Modeling Biological Processes for Reading Comprehension

Vivek Srikumar
University of Utah (Previously, Stanford University)

Jonathan Berant, Pei-Chun Chen, Abby Vander Linden,
Brittany Harding, Brad Huang, Peter Clark and Christopher D. Manning

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

Reading comprehension is hard!

Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. Light absorbed by chlorophyll drives a transfer of the electrons and hydrogen ions from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Reading comprehension is hard!

Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. Light absorbed by chlorophyll drives a transfer of the electrons and hydrogen ions from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Reading comprehension is hard!

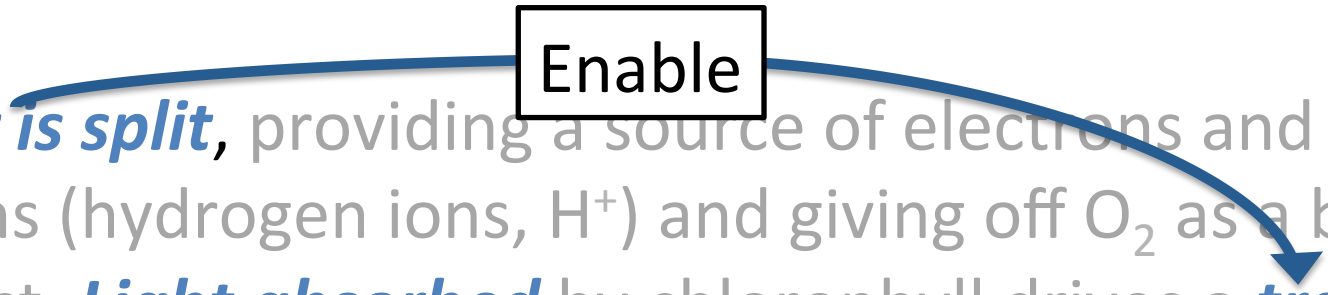
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Reading comprehension is hard!



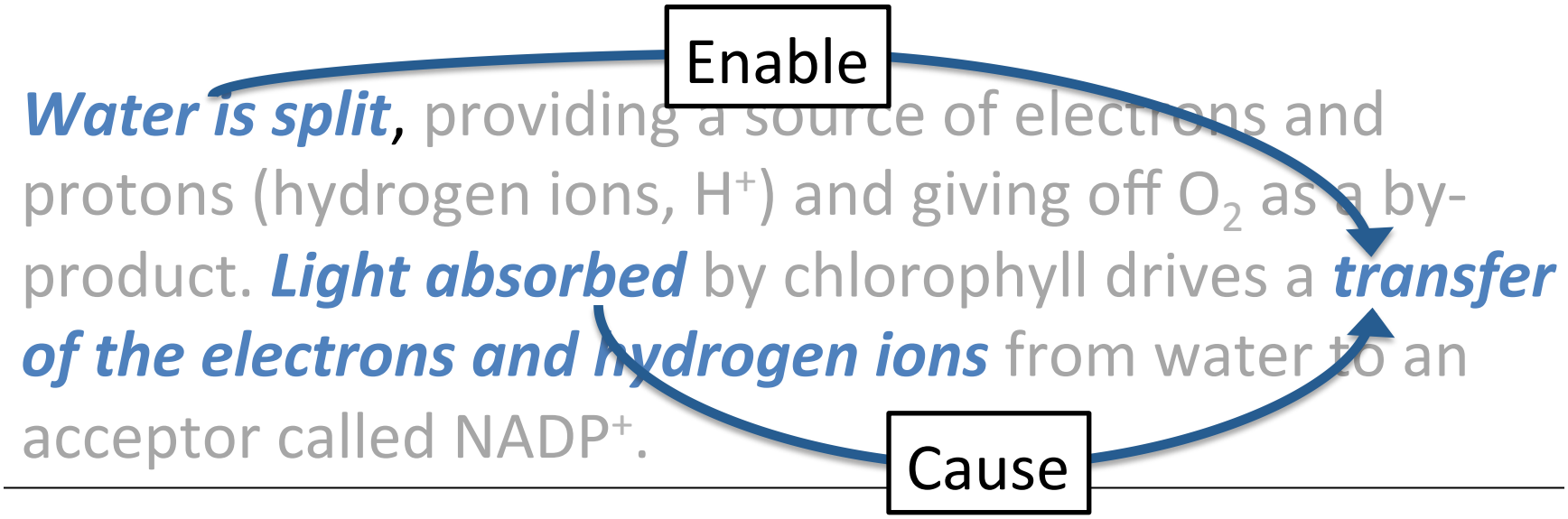
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Reading comprehension is hard!



What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Contributions

1. A new reading comprehension task requiring reasoning over processes
   - Processes are fundamental in many domains

2. A new dataset `ProcessBank` consisting of descriptions of biological processes with
   - Rich process structure annotated, *and*
   - Multiple-choice questions

3. A new end-to-end system for reading comprehension
   - Predict structure and treat it as a knowledge base
   - Parse question as query to this KB (semantic parsing)

A new dataset: `ProcessBank`

Motivation: macro vs. micro reading

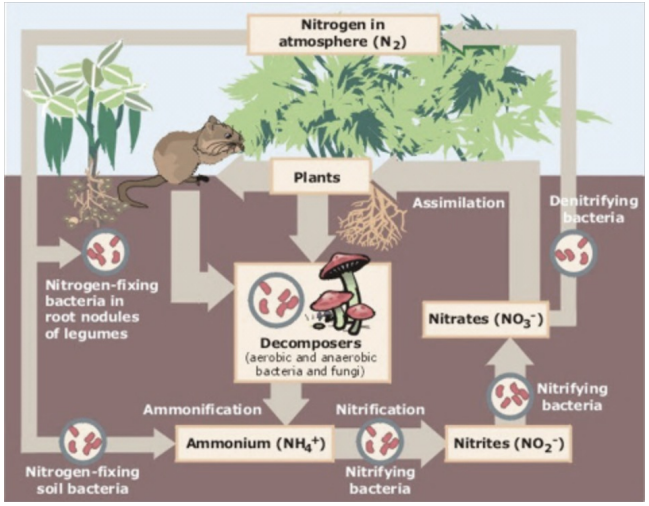
- Macro reading:
  - Exploits web-scale redundancy
    [Etzioni et al., 2006, Carlson et al., 2010, Fader et al., 2011]
  - Factoid questions
    [Berant el al., 2014, Fader et al., 2014]

- Micro reading:
  - Single document
  - Requires reasoning
  - Non-factoid questions
    [Richardson et al., 2013, Kushman et al., 2014]

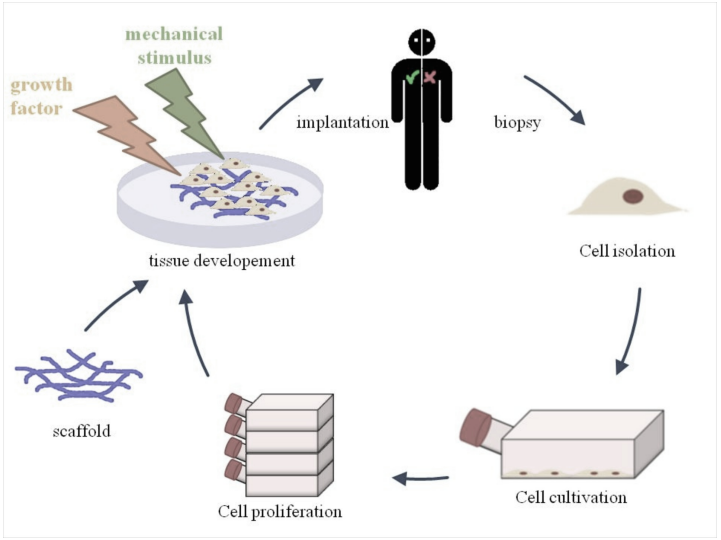
Chosen domain:

Biological process descriptions

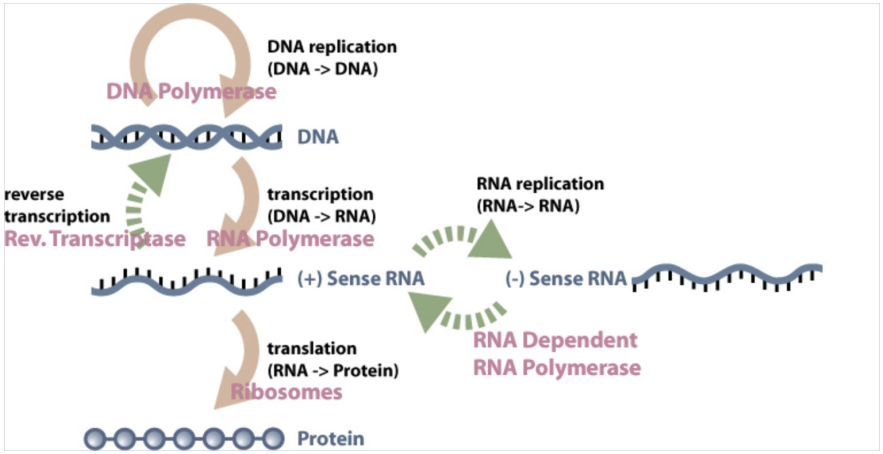
Processes abound in biology



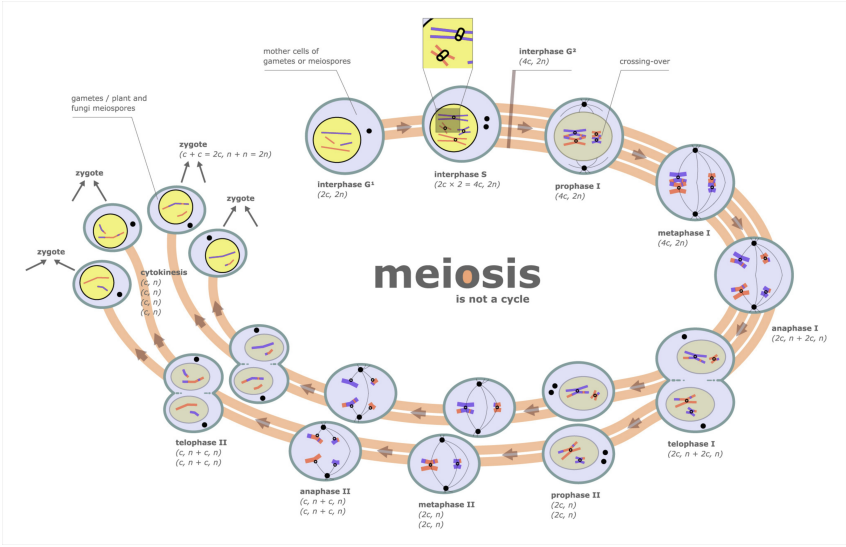
Nitrogen Cycle



Tissue Engineering



Central Dogma of Molecular Biology



Creating a difficult reading comprehension task

200 paragraphs from the textbook *Biology*

Extending [Scaria, et al. 2013] [Campbell & Reese, 2005]

Desiderata

1. Test understanding of inter-relations between events and entities
2. Both answers should have similar lexical overlap:
   - Trump shallow approaches
   - Sidestep lexical variability

Reading comprehension annotation

- Annotation instructions: Ask questions about events, entities and their relationships
  - 10 examples provided
  - Two answer choices, only one unambiguously correct
- 200 paragraphs $\rightarrow$ 585 questions
- Second annotator answered the questions
  - 98.1% agreement

Examples of annotated questions

Dependencies between events/entities (70%)

Q: *What can the splitting of water lead to?*

A: Light absorption
B: Transfer of ions

Temporal ordering of events (10%)

Q: *What is the correct order of events?*
A: PDGF binds to tyrosine kinases, then cells divide, then wound healing

B: Cells divide, then PDGF binds to tyrosine kinases, then wound healing

True-False questions (20%)

Q: *Cdk associates with MPF to become cyclin*

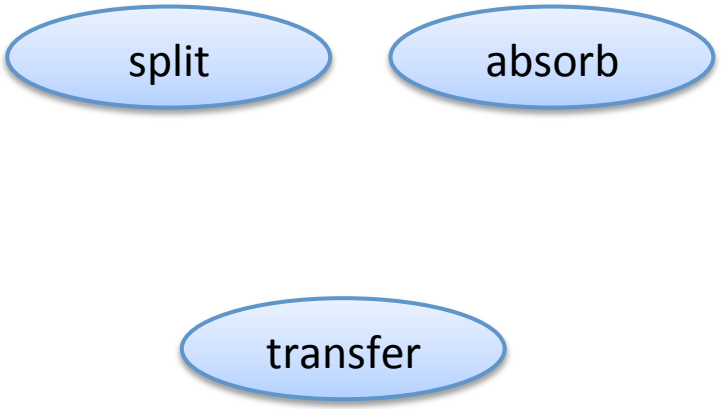
A: True
B: False

A second layer of annotation: Process structures

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 $NADP^+$.

