A Study of Some "Hard to Formulate" Biology Questions

Working Note 33

Peter Clark, Boeing Research, June 2009

Introduction

The below set of 22 questions was selected from the 50 final evaluation questions to focus question formulation research during the 2009 Halo work on question-asking. Specifically, we will be designing how domain knowledge, better dialog, and paraphrasing can improve the question formulation process. The goal of this document is not to design a solution, but to perform an in-depth analysis of some of the problems that the solution will need to address.

Although the question-answering scores were relatively high in biology in the final evaluation (typically in the 50%-70% range), and non-experts generally scored as high as experts, the fidelity of the formulations was very low: on average, only 18% of the words in the original question appeared in the CPL formulation, and, based on a random sample, the users failed to formulate some aspect of the majority of questions in the CPL version. There is thus substantial room for improvement in the formulation of these questions.

The below set of 22 questions was selected by identifying those where the non-experts' formulation received (on average) a lower[1] score than the expert's formulation (using the EE Biology KB). The goal was to select questions where the KB in principle could answer the question (i.e., high expert score), but non-experts had difficulty (i.e., lower non-expert score). In fact, the goal that the KB can in principle answer is only partially met by this method - it turns out that in some cases the KB is unable to answer (or even represent) the original question, yet the user obtained a good score by asking either a slightly different or more general question. We give examples of this and how the users were sometimes able to re-express questions in this way, and illustrate other problems and solutions that arose. We finish with some concluding remarks about the phenomena observed both in the original questions and the users' formulation attempts.

Below, "EE expert" refers the (SRI) biology expert who posed questions against the expert-built (Q1) KB. "EN non-experts" refer to the three non-experts (called Bio1, Bio2, Bio3) who posed questions against the same KB. In particular we looked at the questions posed by Bio2, who had the highest overall score of the three.

The Question Set

Question 2

Which of the following is NOT a characteristic of a prokaryotic cell?

[1] Specifically, where the EE expert score was > 0.3 higher than the average of the three EN scores.

a. A semi-fluid region consisting of cytosol
b. Membrane-bound cellular organelles
c. Plasma membrane consisting of lipids and proteins
d. Ribosomes that synthesize proteins
e. Rigid cell wall
The correct answer is: b.
Relevant sections: pp. 112-126

Comments:
This question nicely illustrates some of the challenges with the current CPL and the value of some of the proposed extensions. In this question, the term "consisting of" is vague, meaning "parts" or "encloses" - alignment with the KB would help with this disambiguation. However, CPL currently does no such alignment, and instead uses WordNet to map "consist" to the nearest CLib concept - here, Event - and thus direct entry of the original question phrase "a region consisting of cystol" would be interpreted as "An Event relating a Region and Cystol", a largely nonsensical interpretation. Rather, CPL should interpret “consisting of” as the correct relationship, or an underspecified relationship. A similar analysis holds for the vague term "characteristic."

Some KB Content:
In fact, the information in the KB does not perfectly align with the syntactic structure above. The nearest thing in the KB to "A prokaryotic cell has a semi-fluid region consisting of cytosol" is:

[Prokaryotic-cell] -has-part→ [Cytosol]

so an extra inference step would be needed to realize if X has-part Y, then X has a region containing Y. Similarly for "A plasma membrane consisting of lipids and proteins", the nearest thing in the KB is:

[Plasma-membrane] -has-basic-structural-unit→ [Phospholipid]
\-has-part→ [Protein]

thus "consisting of" should interpreted in a different way for lipids and proteins.

Question 5
In cells, which of the following can catalyze reactions involving hydrogen peroxide, provide cellular energy, and make proteins, in that order?
a. Peroxisomes, mitochondria, and ribosomes
b. Peroxisomes, mitochondria, and lysosomes
c. Peroxisomes, mitochondria,and Golgi apparatus
d. Lysosomes, chloroplasts, and ribosomes
e. Smooth endoplasmic reticulum, mitochondria, and ribosomes
The correct answer is: a.
Relevant sections: pp. 112-126

Comments:
Here, the KB has no knowledge of what "catalyze" means. The EE expert gets round this by talking about "synthesize" instead, thus matching the KB content.

