Supplementary Material: Modeling Biological Processes for Reading Comprehension

Jonathan Berant[*], **Vivek Srikumar**[*], **Pei-Chun Chen, Brad Huang and Christopher D. Manning**
Stanford University, Stanford

Abby Vander Linden and Brittany Harding
University of Washington, Seattle

1 Features for Trigger Classifier

The following binary features are extracted for a trigger candidate for the trigger classifier:

1. The WordNet synsets of the candidate
2. The Nomlex nominalization type
3. Whether the trigger is an auxiliary verb
4. Clustering of trigger candidate using WordNet to one of the direct hyponyms of the synset *Entity#n#1*
5. Levin verb classes of trigger candidate
6. Adverbial modifiers of trigger candidate
7. Lemma and POS tags of preceding two words
8. Lemma and POS tags of following two words
9. dependency path to root of sentence
10. POS tag of dependency parent conjoined with POS of trigger candidate and the connecting dependency edge relation
11. For each dependency child, POS tag of dependency child conjoined with POS of trigger candidate and the connecting dependency edge relation
12. Indicator if dependency parent is a nominalization
13. indicator if there is a dependency child that is a nominalization
14. Indicator if trigger candidate is a nominalization whose dependency parent is a verb
15. Indicator if the trigger candidate is in a gazetteer of biological processes compiled from Wikipedia

2 Features for Argument Identifier

For an argument candidate of a trigger, we extracted the following features for our argument identifier:

1. The syntactic category of the argument candidate
2. The POS tag of the trigger
3. Conjunction of trigger POS tag and argument category
4. Whether the argument node contains a sentence category
5. Indicator if there is a dependency relation between the trigger and argument
6. Dependency path from trigger to argument
7. Length of dependency path from trigger to argument

3 Features for Joint Model

Event-Argument features For a given trigger and argument candidate, we extract the following features:

1. The head word, its lemma and POS tag,
2. The parse tree node that covers the argument,
3. The subcategorization frame,
4. Indicator for whether the argument contains an SBAR node,
5. Dependency path from the trigger to the head of the argument,
6. Length of the dependency path,
7. Path from argument node to the root of the constituency tree
8. Indicator for whether the argument is before or after the trigger,
9. Number of tokens between the trigger and the argument,
10. Lemma of head of argument
11. Lemma of head of argument conjoined with POS tag of trigger
12. Lemma of head of argument conjoined with lemma of trigger
13. Lemmas of words between the trigger and argument candidate
14. From the verb (PropBank) and nominal (NomBank) semantic role annotation, an indicator for whether the trigger is an predicate and the argument candidate is an argument of that predicate,
15. For the previous case, the label of the argu-

[*] Both authors equally contributed to the paper.

ment, and
16. The previous two features, when the argument heads match.

Event-Event Relation features Given a pair of triggers, we extract the following features:

1. Lemmas of two words preceding each trigger
2. Lemmas of two words following each trigger
3. Lemmas of both triggers,
4. Indicator for whether the two triggers have the same lemma,
5. Word and sentence distance between the triggers (word distance is binned into buckets of <5,6-7, 8-10, 11-15,16-30, >31) ,
6. The determiner of the trigger
7. Conjunctions of the following features:
   - Lemmas and POS tags of the triggers and
   - Cluster identifiers if both triggers are contained in a cluster, using EXCHANGE clustering.
8. Adverbial modifiers for triggers
9. Lowest common ancestor of triggers in constituency tree, if it exists
10. Dependency path between the triggers
11. Length of dependency path between the triggers
12. Whether first trigger dominates the second in the dependency tree
13. Whether second trigger dominates the first in the dependency tree
14. The child of a *mark* dependency relation, if one exists
15. Preposition lexeme, if in a prepositional phrase
16. Whether triggers share a dependency child
17. For each trigger, indicator for whether the trigger is an SRL predicate
18. Whether triggers share an SRL argument
19. Whether the triggers are adjacent in the paragraph
20. Whether triggers are adjacent and which trigger is first
21. Word between triggers
22. Whether first trigger is a noun and the first word in the paragraph

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

1 Introduction

We have the ability to read text that describes a biological process (that is, a collection of interconnected 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 it can be acted upon in the event, it could be generated in the event, and so on.

We have identified the following templates 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?

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?