A second layer of annotation: Process structures

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 $NADP^+$.



Triggers: Tokens denoting occurrence of an event

A second layer of annotation: Process structures

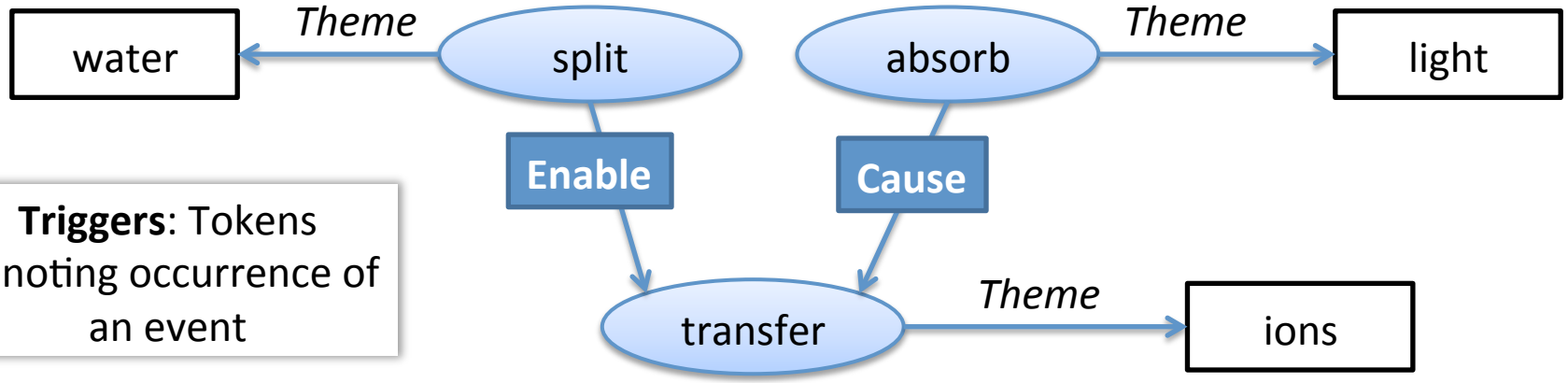
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

Triggers: Tokens denoting occurrence of an event

Arguments: *Agent, Theme, Source, Destination, Location, Result* and *Other*

A second layer of annotation: Process structures

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 $NADP^+$.



Triggers: Tokens denoting occurrence of an event

Arguments: *Agent, Theme, Source, Destination, Location, Result* and *Other*

E-E relations: *Cause, Enable, Prevent, Super, Same*

Process structure data

- Same 200 paragraphs from *Biology*
  - Paragraphs annotated and verified
- Three annotators
  - Biologists
  - Independent from QA annotator
    - Potentially conflicting with questions
- More nuances
  - Eg: No temporal ordering of events
    - Contrast with [Scaria et al 2013]

What is ProcessBank?

- 200 paragraphs from the textbook *Biology*
  - Manually chosen to represent biological processes
- Each paragraph annotated with
  - Non-factoid reading comprehension questions
  - Process structures

Answering questions: Overview

System in a nutshell

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 $NADP^+$.

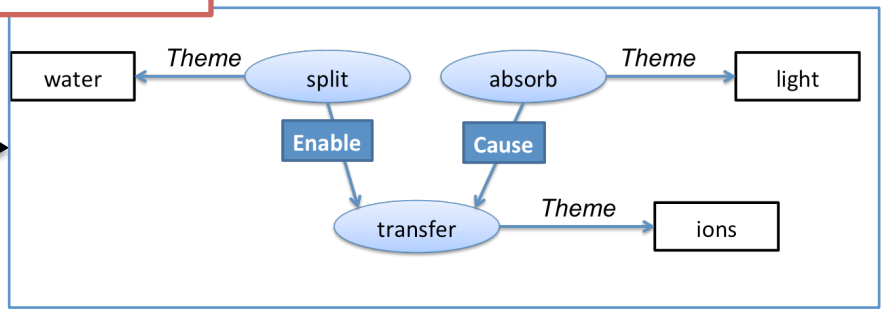
What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions

System in a nutshell

Process Structure Prediction

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 $NADP^+$.

Step 1



What can the splitting of water lead to?

A: Light absorption

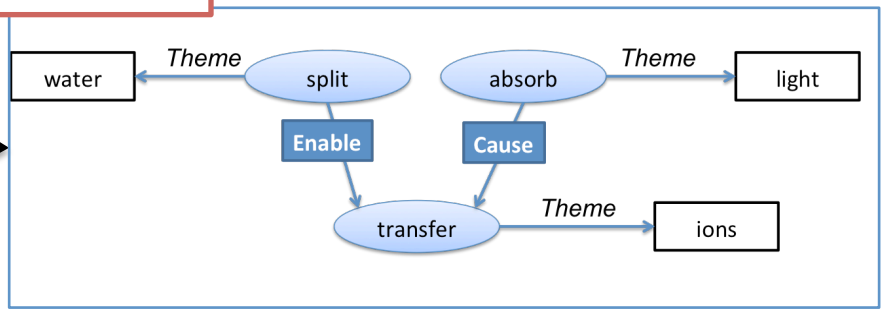
B: Transfer of ions

System in a nutshell

Process Structure Prediction

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 $NADP^+$.

Step 1



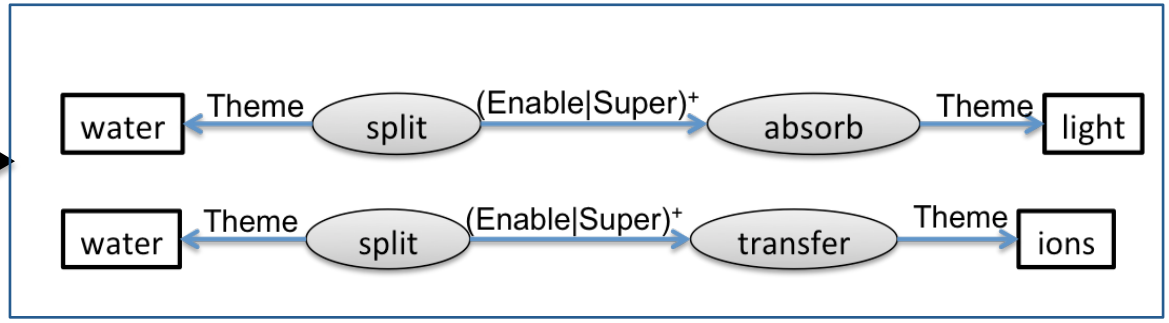
What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Step 2

Question Parsing

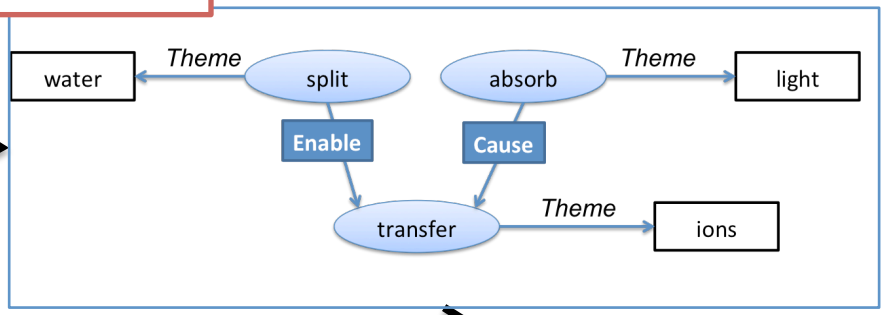


System in a nutshell

Process Structure Prediction

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 $NADP^+$.

Step 1



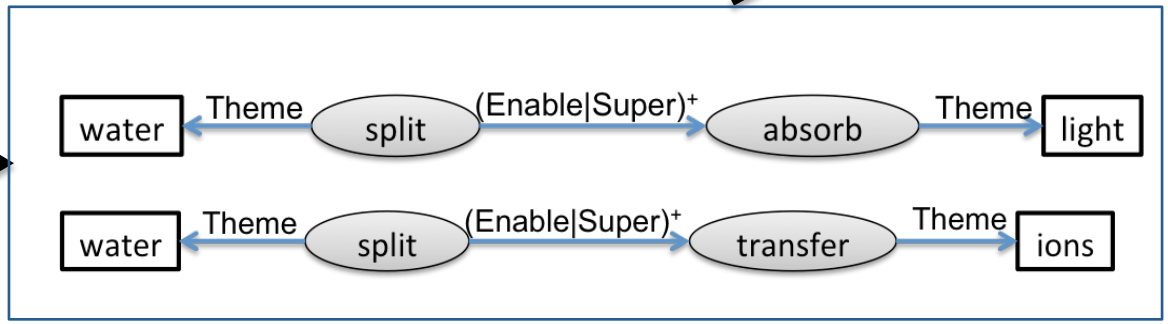
What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions

Answering Question

Step 3: Answer = **B**

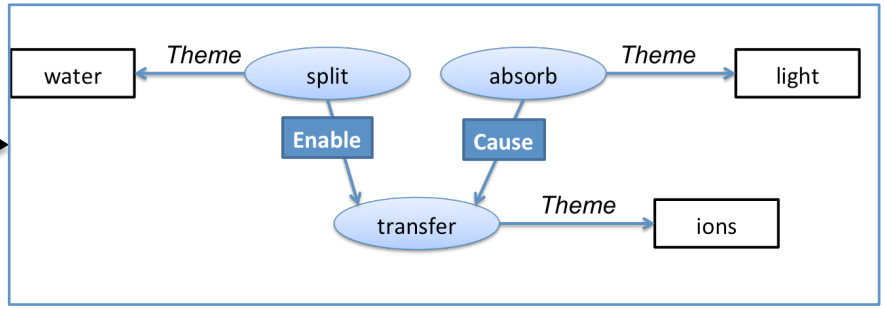
Step 2

Question Parsing



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 $NADP^+$.

Step 1



Predicting process structures

Process structure prediction

1. Train trigger identifier

   Logistic regression; features from words, lists

2. Joint learning and inference over arguments and event-event relations using predicted triggers

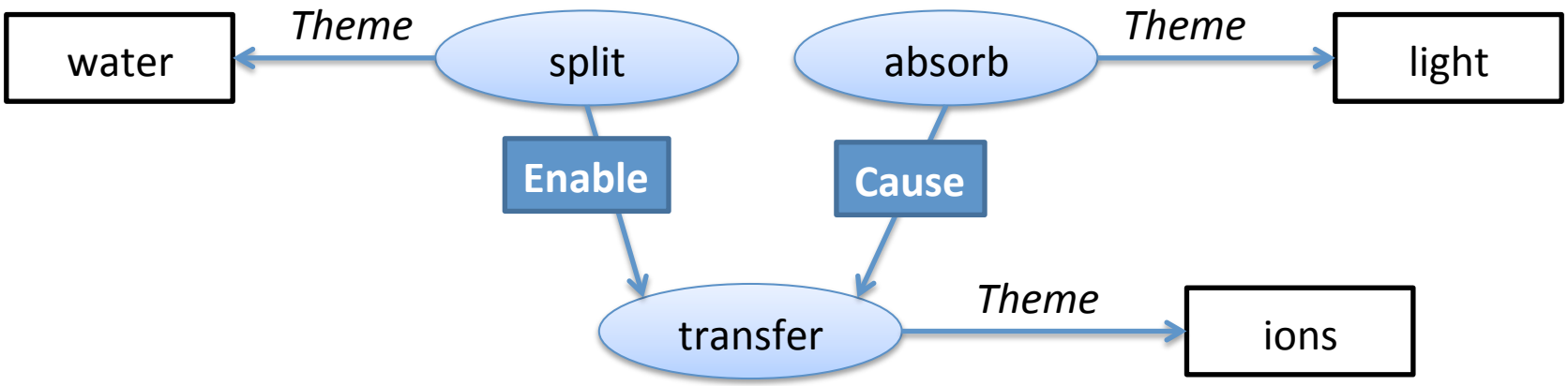
Joint inference over (1) and (2) increases runtime

… without noticeable performance gains

Fast feature engineering cycles are important!

