yes". The motivation for this is to help the user, and our definition of "partial answer" is an attempt to formally specify something that

we think will achieve that goal. There may be other types of information that a system could provide to help the user when a "no" answer is returned – the notion of a "partial answer" here is just a guess at one such type.

Strictly by "partial answer" we mean a full answer to a specialized question Q_S, where Q_S tests only a subset of the objects tested by the original question Q. Formally, if a question Q tests if all objects x satisfying F(x) have properties G(x), and question Q_S tests if all objects x satisfying $F_S(x)$ have properties G(x), and $\forall x\ F_S(x) \rightarrow F(x)$, then we say Q_S is a specialized question of Q.

Consider the question:

(39) Does every cell contain a nucleus?

with semantics

(40) is-it-true[$\forall$x isa(x,Cell) $\rightarrow$ $\exists$y isa(y,Nuc) $\wedge$ has-part(x,y)]

The biologically correct answer to this question is "no" (as only eukaryotic cells have nuclei). Now note the parenthetical information just stated; intuitively, if the properties being queried are only true of some subset of the class being universally quantified over, then it would be useful for the system to report this, e.g.,:

Qn: Does every cell contain a nucleus?
Answer: No. But every eukaryotic cell contains a nucleus.

We call the "But..." part a ***partial answer*** as it characterizes the subset of things for which the queried properties are true, and we call the specialized class (here, "eukaryotic cell") the ***partial answer class***. There may be more than one partial answer class.

Formally, a partial answer class is a subclass C of the class being universally quantified over (here, EukCell), for which the question, with that class replaced by the subclass, has the answer true. In this case, the set of partial answer classes can be defined:

(41) { C | $\forall$x isa(x,C) $\rightarrow$ isa(x,EukCell) *i.e., C is a subclass of EukCell*
$\wedge$ $\forall$x isa(x,C) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z) }

In AURA, $\forall$x isa(x,C) $\rightarrow$ isa(x,EukCell) is refied as is-subclass-of(C,EukCell), thus we can rewrite (41) as:

(42) { C | is-subclass-of(C,EukCell)
$\wedge$ $\forall$x isa(x,C) $\rightarrow$ $\exists$yz isa(y,Nuc) $\wedge$ isa(z,DNA) $\wedge$ has-part(x,y) $\wedge$ encloses(y,z) }

Again as a cosmetic nicety, we should remove redundant members of this set when presenting it to the user, where a class C is redundant if it is a subclass of some other member of the set (i.e., we show the user only the most general partial answer classes).

Although reasoning is out of scope of this document, we will mention that AURA's new question-answering facility finds partial answers by retrying the original query replacing EukCell with some subclass C, stepping through every subclass in turn (in some sensible order).

In the above analysis, we only considered classes that are explicitly reified in the KB. A generalization of this is to consider unreified classes, i.e., that can be defined by some formula F(x). In this case, the set of partial answer classes (or more precisely, their definition), can be expressed by:

(43) { F(x) | $\forall x\ F(x) \rightarrow isa(x,EukCell)$
$\wedge\ \forall x\ F(x) \rightarrow \exists yz\ isa(y,Nuc) \wedge isa(z,DNA) \wedge has\text{-}part(x,y) \wedge encloses(y,z)$ }

For example:

Qn: Does every cell photosynthesize?
Answer: No. But every cell *containing a chloroplast* photosynthesizes.

In this example, the partial answer class is (the one defined by):

"cells containing a chloroplast"
$F(x) = isa(x,Cell) \wedge \exists y\ isa(y,Chloroplast) \wedge encloses(x,y)$

The space of possible formulae F(x) is potentially unbounded. In addition, not all F(x) would serve as useful answers to a user. Thus some means would be needed to restrict F(x) to "useful" formulae, e.g., by restricting F(x) to contain at most one non-isa predicate. This is an area for future work.

4.2 Find-A-Value Questions

We can generalize this analysis to apply it to find-a-value questions. In the situation where no values are returned, a "partial answer" can be found by similarly searching for a subclass of the universally quantified class that returns at least one value, and reporting the specialized question and its answer to the user.