3. For events e_1,...,e_n:
   (a) What is the correct ordering of the events?
   (b) Which may simultaneously occur?

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

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 produces 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 like them. For example, the first question in the three example questions listed above asks what events are necessary to produce an entity (template 6b). Similarly, the second question belongs to the template 1a and the third one belongs to the template 1b.

3.1 Other guidelines

1. Each question should be associated with two answers, where only one is unambiguously correct and the other is unambiguously incorrect.

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

3. Avoid 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.

4. When referring to entities and events in the questions and answers, try to use their names as they appear in the paragraph. However, sometimes the same 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.

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

6. 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 for such sentences.

Guidlines for Annotating Process Structures

1 Annotation guidelines

Annotating processes has four parts:

1. Identifying event triggers
2. Identifying arguments
3. Annotating event-argument relations
4. Annotating event-event relations

1.1 Identifying Event Triggers

A trigger is a span of text that denotes the occurrence of an event. Usually triggers are single words but they can also be phrases. In general, a trigger is the minimal text span that denotes event occurrence.

One case where a triggers are longer than a single word are nominalizations with modifiers. The trigger can be multiple words if its meaning is more than the sum of its parts, that is, it is a named entity that refers to some specific process. For example, *natural selection* is a trigger since it is a particular conventionalized expression. Some more examples:

- *genetic drift*: a compound referring to a particular event.
- *geographic isolation*: compositional, only *isolation* is the trigger.
- *different mutation*: compositional, only *mutation* is the trigger
- *light reaction*: compositional, only *reaction* is the trigger and *light* is an argument.

Another case is when the trigger is composed of a light verb + noun/adjective. A light verb is a verb that has a vague semantic meaning. In these cases, we would mark the noun/adjective as the trigger. Some examples are:

- make cuts
- made permanent
- became elevated

1.2 Identifying Arguments

Often events contain various participants, or arguments. For example, in *after gametes fuse and form...* the word *gametes* participates in the event denoted by the trigger *fuse*.

For arguments the general rule is that we mark the full text span that denotes an argument. For example, *The red blood cells of people with sickle-cell disease become distorted.* The argument is *The red blood cells of people with sickle-cell disease* and the trigger is *distorted*. Notice that we include articles such as *a* and *the* in the span. Similarly, in *protons (hydrogen ions)*, we would mark the entire span as an argument including the words in parenthesis.

Important: We only care about entities when they participate in some event – there is no need to annotate any entity that does not participate in some event through an event-argument relation as specified below.

1.3 Event-Argument Relations

Each argument has a role in an event that is marked by an edge from the trigger to the argument. We define the following roles:

1. AGENT: An argument that is the **doer** or **performer** of the action in the event
2. THEME: An argument that is the entity on which the action is **done/performed**. If an argument is both an AGENT and a THEME (*the cell divides*), we mark it as a THEME.
3. RESULT: An argument that is produced or generated and is the result of the event.

4. SOURCE: In an event that involves movements from point A to B, the entity denoting A.

5. DESTINATION: In an event that involves movements from point A to B, the entity denoting B.

6. LOCATION: The location where the event takes place, if specified.

7. OTHER: Other participants

Note that an argument may have more than one role. For example, *Gene flow occurs both to and from the sub-population*. In this case, the *sub-population* is both the SOURCE and DESTINATION of the event gene flow.

1.4 Event-Event relations

We define the following event-event relations

1. SUPER: An event A is a super-event of event B if B is an event that is part of A. The arrow in this case will go from B to A

2. SAME: Coreference relation between event mentions

3. CAUSE/ENABLE: These relations are types of **dependence**. If event A causes/enables event B in some process, then this means that in order for event B to happen, event A has to happen. The difference between CAUSE and ENABLE is that if event A causes event B, then whenever A happens, B also happens (it is caused by it). On the other hand for ENABLE, if A happens, it is not true that B necessarily happens, it is only the case that B can only happen after A, but it is not directly caused by it. Examples:

   - CAUSE: *He* **threw** *the glass on the floor and it* **broke**.
   - ENABLE: *He* **opened** *the door and the cat* **came** *in.*

4. CAUSE-OR/ENABLE-OR: A further expansion is the use of CAUSE-OR and ENABLE-OR. Sometimes an event C depends on two or more events A and B.