Event-arguments and event-event relations

$$\max \left[\text{Total event-argument score} + \text{Total event-event relation score} \right]$$



[Roth and Yih, 2004, Punyakanok et al. 2008, etc]

Event-arguments and event-event relations

$$\max \sum_{t,a,L}$$

Score that arg-candidate has label L

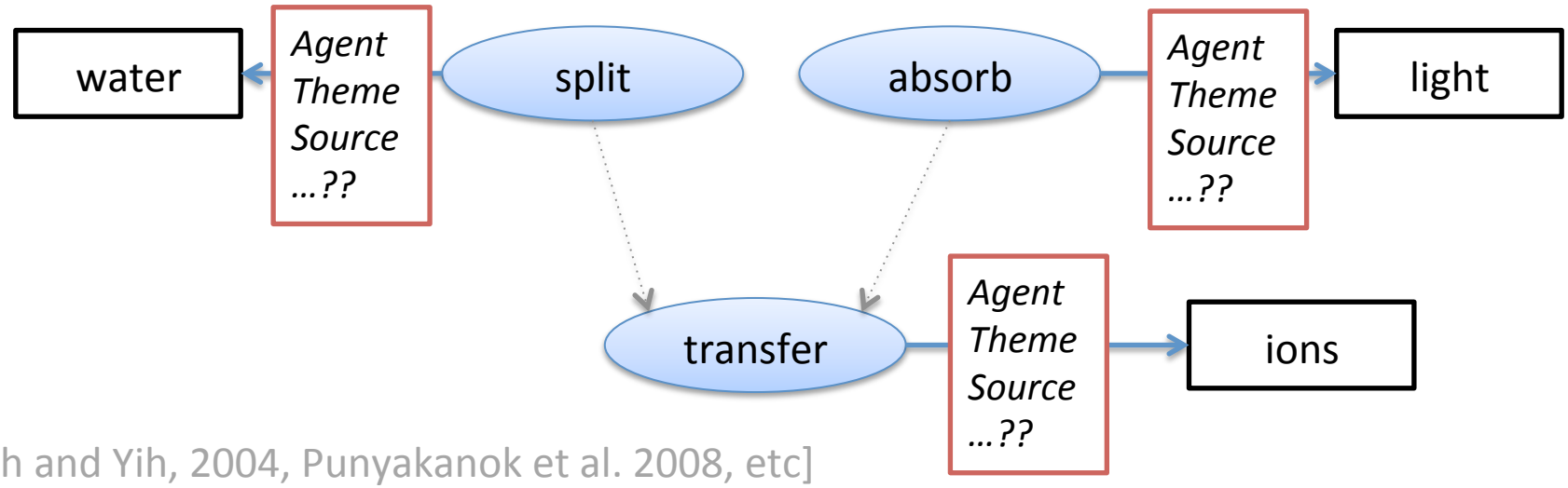
+

Total event-event relation score

Trigger *t*

Argument candidate *a*

Argument label *L*



[Roth and Yih, 2004, Punyakanok et al. 2008, etc]

Event-arguments and event-event relations

$$\max \sum_{t,a,L} \text{Score that arg-candidate has label L} + \sum_{t_1,t_2,R} \text{Score that trigger pair connected by relation R}$$

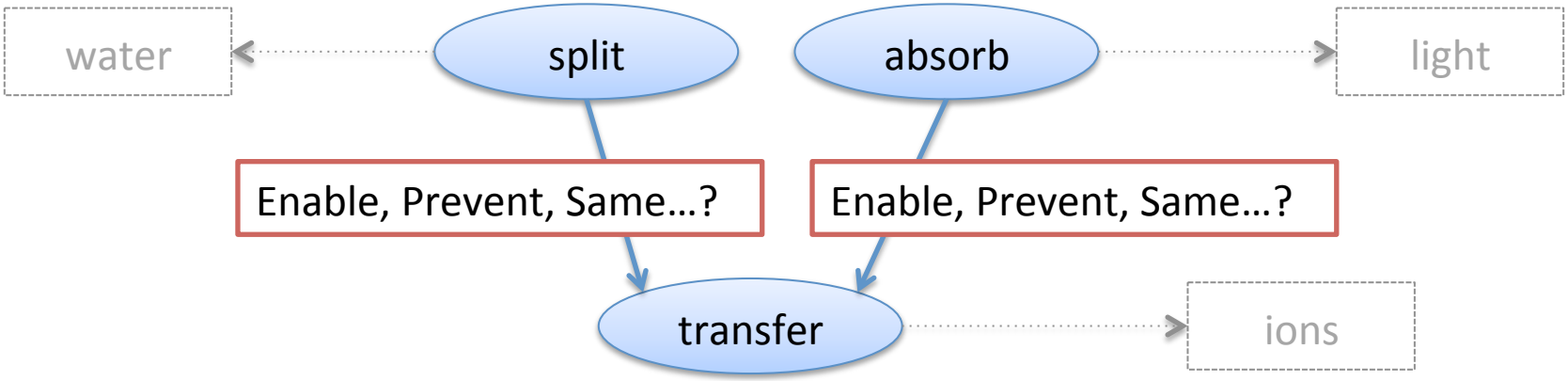
Trigger t

Argument candidate a

Argument label L

Pair of events (t_1, t_2)

Relation label R



[Roth and Yih, 2004, Punyakanok et al. 2008, etc]

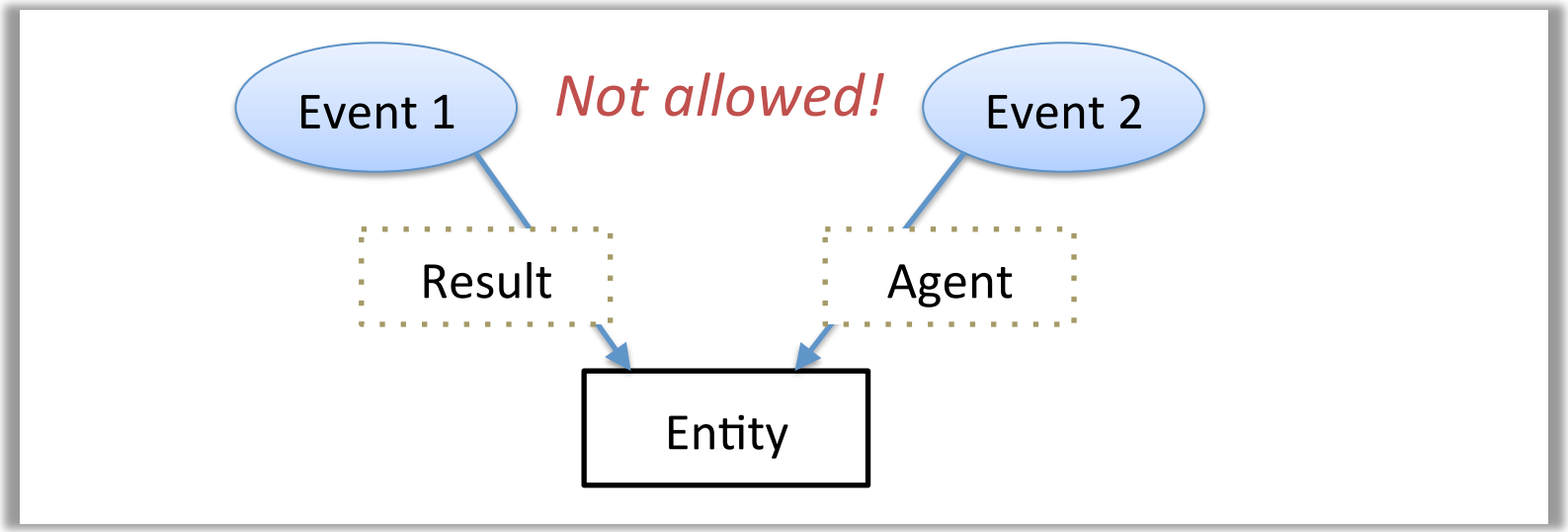
Joint inference with constraints

1. No overlapping arguments
2. Maximum number of arguments per trigger
3. Maximum number of triggers per entity
4. Connectivity
5. Events that share arguments must be related

And a few other constraints

Joint inference with constraints

1.
2.
3.
4.

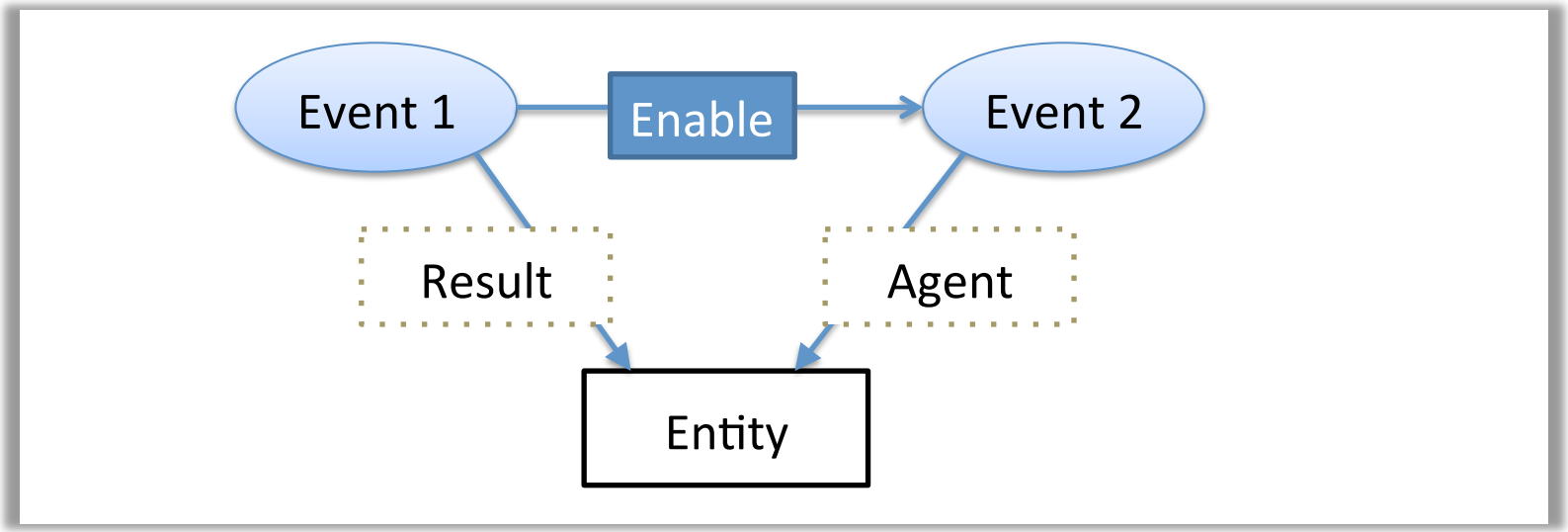


5. ***Events that share arguments must be related***

And a few other constraints

Joint inference with constraints

1.
2.
3.
4.



5. ***Events that share arguments must be related***

And a few other constraints

Learning and Inference

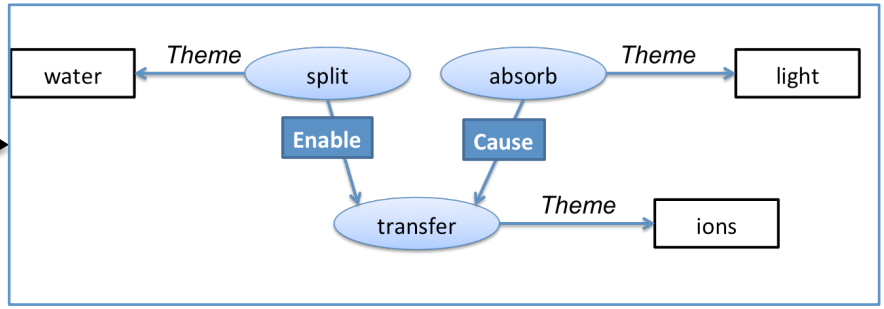
- Linear model to score argument labels and event-event relations
  - Related: Semantic role labeling, information extraction
- Structured averaged perceptron
- Gurobi ILP solver (exact solution)

Answering questions

Where are we?

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 $NADP^+$.