5 Summary and Status

We have discussed the semantics of questions at length, to try and give clarity to the representation that the question interpreter should pass on to the reasoner to answer. Although the analysis is largely example-driven and incomplete in places, there are several items of significance which we summarize:

1. Questions often come with **presuppositions**, and these are important for manipulating the formal semantics of the question. In particular, the presupposition of unique existence (UE) and existence (E) are important, and we have introduced the notion of a nonsensical question with an undefined answer when its presupposition is violated.
2. We have shown some **important equivalences** between universal and existential questions, specifically for:

   (10a) Does every eukaryotic cell have a nucleus containing DNA?
   (10b) Does the nucleus of every eukaryotic cell contain DNA?
   (8) A Eukaryotic cell has a nucleus. Does that nucleus contain DNA?

   We have shown that these three have slightly different semantics, but that if "unique existence" of the eukaryotic cell's nucleus is true then the semantics will be the same, expressible as:

   $is\text{-}it\text{-}true[\forall x\ isa(x,EukCell) \rightarrow \exists yz\ isa(y,Nuc) \wedge isa(z,DNA) \wedge has\text{-}part(x,y) \wedge encloses(y,z)\ ]$

   This is important, because users often reformulate questions in one of these styles into one of the other styles when posing questions to AURA.

3. We have provided semantics for find-a-value questions, in particular highlighting the role of the description desc(x,d) of objects in finding answers.
4. We have introduced the notion of "partial answers" (or more precisely, full answers to partial questions), as a useful tool for providing a more informative response to questions whose answer is "no" (true/false) or "nothing" (find-a-value).

There are still holes in this analysis. The analysis has largely worked through examples, but not created a fully general framework. I remain unsure about the existence presupposition (E) as a way to characterize the presupposition behind universal questions (Section 2.3), but for now this seems to be the best way to capture it. I also do not have a formal semantics for questions about possibility (Section 4.6). The connection between this paper and prior work in linguistics on question semantics is also unclear, and more exploration is needed. Despite this, this paper (hopefully) helps to clarify the representation that the question interpreter should pass on to the deductive reasoner to answer.

References

Barker, K., Chaudhri, V., Chaw, S., Clark, P. Fan, J. Israel, D., Mishra, S., Porter, B., Romero, P. Tecuci, D., Yeh, P. 2004. A Question-Answering System for AP Chemistry: Assessing KR&R Technologies. In *Proc 9th International Conf on Knowledge Representation and Reasoning (KR'04)*, 2004, AAAI Press.

Chierchia, G. 1993. Questions with Quantifiers. Natural Language Semantics, 1, 181-234.

Clark, P., Chaudhri, V., Mishra, S., Thomere, J., Barker, K., Porter, B. 2003. Enabling domain experts to convey questions to a machine: a modified, template-based approach. Proc. KCap'03.

Clark, P., Chaw, J., Barker, K. Harison, P. Chaudhri. 2007. Capturing and Answering Questions Posed to a Knowledge-Based System. In: *Proc 4th Int Conf on Knowledge Capture (KCap'07).*

Cucina, R., Shah, M., Berrios, D., Fagan, L. 2001. Empirical Formulation of a Generic Query Set for Clinical Information Retrieval Systems. MEDINFO 2001, Amsterdam:IOS.

Gunning, D., Greaves, M., Chaudhri, V., et al., 2010. Project Halo Update: Towards Digital Aristotle. Submitted.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M., Lin, C-Y. 2001. Question-Answering in Webclopedia. Proc TREC-9, NIST.

Karttunen, L. 1977. Syntax and Semantics of Questions. Linguistics and Philosophy, 1, 3-44.

Lester, J., Porter, B. 1996. Scaling Up Explanation Generation: Large-Scale Knowledge Bases and Empirical Studies. AAAI'96.

Lester, J., Porter, B. 1997. Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments. *Computational Linguistics*, 23(1), pp. 65-101.