Some KB Content:
[Peroxisome] -has-part→ [Enzyme] ←agent- [Synthesize]
\-raw-material→ [Hydrogen-peroxide]
\-result→ [H2O] [O2]
[Mitochondria] ←site- [Synthesis] -result→ [ATP]

Example CPL from the Final Evaluation:
EE: There is a synthesis of hydrogen peroxide.
Is it true that the peroxisome is the site of the synthesis?
[AURA says "no", but the answer book suggests this was misgraded]
There is a synthesis of ATP.
Is it true that the mitochondria is the site of the synthesis? *[Yes]*
There is a synthesis of protein.
Is it true that the site of the synthesis is a ribosome?
[AURA says "no", but the answer book has a "yes" answer recorded]
(overall answer scores 2 points)

EN-Bio2: What is a X? *for multiple X's. (answer scores 1 point)*

Question 9
Which of the following substances is found in RNA molecules but not in DNA molecules?
a. adenine
b. phosphate group
c. thymine
d. deoxyribose
e. ribose
The correct answer is: e
Relevant sections: pp 292-301, pp 304-311

Comments:
The phrase "found in" here means (if one aligns with the KB) "part". "phosphate group" should map to the concept Phosphate in the KB, whose text string states: "A phosphate group contains one atom of phosphorus, and four atoms of oxygen.". However, the KF SME hasn't added "phosphate group" as a synonym for "phosphate", so CPL will

currently choke on this phrase. Again, this question illustrates how alignment with the KB could help.

Some KB Content:
The KB appears incomplete: it does not state that RNA contains Phosphate (the concept for "phosphate group").

Example CPL from the Final Evaluation:
EE: What is the difference between RNA and DNA?

The answer includes "ribose" for RNA but not for DNA, and hence scores 2 points. Interestingly, the incompleteness mentioned above does not result in lost points - rather, "phosphate group" isn't mentioned for either RNA and DNA, and the grader does not seem to have minded.

Non-experts: Is it true that thymine is part of RNA? *etc.*

Generally this is enough to score points, despite the infidelity.

Question 37
Prokaryotic and eukaryotic cells generally have which of the following features in common?
a. A membrane-bound nucleus
b. A cell wall made of cellulose
c. Ribosomes
d. Similar size and complexity
e. A nucleoid
The correct answer is c.
Relevant sections: pp 112-126

Comment:
"have" is a vague term (similar to "consisting of" in Question 2). Alignment with the KB would help disambiguate this.

Example CPL from the Final Evaluation:
Users avoided explicitly stating the relationships to test by asking:
EE: What are the similiarities between prokaryotic cells and eukaryotic cells?
(answer was good enough to score 1.5)

Question 43
Which of the following organelles is not matched with its corresponding function?
a. Ribosome, protein synthesis
b. Nucleolus, ribosome production

c. Golgi apparatus, sugar synthesis
d. Nucleus, DNA replication
e. Chloroplasts, photosynthesis
The correct answer is: c
Relevant sections: pp. 112-120, pp 317-320

Comment:
The question is asking if an unstated, and possibly indirect, relationship holds between an object and an event. (A literal interpretation of "matched with" would be inappropriate here). Also see question 89 for a similar example.

Example CPL from the Final Evaluation:
EE: What is a ribosome? *etc. (answers scored 1.5 points)*

EN: Non-experts tried a large variety of formulations, e.g.,

What is the purpose of the golgi apparatus?
Is it true that the nucleus replicates DNA?
What is the location of DNA replication?

Here the non-experts are essentially taking stabs in the dark to try and replicate the structure in the KB. This is clearly undesirable!