Step 1



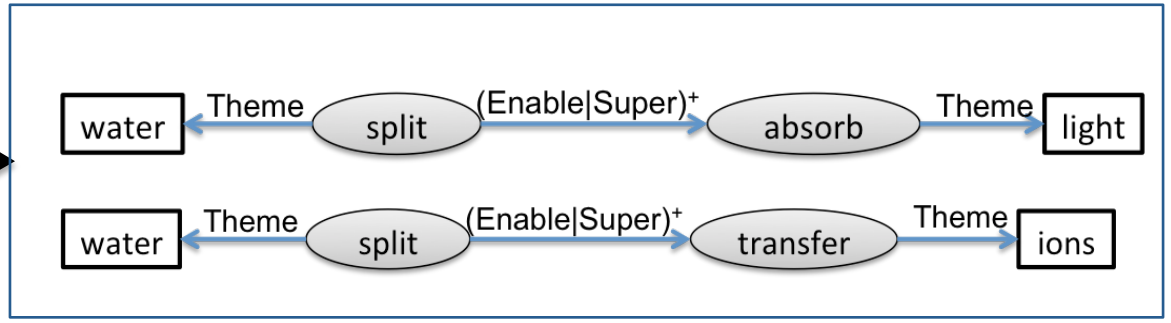
What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Question Parsing

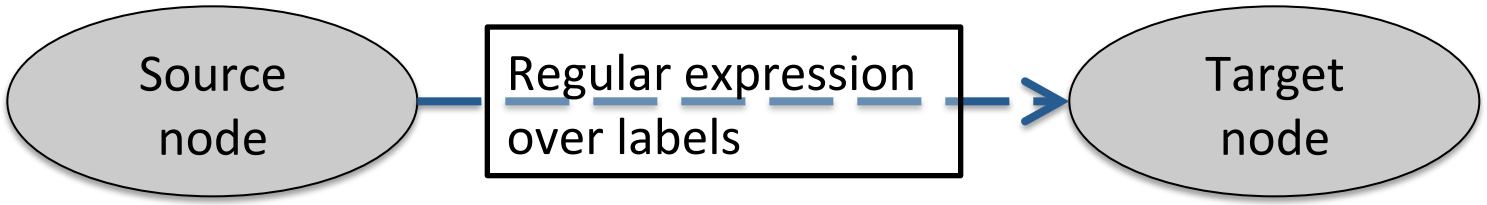
Step 2



Question parsing

- Task: Given a question and two answers, produce two queries
  - One for each answer

- Query structure

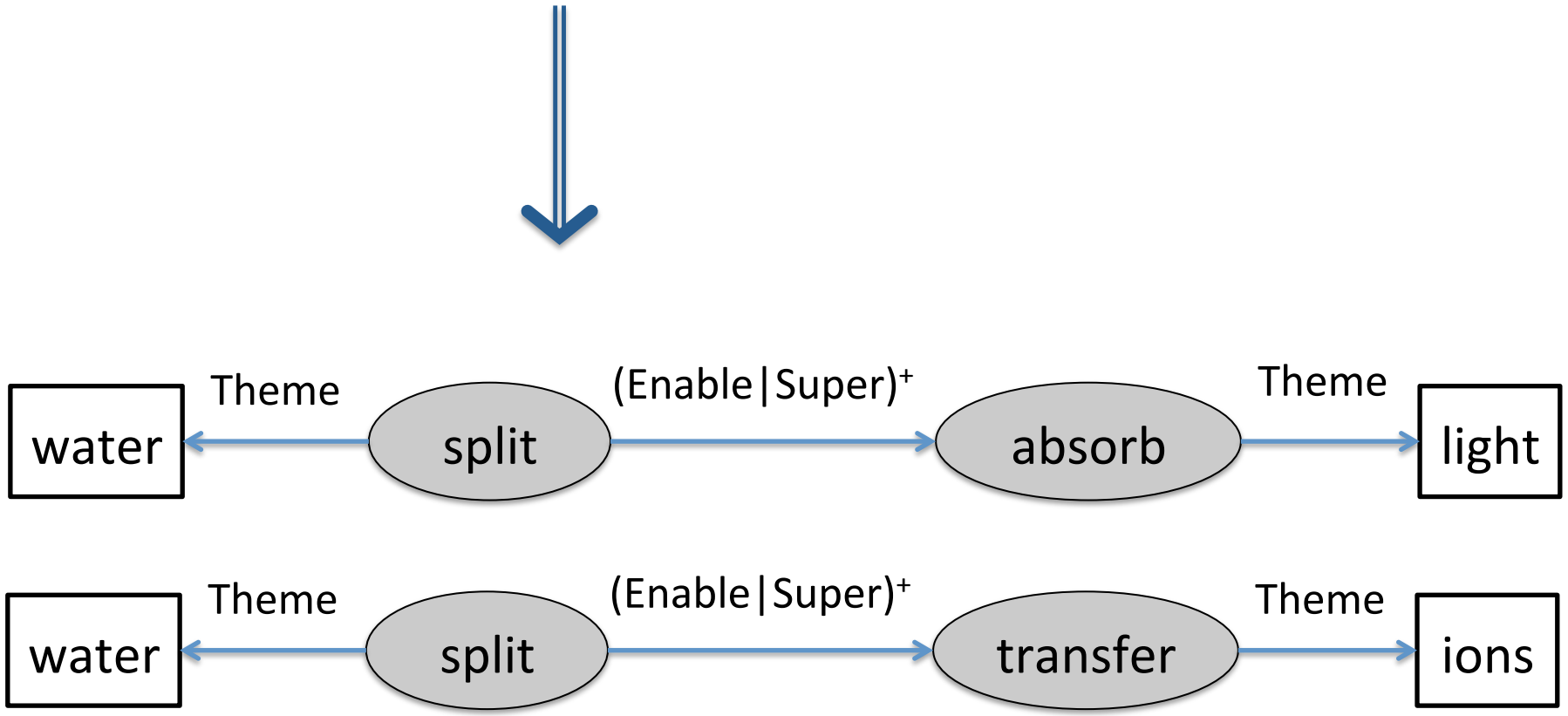


Parsing question to produce formal queries

What does the splitting of water lead to?

A: Light absorption

B: Transfer of ions



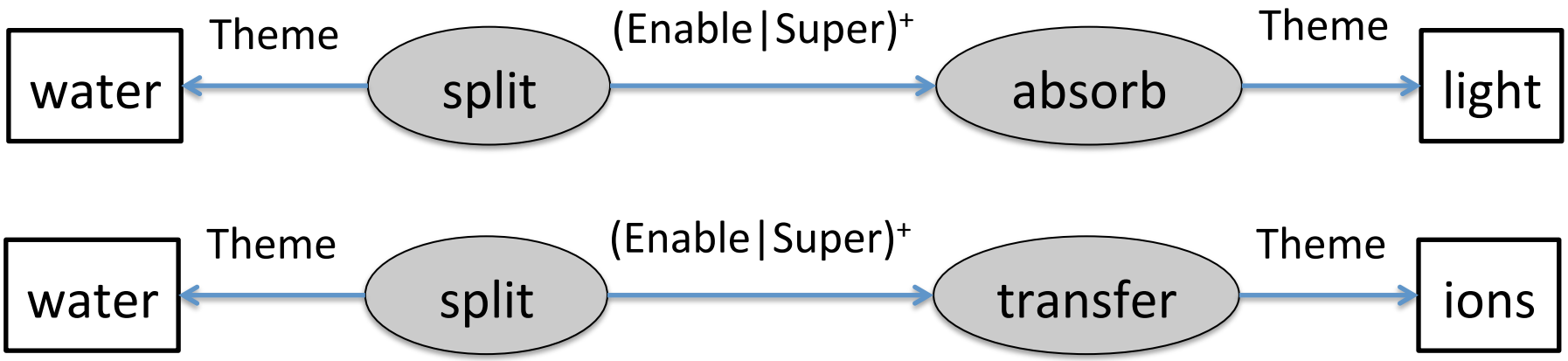
Parsing question to produce formal queries

What does the splitting of water lead to?

A: Light absorption

B: Transfer of ions

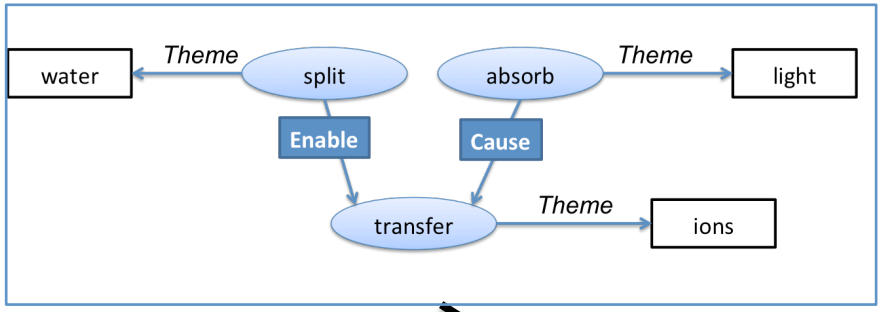
1. Align Q&A triggers and arguments to structure
2. Identify source and target
3. Identify regular expressions
   From small set (~10)



Where are we?

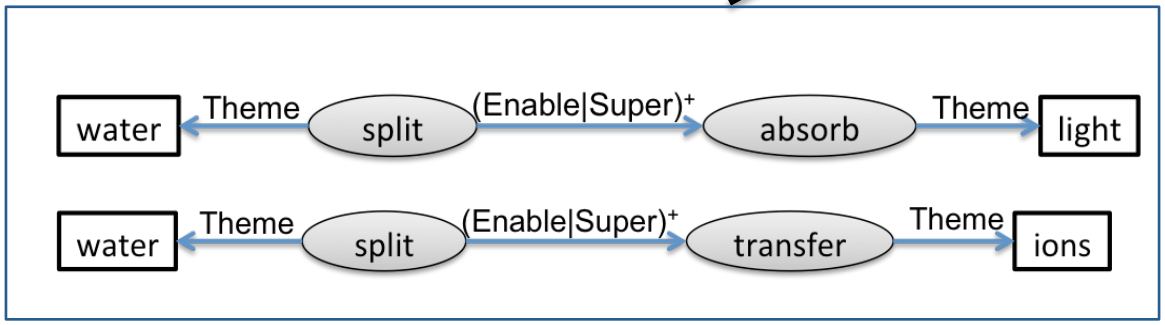
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions



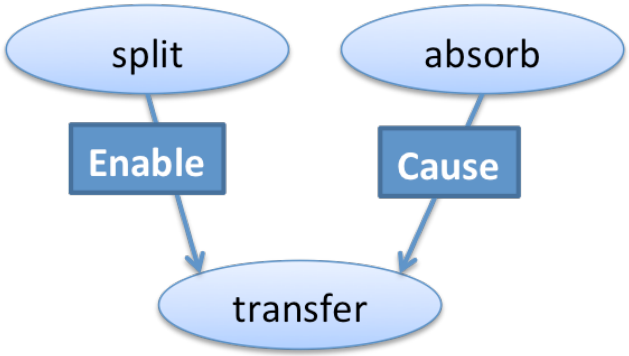
Answering Question

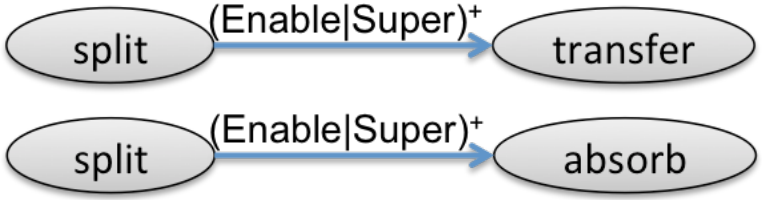
Step 3: Answer = **B**



Step 3: Answering questions

- Given
  - Process structure
  - Two queries

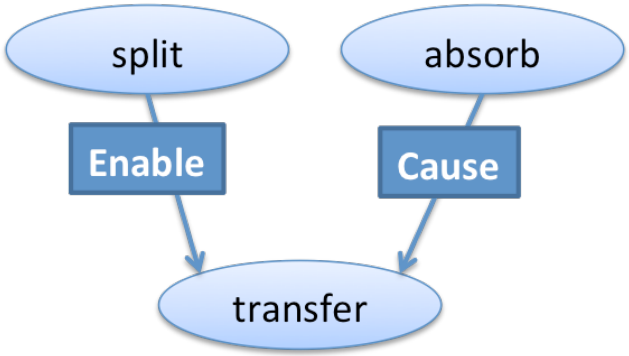


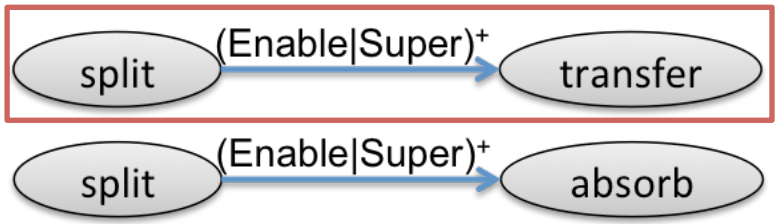


- Answering algorithm

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

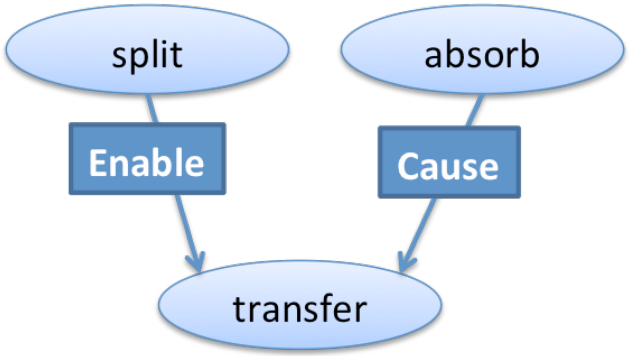


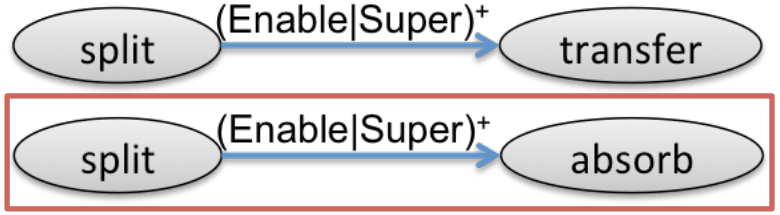


- Answering algorithm
  1. Find matching path (valid proof)

