Creating a Corpus of Questions and Answers about Biological Processes: Annotation Guidelines

Jonathan Berant and Vivek Srikumar

1 Introduction

We have the ability to read text that describes a biological process (that is, a collection of inter-connected events that lead to an end result) and answer complex questions about the relationships between the events. Our goal is to develop systems that can automatically answer complex biology AP style questions in such a reading comprehension setting. We will use a hand created corpus of questions associated with text to train and evaluate the systems.

2 Generating questions and answers

The goal is to generate multiple-choice questions about biological processes that are described in a paragraph of text. The questions should focus on the events and entities participating in the process. Consider the following paragraph (from the textbook *Biology* by Campbell and Reece) as an example:

> The light reactions are the steps of photosynthesis that convert solar energy to chemical energy. Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. Light absorbed by chlorophyll drives a transfer of the electrons and hydrogen ions from water to an acceptor called $NADP^+$ (nicotinamide adenine dinucleotide phosphate), where they are temporarily stored. The electron acceptor $NADP^+$ is first cousin to NAD^+, which functions as an electron carrier in cellular respiration; the two molecules differ only by the presence of an extra phosphate group in the $NADP^+$ molecule.

There are several events described in the paragraph – splitting of water, absorption of light, transfer of electrons and hydrogen ions, etc. These events involve entities like water, electrons and protons, chlorophyll, etc.

We can write several questions about these. Some examples are listed below, with the correct answer marked in **boldface**:

1. A source of electrons and protons are provided after which event?
   (a) **Water is split**
   (b) Light is absorbed
2. Which of the following events is caused by the absorption of chlorophyll?
   (a) **Transfer of electrons and protons into** $NADP^+$
   (b) The splitting of water
3. What event would not happen if water does not provide electrons and hydrogen ions?
   (a) Light absorption by chlorophyll
   (b) **Transfer of ions to** $NADP^+$

3 Guidelines for generating questions and answers

We are primarily interested in questions that depend on the inter-relationships between events. An event can be a *subevent* or a *super-event* of another one. Additionally, an event can *enable*, *cause* or *prevent* another one. Note that these **event-event** relations often imply a temporal ordering between them. For example, if event e_1 causes an event e_2, then e_1 should occur before e_2.

Entities can play different roles in one or more events. For example, an entity can be the performer of an event (or more formally, the `Agent` of the event), or it can be acted upon in the event (that is, the `Patient` of the event), it could be generated in the event (the `Result` of the event), and so on.

We have identified the following classes of questions that verify understanding of these relationships between events and entities:

1. For event e:
   (a) What event will be caused or prevented by e?
   (b) If e does not happen, what else will not happen?
   (c) What event *should* occur after/before e?
   (d) What events are necessary for e to occur?
   (e) What event immediately follows e?
2. For events e_1, e_2:
   (a) Which one happens first?
   (b) What is the relation between them (eg. e_1 causes e_2, e_1 is a super event of e_2, and so on)?

(c) What is the sequence of events between them?

(d) Is some relation between e_1 and e_2 true or false?

3. For events $e_1,\ldots,e_n$:

   (a) What is the correct ordering of the events?

   (b) Which may simultaneously occur?

4. (a) Which entity performs a given role (eg. `Agent`, `Theme`, `Result`) for an event?

   (b) what entity is the result of an event e?

5. What role does an entity perform in an event?

6. For entity a:

   (a) What entities are necessary to produce a?

   (b) What events are necessary to produce a?

   (c) If a is not produced what events will not happen?

   (d) If a is not produced what other entities will not be produced?

7. If a_1 and a_2 are two entities in the process, how does a_1 lead to the production of a_2?

Note that these are only templates for types of questions and the actual questions need not look exactly like them. For example, the first question in the three example questions listed above asks what events are necessary to produce an entity (template 6.2). Similarly, the second question belongs to the template 1.1 and the third one belongs to the template 1.2.

3.1 Detailed guidelines

1. Each question **should** correspond to one of the templates identified above.

2. Each question should be associated with two answers, where only one is unambiguously correct and the other is unambiguously incorrect (and called the *distractor*).

3. It should be possible to answer the question by reasoning about the events and entities and their relationships, as specified in the text.

4. Do not use background knowledge that is not present in the text. In the above example, if the text did not identify that protons are hydrogen ions, represented by H^+, we should not use these names in the questions or the answers.

5. When referring to entities and events in the questions and answers, use their names as they appear in the paragraph. However, sometimes an entity may be referred to by different names (like proton or H^+ in the paragraph). If (and only if) this happens, you can refer to the entity by any of these names. For example, if the text talks only about *blood flow*, do not use the similar phrase *blood stream*.

6. Do not use contractions or drop words in the names of entities unless the text becomes awkward without doing so.

7. We are only interested in events and entities that participate in them. In the paragraph above, the last sentence says "the two molecules differ only by the presence of an extra phosphate group in the $NADP^+$ molecule". Note that this sentence does not describe an event. Do not generate questions based on such sentences.

8. Naturally, the question will contain words that relate the different events and entities mentioned. We have a few conventions on the use of these words:

   (a) The word *cause*, as in "What is caused by event e?" refers to the events that can only happen if e happens. However, a more specific meaning of *cause* is the events that *must* happen if e happens. To refer to the latter case, we will use the phrase *necessarily cause*.

   (b) Assume event e_1 causes event e_2, which in turn causes event e_3. The answer to the question "What is caused by e_1?" is e_2 and e_3. If we want to refer only to e_2 we will use the question "What is immediately caused by e_1".

9. For questions that ask about the order of events, use only English in the answers and not symnbols such as $\rightarrow$. A possible answer is "First, event e_1 happens, then e_2, and finally e_3" rather than "$e_1 \rightarrow e_2 \rightarrow e_3$".

10. Do not ask about the state of entities, that is about properties of entities even as a result of some event. For example, the question "What is the end result of prometaphase?" and the answer "the two centrosomes are at opposite ends of the cell" are not good since they ask about the state of the centrosomes rather on the activities expressed in the text.

11. Phrase grammatical questions – use "What does e lead to?" rather than "e leads to what?"

12. Do not ask questions that require linguistic knowledge that is not in the text such as opposites. For example, if the text says "e causes the enzyme to become inactive", don't ask "What causes the enzyme to stop being active?"

13. In general, answers often contain an evet such as "Water is split". It is allowed but not recommended to create distractors by combining an event word such as "split" with

another entity that appears in the text such as “NADP+”, and so “NADP+ is split” is a plausible distractor. These questions are fine if it is not easy to understand from the text which entity is related to the event. Other than that, creating distractors by changing a word or a phrase in the correct answer is not allowed.