Question 53
The exchange of segments of chromatids during synapsis is known as
a. cross linkage
b. gene linkage
c. sex linkage
d. crossing over
e. gametogenesis
The correct answer is: d
Relevant sections: pp. 239-245

Some KB Content:
The KB does not seem to have an adequate representation of Synapsis to answer this question, containing only:

[Synapsis] -isa→ [Come-Together]
\-object→ [Chromosome]
\-destination→ [Homologous-Chromosome]
\-site→ [Centromere]

It also has:

[Crossing-Over] -isa→ [Transfer]
\-donor→ [Chromatid]
\-recipient→ [Homologous-Chromosome]
\-result→ [Recombinant-Chromosome]

Despite this, the EE expert scored 2 points by asking simply "What is a crossing over?", even though "exchange" and "segments" and "synapsis" weren't mentioned in the answer.

Question 56
Chemical analysis indicates that the cell membrane is composed mainly of
a. proteins and starch
b. proteins and cellulose
c. lipids and starch
d. lipids and proteins
e. proteins and mitochondria
The correct answer is d.
Relevant sections: pp 112-126

Comments:
"composed of" again needs to be interpreted with respect to the KB. CPL currently interprets "compose" as Write (!) (as in "compose a letter", using WordNet to climb the hypernym tree).

Some KB Content:
"starch" is not directly known to the KB, but WordNet nicely maps it to Polysaccharide, the general class of which starch is a member.

Example CPL from the Final Evaluation:
Non-experts tried a variety of relationships, including "contains", "has", and "is part of", between a cell membrane and the chemicals mentioned, trying to find the right one.

Question 68
Which of the following bonding examples is not possible?
a. A DNA adenine to a DNA thymine
b. A DNA thymine to an RNA thymine
c. A DNA guanine to an RNA cytosine
d. A DNA adenine to an RNA uracil
e. A DNA guanine to a DNA cytosine
The correct answer is: b
Relevant sections: pp. 239-245

Comments:
This currently requires the user translating "possible pairing" to "complement", a very specific transformation.

Example CPL from the Final Evaluation:
EE: Is it true that adenine is the complement of thymine? *(etc)*
(answer scores 2 points)
EN-Bio2: same. (answer scores 2 points)

Question 73
Which of the following best describes the DNA molecule?
a. Two parallel strands of nitrogen bases held together by hydrogen bonding
b. Two complementary strands of deoxyribose and phosphates held together by hydrogen bonding
c. Two antiparallel strands of nucleotides held together by hydrogen bonding
d. A single strand of nitrogen bases coiled upon itself by hydrogen bonding
e. A single strand of nucleotides coiled into a helix.
The correct answer is c.
Relevant section: pgs 292-301

Comments:
This is a challenging question concerning the spatial and physical structure of DNA.

Some KB Content:
The knowledge being tested for in this question is largely absent in the KB.

Example CPL from the Final Evaluation:
EN: Users used a variety of "What is a X?" questions, e.g.,
What is DNA?
What is a double helix?
What is hydrogen bonding?
What is a DNA strand?

Question 74
In DNA replication, the role of DNA polymerase is to
a. bring two separate strands back together after new ones are formed
b. join the RNA nucleotides together to make the primer
c. build a new strand from 5` to 3`
d. unwind the tightly would helix
e. join the Okazaki fragments
The correct answer is c.
Relevant section: pps 292-301

Comments:

Again, these are complex processes being described in the question, for which alignment with the KB is necessary as part of either the question interpretation or question answering process.

Example CPL from the Final Evaluation:
EE: What is DNA-polymerase? *(answer scores 2 points)*
EN-Bio2: What is DNA polymerase?
(scores 0 points because of a CPL bug, HLO-2712).