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

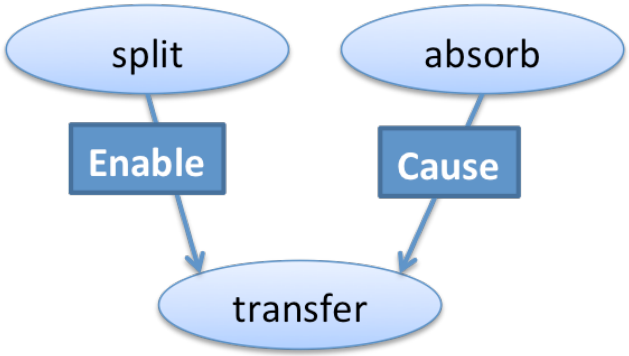


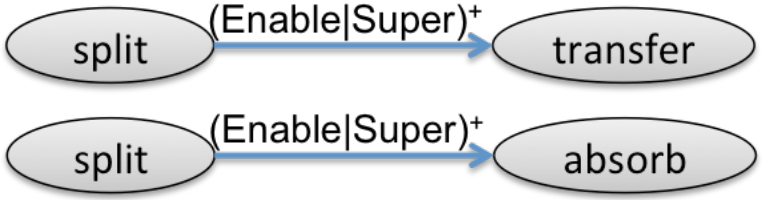


- Answering algorithm
  1. Find matching path (valid proof)
  2. Else, find contradiction of causality (refutation)

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

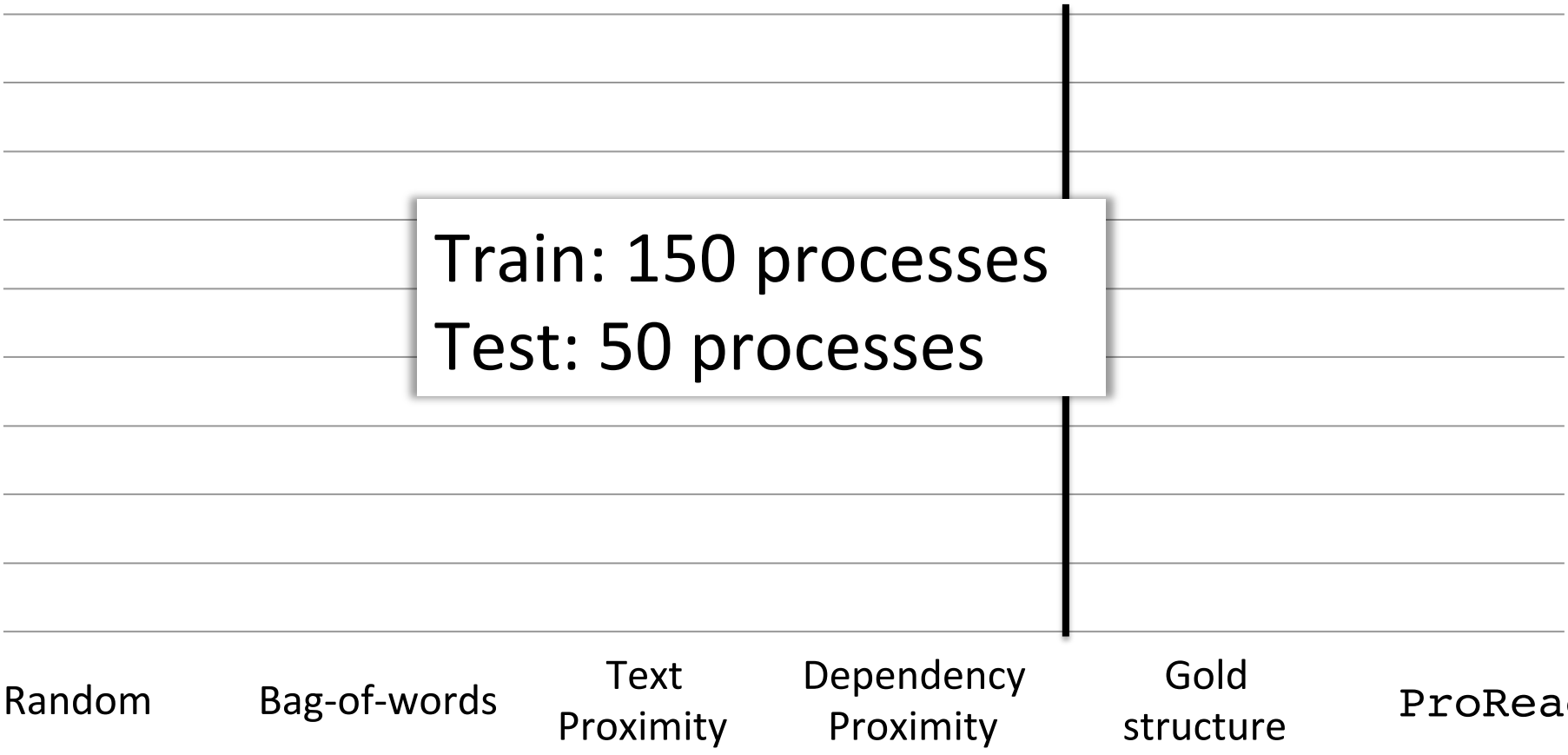




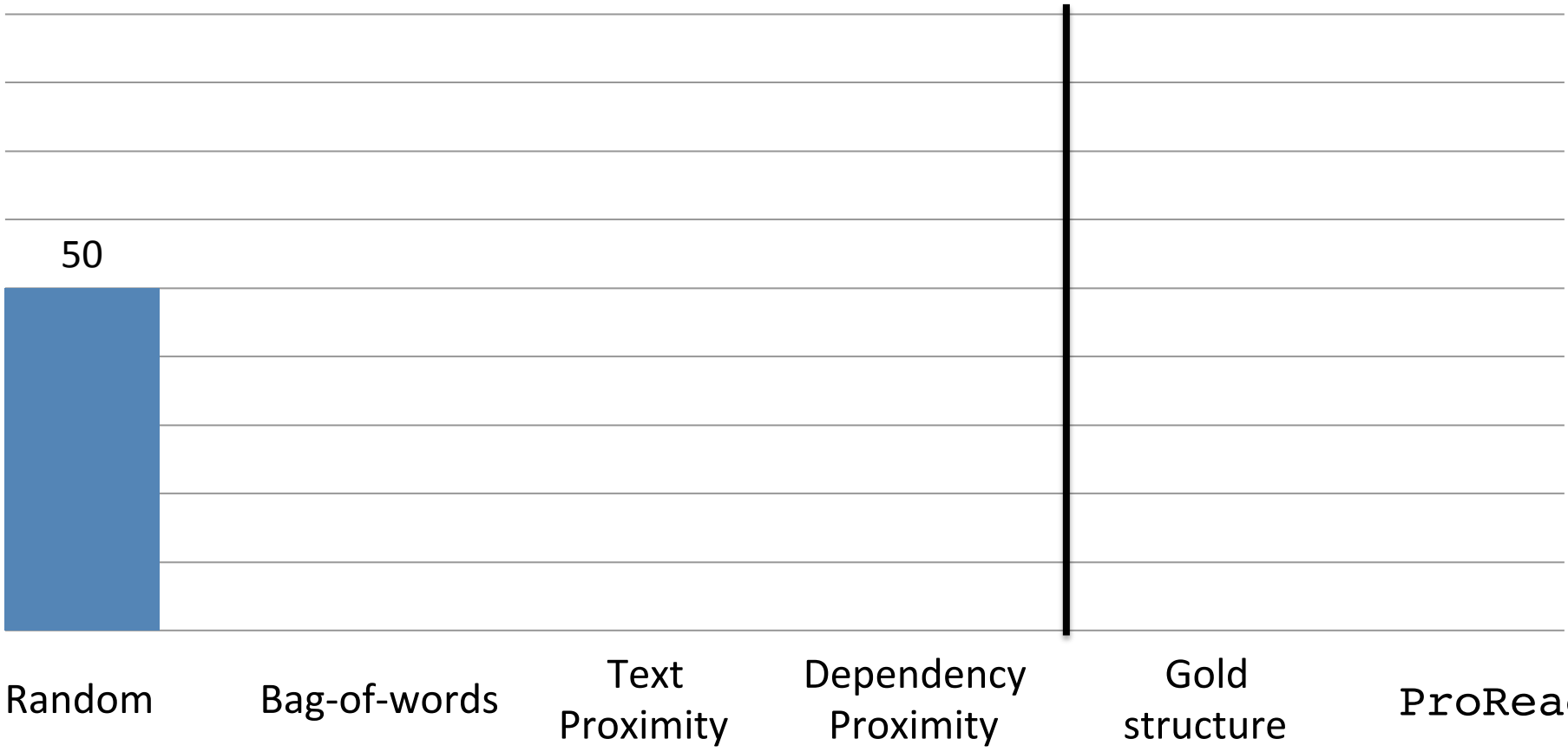
- Answering algorithm
  1. Find matching path (valid proof)
  2. Else, find contradiction of causality (refutation)
  3. Back off to baseline