   If both A and B have to happen for C to happen we mark this by A–CAUSE/ENABLE→C and B–CAUSE/ENABLE→C.

   If either A or B have to happen in order for C to happen, then we mark it as A–CAUSE-OR/ENABLE-OR→C and B–CAUSE-OR/ENABLE-OR→C.

   Similarly, if A ENABLE/CAUSE B or C, we annotate by A–ENABLE-OR/CAUSE-OR→B and A–ENABLE-OR/CAUSE-OR→C.

5. PREVENT: The opposite of CAUSE. If A happens, then B does not happen.

6. There is also PREVENT-OR analogously to ENABLE-OR and CAUSE-OR

2 Stative events

The general rule is that we are interested in events with participating entities, and not in the state or properties of entities. However, sometimes an event is expressed by describing a state that has been created:

> *An increase in hormones causes high blood pressure which then causes ...*

In this case we need to annotate *high blood pressure* as an event since the text implicitly refers to the creation of a new state, that later causes other events. The general rule is, we annotate a stative event if there is an event of creation of that state, and there is no other trigger that accounts for this creation more explicitly.

However, this does not mean that all mentions of a state need to annotated – here are some examples where we should not annotate a stative word as an event

1. *The smaller group is isolated, and then establishes a population with a gene pool that differs...*: In this case, *differ* is not a trigger since the event is establishing a population that has that has a property, and so the trigger is *establish*.

2. *lava cools and then radioisotopes become trapped. The trapped isotopes...*: In this case there is no need to annotate the second mention of *trapped* as an event, since it describes a property of the isotopes whose creation has been already mentioned in the previous trigger. So this stative-event should not be annotated.

Last, note that the construction of an entity with a verb that describes it may sometimes have a regular event in which case it should be annotated. For example, *The mRNAs hybridize with a probe recognizing beta-globin.* In this case, *recognizing* is in a relative clause but denotes a non-stative event that should be annotated. A test for that is if this can be paraphrased in a more clear way such as *that then recognizes*.

3 Further Guidelines

1. For SUPER events, mark all sub events. No need to mark hierarchical sub-event. That is, if A is super-event of B and B is super-event of C, no need to mark that A is super-event of C.

2. SAME is a transitive relation. If A is the same as B and B is the same as C, then A is the same as C.

3. The events in a process need to be connected to one another. If they are not, then this means there is more than one process and we should be notified.

4. Some sentences with no annotation: if the sentence does not discuss the process, talks about something else, explains the structure of some entity, etc.

5. For arguments such as LOCATION, SOURCE, and DESTINATION, if it is expressed with a preposition (*in the cell*) mark only the noun phrase as the argument (*the cell*).

6. Distinguishing CAUSE from ENABLE. It is important to note that we dont handle "probabilistic causation/enablement". That means that if we have a condition such as"A may cause B" we annotate this as CAUSE and not ENABLE. The enable relation signals that A creates conditions for B to happen, but not that A increases the probability of B happening. We ignore probabilities in this annotation.

7. In general if we have an expression such as a *protein (Ced-9)*, then we annotate both as entities that have a coreference relation. However if we have a *protein (Ced-9) in the membrane*, then it can not be separated and we annotate the whole span as an entity.

8. Attachment of LOCATION, SOURCE, DESTINATION to entity rather than event: Note that when we annotate the mentioned role, they should describe the event rather than the entity. For example, in the sentence *scientists infer that bacteria in the body...*, the phrase *in the body* refers to the location of the bacteria and not the *infer* event. This of course means that they are part of the entity and should not be annotated as a location of the event. Another example is *isotopes from the environment become trapped*. Again, the *environment* is the source of the isotopes, not the source of the *trapped* event. This requires attention when annotating these roles.

3.1 Some Tests for Deciding Hard Cases

(a) A verb is not a trigger, but expresses a PREVENT relation or a T-NO on another trigger if it can be replaced with *does not happen*. For example, *Gene flow is blocked* denotes that the event *Gene flow does not happen*.

(b) When it is not clear if events ENABLE/CAUSE another, or are the SAME, there are two tests:

   i. If the events have contradicting roles, it is not SAME. For example, if they dont have the same THEME.

   ii. If there is a point in time where event a has started but event b has not, it is ENABLE/CAUSE and not SAME.