Question 77
Regarding mitosis and cytokinesis, one difference between higher plants and animals is that in plants
a. the spindles contain cellulose in addition to microtubules whereas in animals they do not.
b. sister chromatids are identical, whereas in animals they differ from one another.
c. a cell plate begins to form at telophase, whereas in animals a cleavage furrow is initiated at that stage.
d. chromosomes become attached to the spindle at prohpase, whereas in animals chromosomes do not become attacked until anaphase.
e. spindle poles contain centrioles, whereas spindle poles in animals do not
The correct answer is: c
Relevant sections: pp. 217-224

Comments:
The KB does not distinguish between mitosis (etc.) in plant vs. animal cells, there's just a single representation and so seems inherently incapable of answering this question faithfully.

There's an interesting example of CPL over-generalization in one formulation: The user asks "Is it true that an animal cell has a cell plate?" (the correct answer is yes). "cell plate" is not in the KB, so CPL interprets "plate" as "Tangible-Entity", so the question is interpreted as "Is it true that an animal cell has a tangible-entity?". AURA answers yes, so scores a point by luck.

Some KB Content:
[Animal-Cell] -has-part→ [Clevage-furrow]
[Telophase-II] ←agent- [Clevage-furrow]
[Cytokinesis] ←agent- [Clevage-furrow]
"cell plate" is not in the KB.

Example CPL from the Final Evaluation:
EE: Is it true that an animal cell has a clevage furrow? *AURA: Yes - 1 pt*
EN-Bio2: What are the subevents of cytokinesis?

What are the subevents of mitosis?
AURA mentions Clevage-furrow - 1pt

Question 79
All of the following organelles are associated with protein synthesis EXCEPT:
a. ribosomes
b. Golgi bodies
c. the nucleus
d. the rough endoplasmic reticulum
e. the smooth endoplasmic reticulum
The correct answer is: e.
Relevant sections: pp 112-126

Comments:
"Golgi bodies" isn't in the KB, the EE expert rephrased this as "Golgi apparatus". The notion "associated with" is vague, and again needs to be rephrased.

Example CPL from the Final Evaluation:
EE: What is the relationship between protein synthesis and a ribosome? *(etc)*
What is the relationship between a golgi apparatus and a protein?
[→ an answer for all but the smooth endoplasmic reticulum, score 2 points]
EN-Bio2: more specific questions, e.g.,
What is the function of the rough endoplasmic reticulum?
What is the function of the golgi apparatus? etc. *(answer scores 2 points)*

Question 80
Which of the following best corresponds to vesicles that serve to break down cellular debris?
a. Smooth ER
b. Lysosome
c. Cell wall
d. Plastids
e. Nucleolus
The correct answer is: b.
Relevant sections: pp 114-119, 121-122

Comments and Some KB Content:
"serve to break down cellular debris" is a complex notion. The closes thing that the KB contains is:

[Lysosome] -site-of→ [Take-Apart] -object→ [Subcellular-Entity]

Here, "Taking apart subcellular entities" appears to be the closest the KFE was able to get to saying "breaking down cellular debris".

Example CPL from the Final Evaluation:
EE: The EE expert, with benefit of knowing the KB encoding, was able to re-express the question to exactly match his/her structure:

> There is a take apart.
> The object of the take apart is a subcellular entity.
> Is it true that the lysosome is the site of the take apart?
> *Answer: Yes (answer scores 2 points)*

EN-Bio2: EN-Bio2 was able to score 1 point by asking:
> What is a lysosome?

Question 81
Which of the following best corresponds to a semi-rigid structure that lends support to a cell?
a. Smooth ER
b. Lysosome
c. Cell wall
d. Plastids
e. Nucleolus
The correct answer is: c.
Relevant sections: pp 112-126

Comment:
CLib has no notion of "rigid" and so cannot represent this knowledge. The notion of "support" here is also complex: CLib contains a structural notion of "support" (e.g., a pillar), but not the subtler notion of (something like) "provide resistance to deformation". WordNet also does not have this sense of support.

Some KB Content:
The knowledge being queried here is not represented in the KB, and is outside the representational capabilities of the current ontology.