Experiments and Results

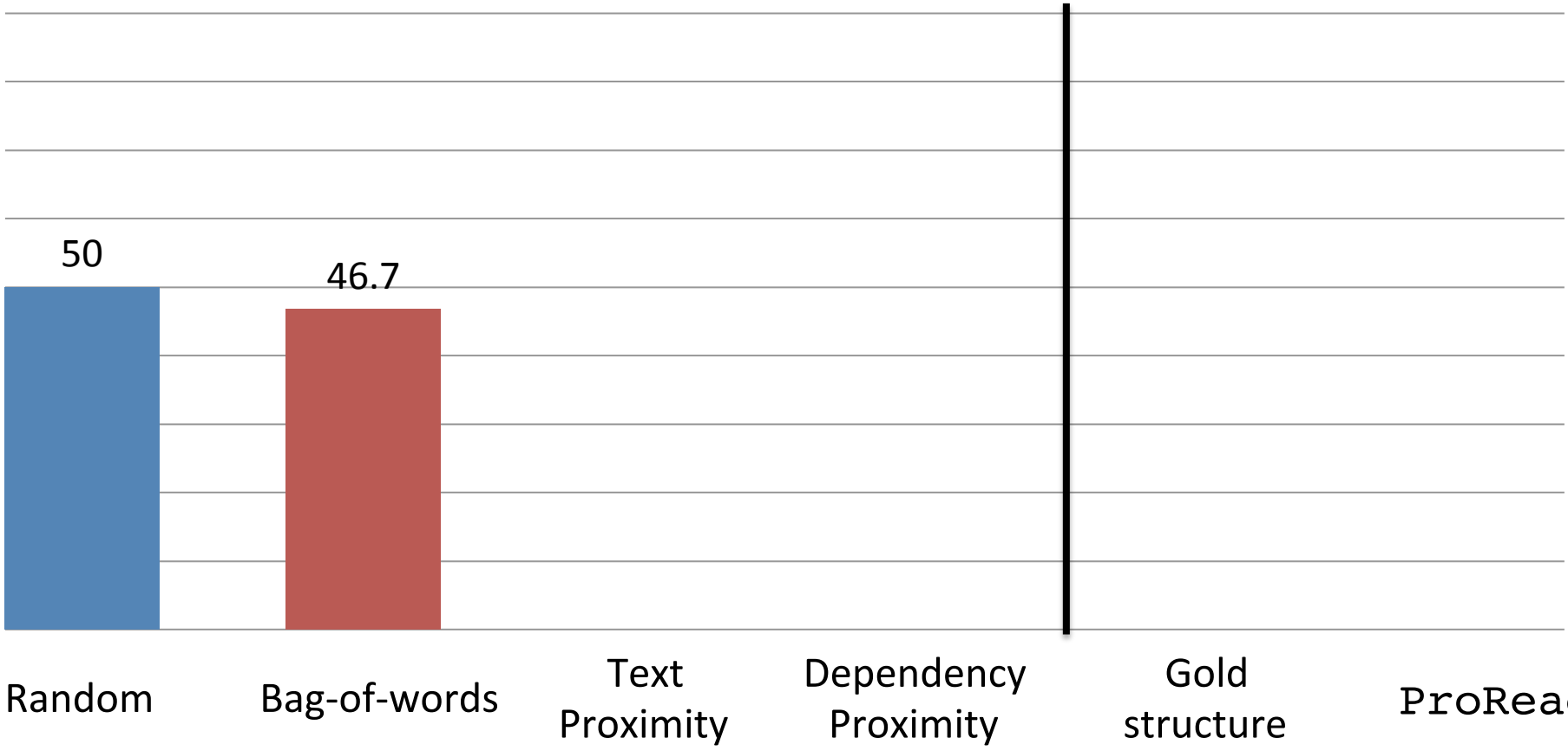
Question Answering Accuracy



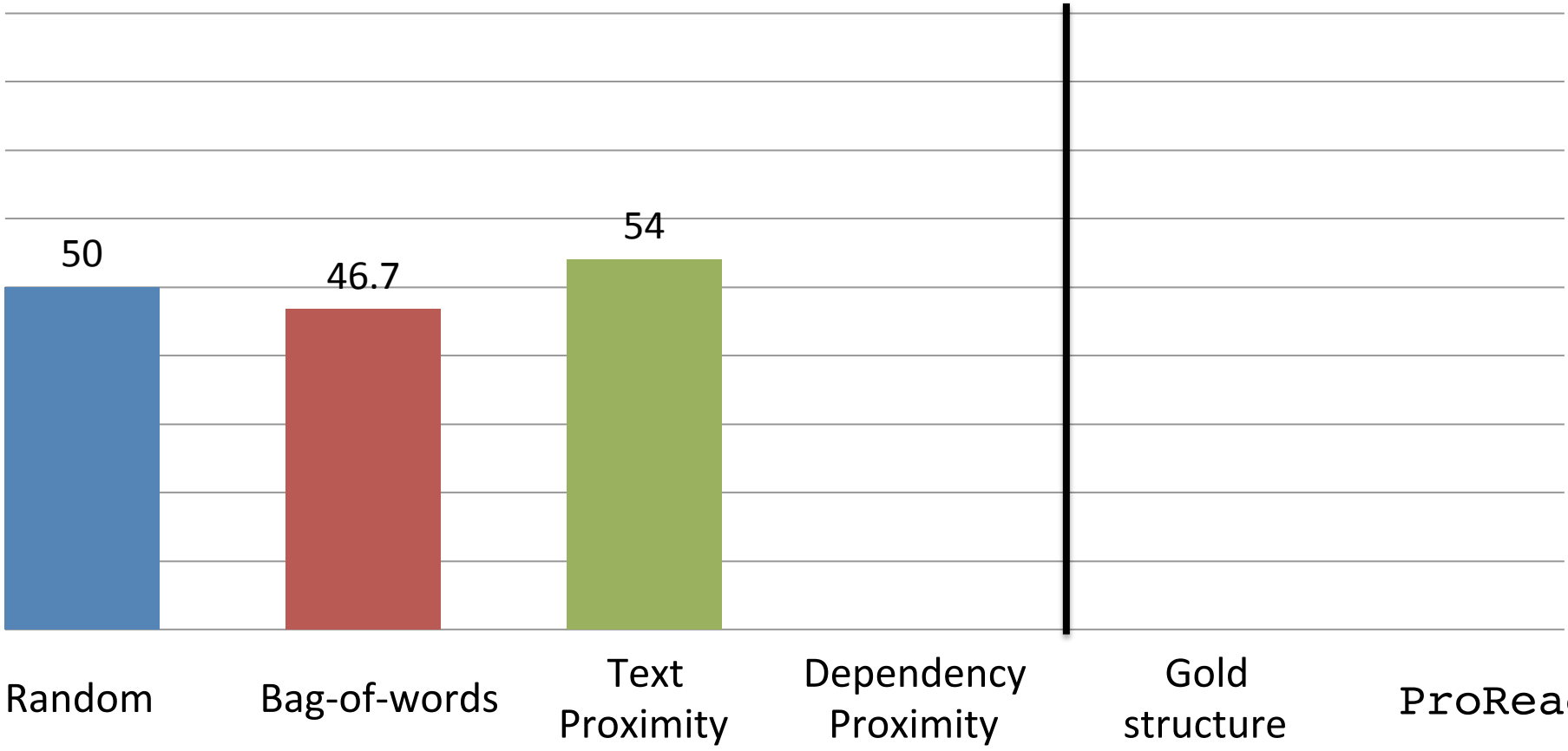
Question Answering Accuracy



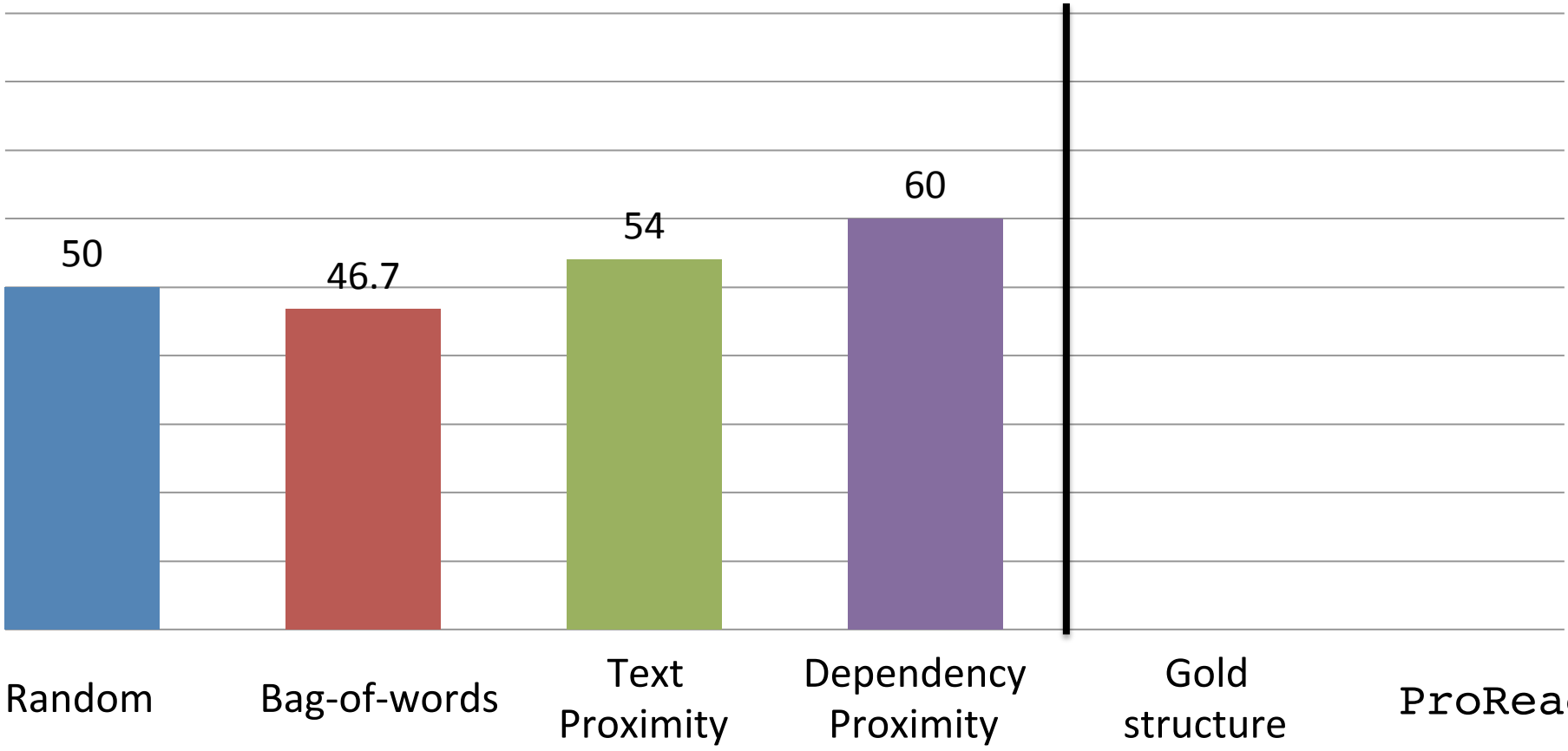
Question Answering Accuracy



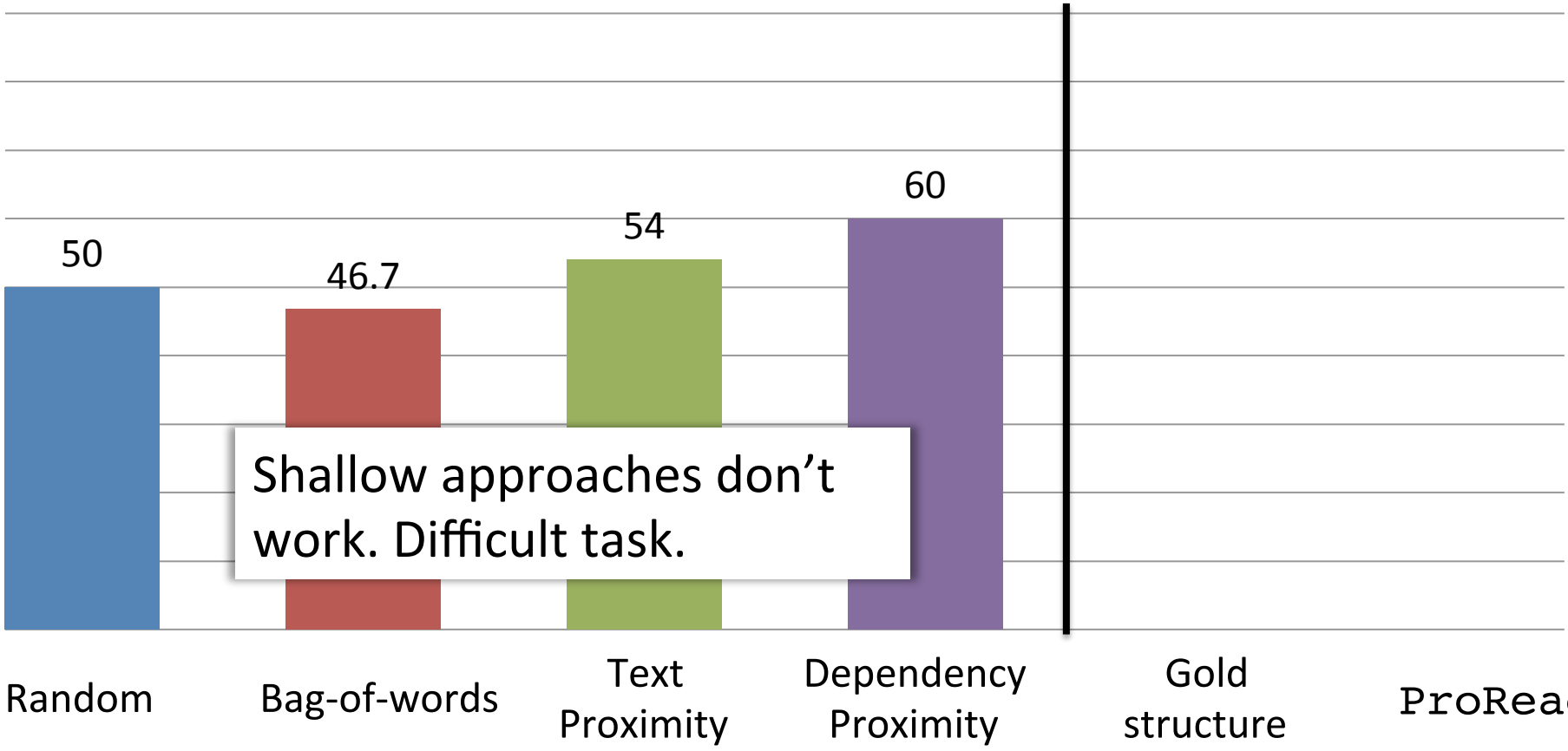
Question Answering Accuracy



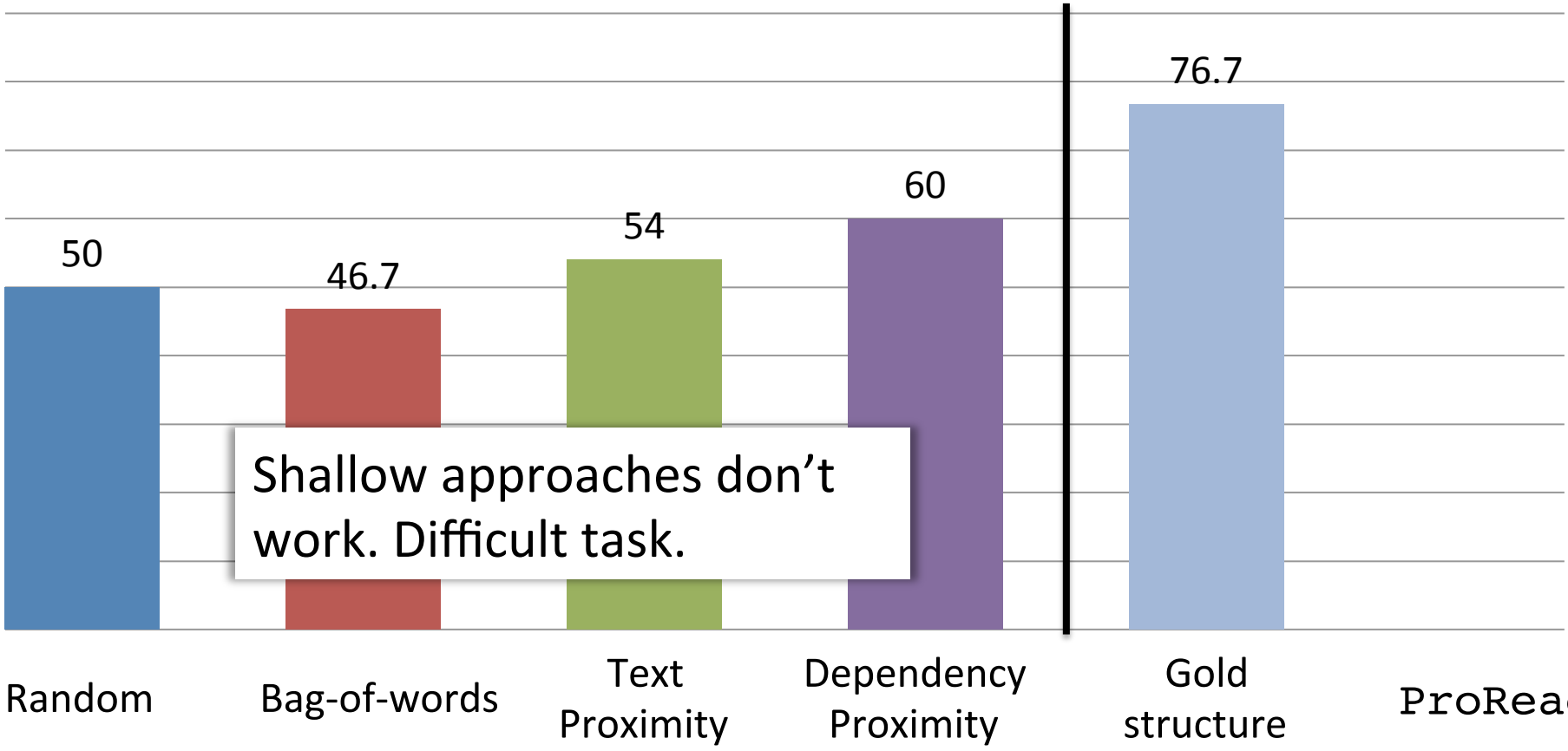
Question Answering Accuracy



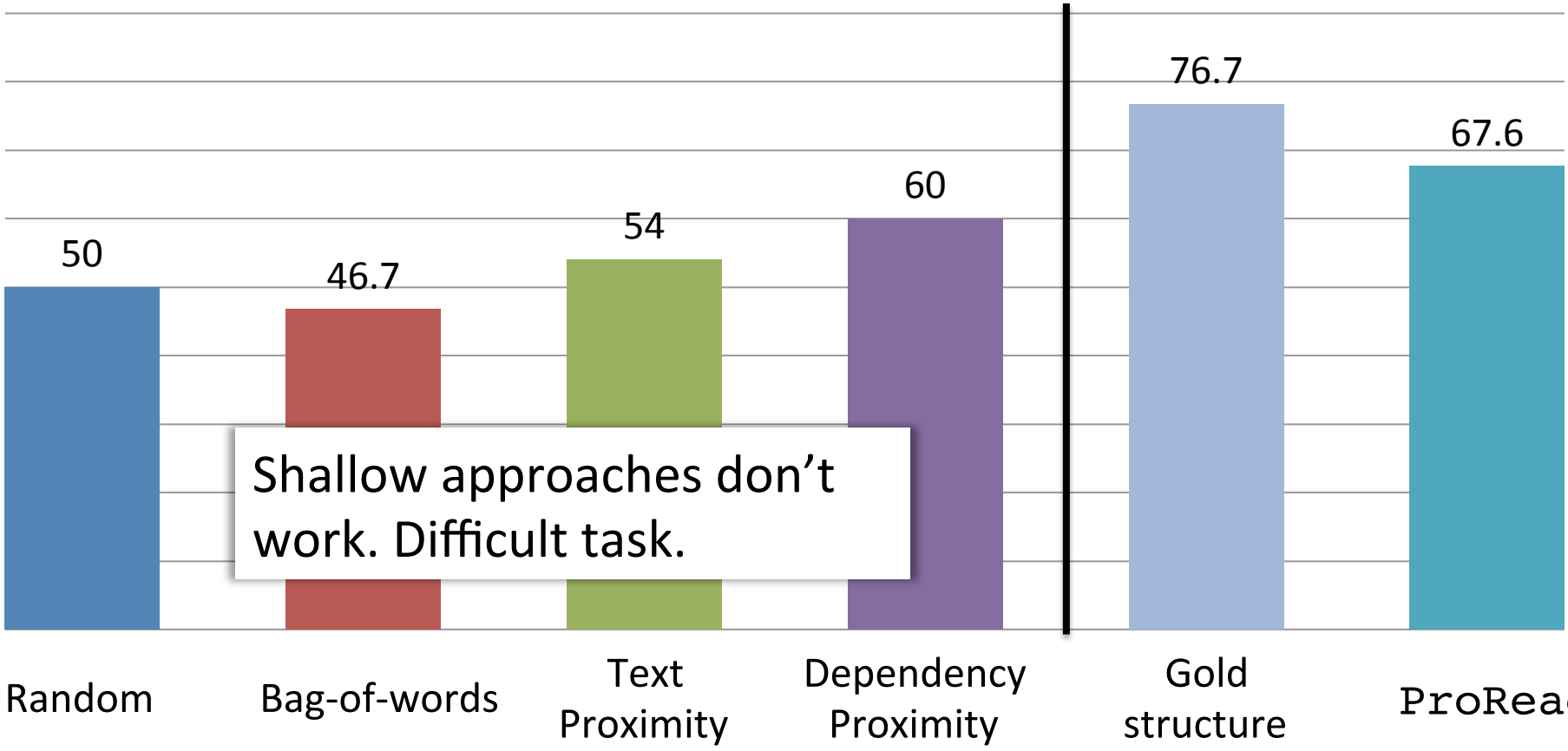
Question Answering Accuracy



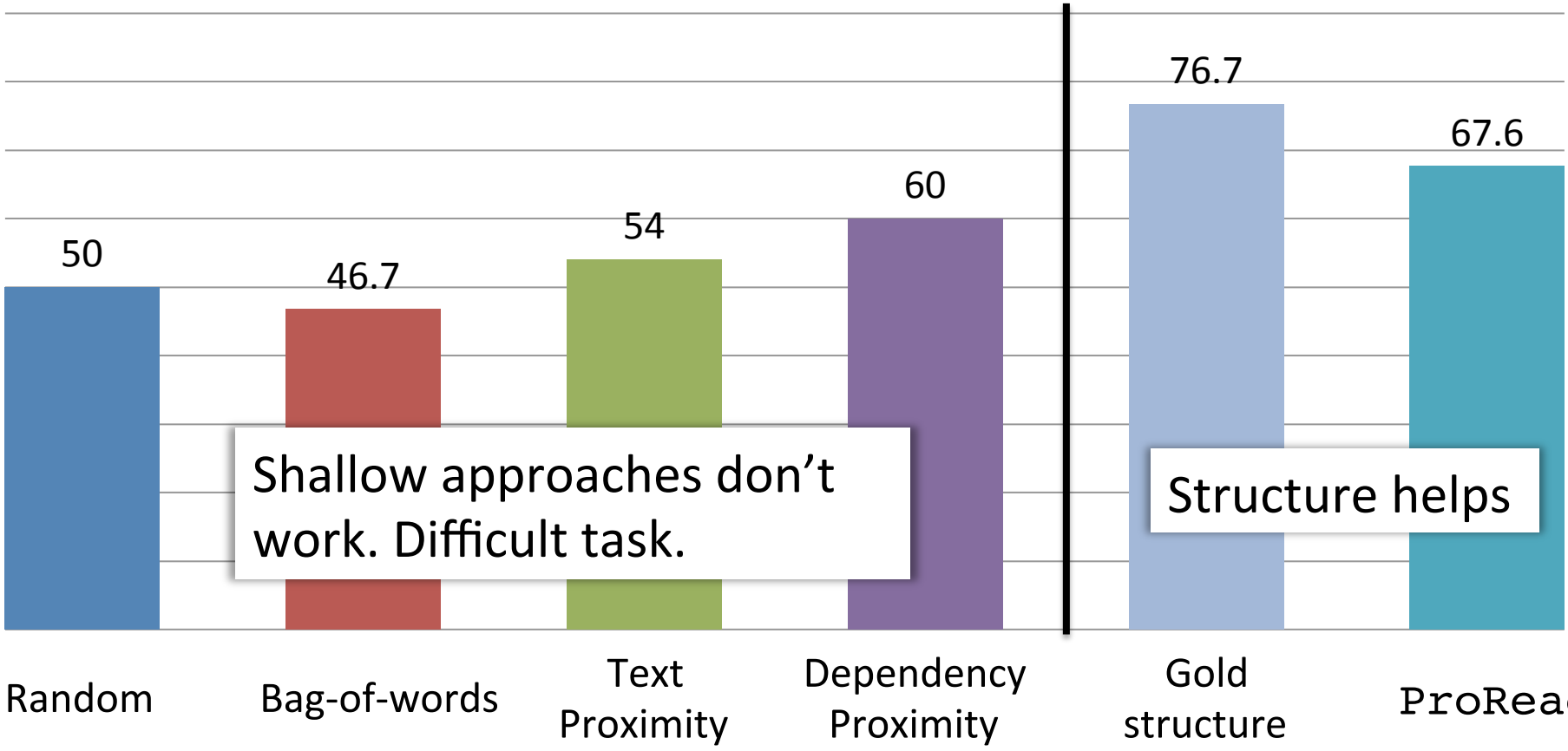
Question Answering Accuracy



Question Answering Accuracy



Question Answering Accuracy



Reading comprehension errors

- Errors with predicted structure
  - Structure error (55%)
  - Alignment (10%)
  - Annotation mismatch (10%)
  - Entity coreference (10%)

- Errors with gold structure
  - Alignment (35%)
  - Annotation mismatch (25%)
  - Entity coreference (20%)

Summary

This work

- A new reading comprehension task
- A dataset of structures with Q&A: `ProcessBank`
- An end-to-end system for question answering via predicted structures
- Rich entity and event structure helps

Summary

This work

- A new reading comprehension task
- A dataset of structures with Q&A: `ProcessBank`
- An end-to-end system for question answering via predicted structures
- Rich entity and event structure helps

Bigger picture

- Wanted: Programs that read, understand & reason about text
- Processes are complex phenomena that are extensively described in text $\Rightarrow$ Great test-bed
- Ties together many NLP strands: information extraction, semantic role labeling, semantic parsing and reading comprehension
- Towards machine reading

Summary

This work

- A new reading comprehension task
- A dataset of structures with Q&A: `ProcessBank`
- An end-to-end system for question answering via predicted structures
- Rich entity and event structure helps

Bigger picture

- Wanted: Programs that read, understand & reason about text
- Processes are complex phenomena that are extensively described in text $\Rightarrow$ Great test-bed
- Ties together many NLP strands: information extraction, semantic role labeling, semantic parsing and reading comprehension
- Towards machine reading

Thank you!