Example CPL from the Final Evaluation:
EE: What is a cell wall?
> *answer scores 2 points by simply reciting the documentation string (not interesting from our perspective here).*

Question 85

During which stage of mitosis do nucleoli reappear?
a. telophase
b. anaphase
c. early prophase
d. late prophase
e. metaphase
The correct answer is: a
Relevant sections: pp. 217-224

Comments:
The notion of "reappear" is a complex notion not in CLib. The KB does not directly state that "nucleoli appear" - rather, that nucleoli are created during telophase (see KB structure below). The EE expert specifically writes CPL to match that target KB structure. However, we'd prefer for either CPL or the problem-solver to bridge the mismatch between the question structure and the KB structure.

Some KB Content:
Mitosis] -subevent→ [Telophase] -first-subevent→ [Create] -result→ [Nucleolus]

Example CPL from the Final Evaluation:
EE: There is a create.
The result of the create is the nuclear envelope.
The second result of the create is the nucleolus.
Is it true that the create is a subevent of telophase? *AURA: Yes. (score 2 points)*
EN-Bio2: What are the subevents of mitosis? *(score 0 points)*

Question 89
Which of the following cellular organelles is most closely associated with the transcription activity of RNA?
a. mitochondria
b. nucleus
c. ribosomes
d. Golgi apparatus
e. lysosome
The correct answer is: b
Relevant sections: pp. 304-311

Comments:
The vague notion of "associated with" needs to be expanded, either as part of the question interpretation or question answering process. The question essentially asks whether some unspecified and possibly indirect relationship exists between an object and an activity. Also see question 43 for a similar example.

Question 91
During meiosis there are two rounds of all of the following stages EXCEPT
a. prophase
b. metaphase
c. anaphase
d. telophase
e. interphase
The correct answer is: e
Relevant sections: pp. 239-245

Comments:
The phrase "two rounds" is important but makes this question difficult to formulate. Really, the users should be asking "How many times does prophase happen during meiosis?", or "How many prophases are subevents of meiosis?" to capture the "two" notion. In the final evaluation the users generally worked around this.

Example CPL from the Final Evaluation:
EE: What are the steps of meisois?
What are the steps of meisois I?
What are the steps of meisois II? *(answer scores 2 points)*
EN-Bio2: What are the differences between meiosis-i and meiosis-ii?
(answer scores 2 points)

Other non-experts tried and failed to encode the notion of "two rounds" in various ways (answers score 0 points)

Question 103
The part of a cell that is in most direct contact with the environment is the
a. nucleus
b. cell membrane
c. mitochondrion
d. centrioles
e. chloroplasts
The correct answer is b.
Relevant sections: pp 112-126

Comment:
There is no notion of "in contact with the environment" in the KB; a better way of rephrasing this, within the constraints of the ontology, would be "which part is not contained in anything else?". This question would be hard to rephrase properly. The question in fact requires somewhat complex spatial reasoning to answer.

Example CPL from the Final Evaluation:

EE: What is a cell membrane?

answer scores 2 points by simply reciting the documentation string (not interesting from our perspective here).

Question 104

Plant cell organelles that contain photosynthetic pigments are
a. chloroplasts
b. centrioles
c. nuclei
d. cell walls
e. mitochondria
The correct answer is a.
Relevant sections: pp 112-126

Comment:

The notion of "photosynthetic pigments" does not exist in the KB. The EE expert cleverly tweaks the question from "containing photosynthetic pigments" to "performing photosynthesis." (below).

Example CPL from the Final Evaluation:

EE: There is a photosynthesis.
Is it true that the site of the photosynthesis is the chloroplasts?

Answer: yes (scores 2 points)

Question 106

The sites of protein synthesis in the cytoplasm are the
a. ribosomes
b. lysosomes
c. nuclei
d. centrioles
e. cell membrane
The correct answer is a.
Relevant sections: pp 112-126

Comments and Some KB Content:

It seems like "site" should translate directly to the slot "site" in CLib. However, for the correct answer (a) the relationship in the KB is in fact "agent":

[Ribosome] -agent-of→ [Synthesis] -result→ [Protein]

Example CPL from the Final Evaluation:

EE: The EE expert rewords "site" to "agent" in his/her CPL:

There is a synthesis of protein.
Is it true that the agent of the synthesis is a ribosome?
Answer: yes (scores 2 points)

Note that the EE expert writes "synthesis of protein" rather than "protein synthesis". CPL interprets "synthesis of protein" as [Synthesis] -result→ [Protein], but would interpret "protein synthesis" as [Protein-synthesis]. It is not clear if the problem-solver can then reason from [Protein-synthesis] to [Synthesis] -result→ [Protein] in the context of matching, in order to answer this question.

Question 121
The synaptonemal complex forms during
a. anaphase I of meiosis
b. interphase of any cell
c. anaphase II of meiosis
d. telophase
e. prophase I of meiosis
The correct answer is e.
Relevant section: pgs 239-245

Comment:
"synaptonemal complex" is unknown to the KB. The EE expert gets away by rephrasing the question in terms of "synapsis", but Synapsis does not have information in the formal representation to answer this question, and so is uninteresting (EE expert scores 2 points from use of the documentation string on Synapsis).

Question 122
Homologous chromosomes separate during
a. prophase I
b. prophase II
c. anaphase of mitosis
d. anaphase I
e. anaphase II
The correct answer is d.
Relevant section: pgs 239-245

Comments and Some KB Content:
This looks like it would directly and straightforwardly translate to CPL. However, there is a subtle structural mismatch: The question suggest the separation is a *subevent* of Anaphase I, while the KB states that Anaphase I *is* the separation event:

[Anaphase-I] -isa→ [Move-Apart] (i.e., separate)
\-object→ [Homologous-Chromosome]

This mismatch thwarted some of the non-experts, as their use of the word "during" translated to a subevent link, not present in the KB.

Example CPL from the Final Evaluation:
EE: The EE expert subtly rephrased this to match the KB structure directly:

> There is a move apart event.
> The object of the event is a homologous chromosome.
> What is the event?

Note that the EE expert could have been even more faithful and used the word "separate" instead of "move apart", as CPL interprets "separate" as Move-Apart.

Some Concluding Comments

There are several interesting things that emerge from this analysis:

1. There is often a mismatch between the knowledge structure suggested by the question and the actual structure in the KB (e.g., see questions 5, 80, 85, 122). The EE expert managed to get round this by rephrasing the question to match the KB structure. The log files show the non-experts hunting around different wordings to find one that matched the KB structures.

   Clearly this is undesirable, and AURA itself needs to bridge between the question and KB structures, either as part of question interpretation, or question answering, or both. This is a major topic for investigation this year. The boundary between where question interpretation ends, and question answering ends, or even if there should be a boundary, are important questions to consider.

   Closely related to this is the challenge of interpreting a "no" answer from AURA. If AURA is sensitive to wording, then the user has little guidance as to whether a "no" really means "no" or just means that the wording didn't quite match a structure in the KB that would otherwise produce a "yes" answer.

2. There is often terminology in the questions which does not easily translate into KB concepts. Occasionally there are good examples of WordNet bridging the gap, e.g., “starch” → Polysaccharide (q56), "separate" → Move-Apart (q122), but often the onus has been on the user to replace the terminology with known terminology, e.g., "Golgi apparatus" for "Golgi bodies" (q79), "synthesize" for "catalyze" (q5), "subcellular entities" for "cellular debris" (q80). Improved methods of helping the user in this task would make formulation easier.

3. There are some examples of paraphrases in the text above also, e.g., "consisting of" (q2), "found in" (q9), and "has" (q77) all can denote a part-of relationship. However, in this question set at least, the number of paraphrases seems relatively small, and hence use of a paraphrasing database may have only a small impact on

the overall question formulation problem in this domain. This has still to be investigated and explored.