`http://nlp.stanford.edu/software/bioprocess/`

Extra slides

Non-factoid questions: AP Exams

In the development of a seedling, which of the following will be the last to occur?

A. Initiation of the breakdown of the food reserve
B. Initiation of cell division in the root meristem
C. Emergence of the root
D. Emergence and greening of the first true foliage leaves
E. Imbibition of water by the seed

Actual order: E A C B D

Temporal vs. Causal dependencies

Related: Temporal relations between biological events

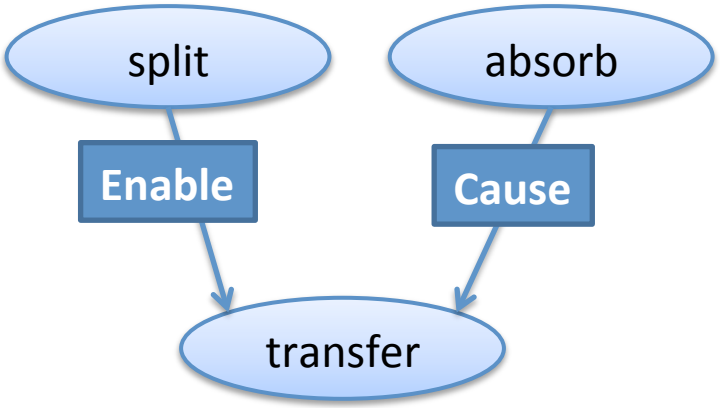
[Scaria, et al, 2013]

- Causal dependencies $\rightarrow$ Relevant temporal relations

  split and **absorb** before **transfer**

- What about the others?

  Temporal ordering of **split**, **absorb** unspecified

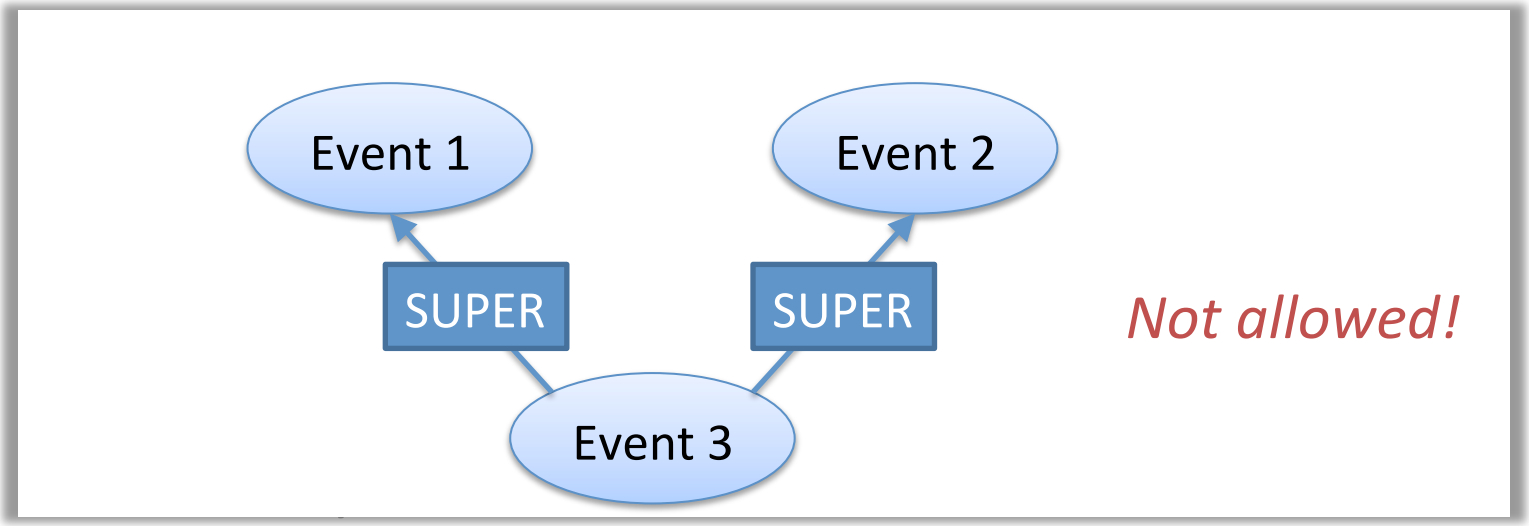


Joint inference with constraints

1. No overlapping arguments
2. Maximum number of arguments per trigger
3. Maximum number of triggers per entity
4. Connectivity
5. Unique SUPER parent
6. Events that share arguments must be related

Joint inference with constraints

1.
2.
3.
4.

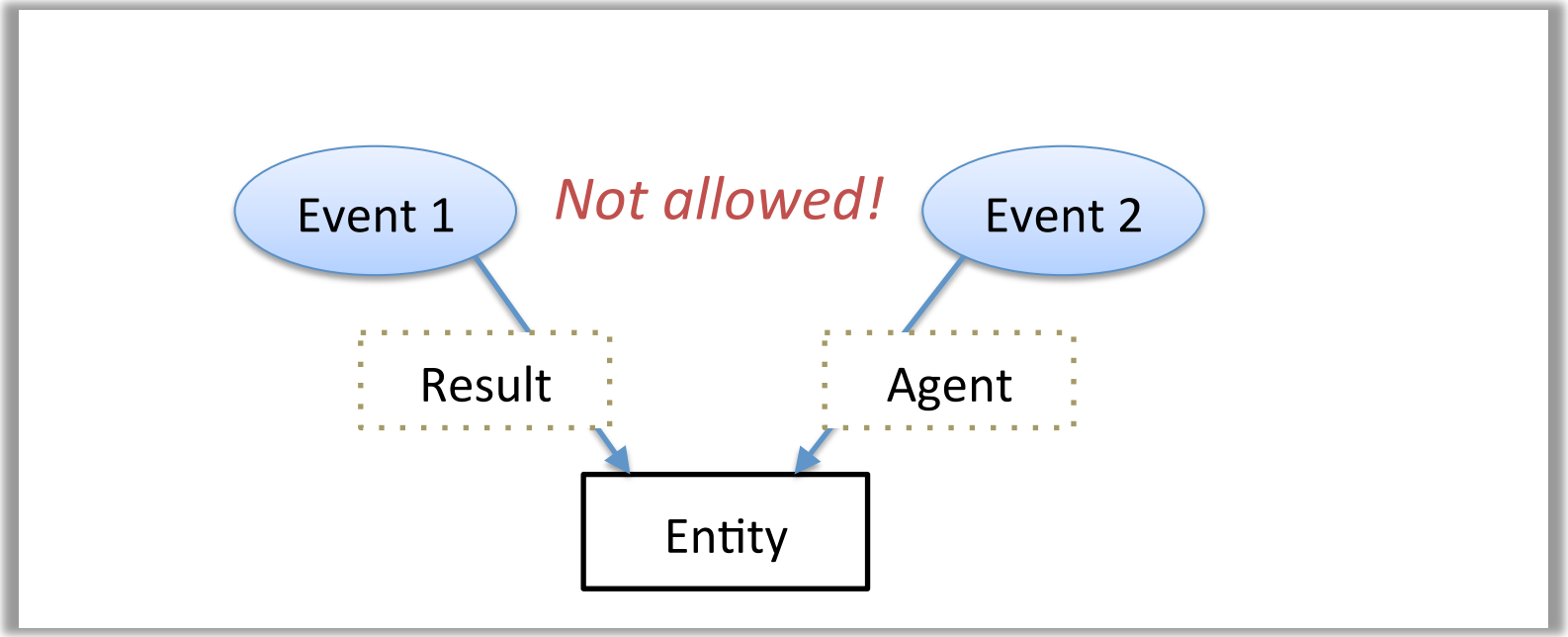


5. ***Unique SUPER parent***

6. Events that share arguments must be related

Joint inference with constraints

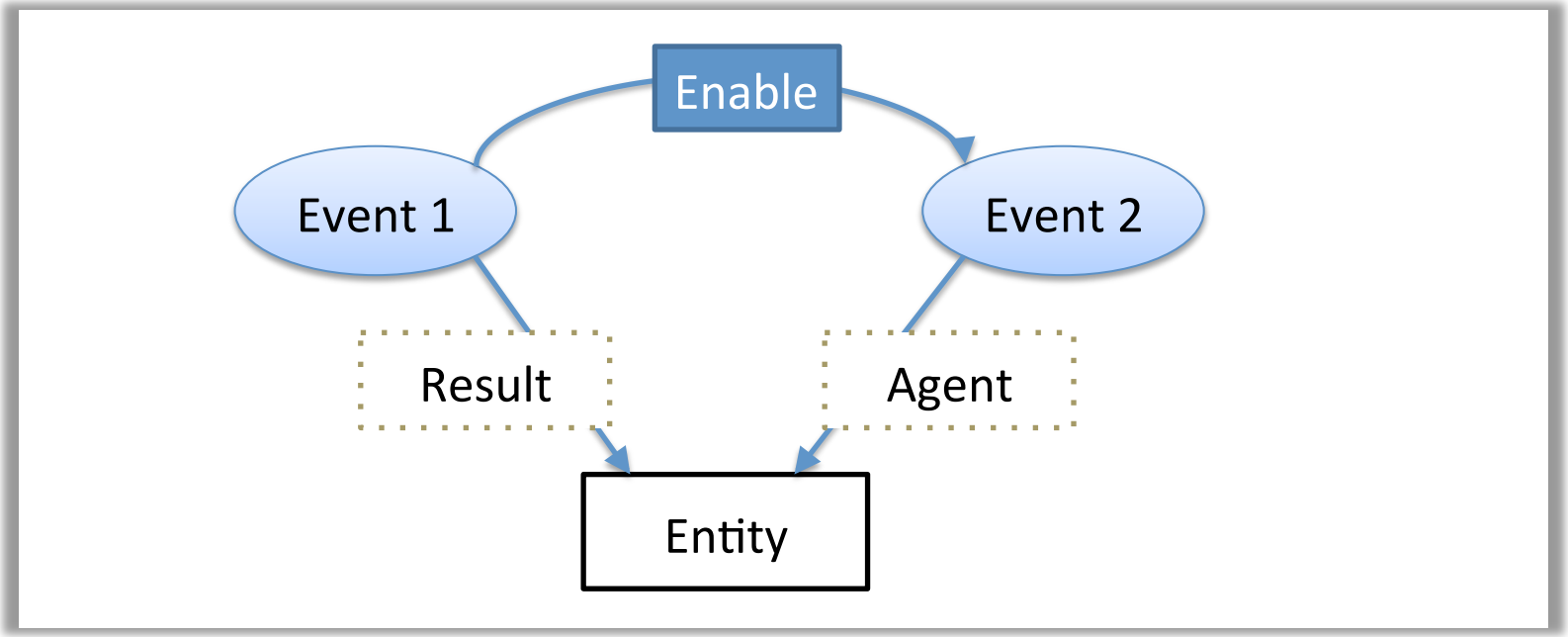
1.
2.
3.
4.
5.



6. ***Events that share arguments must be related***

Joint inference with constraints

1.
2.
3.
4.
5.



6. ***Events that share arguments must be related***

Trigger classification

- Regularized logistic regression
- Features
  - Syntactic: POS tag, path to root, …
  - Semantic
    - WordNet
    - NomLex
    - Levin verb classes
    - Gazetteer from Wikipedia
  - Syntactic and linear context

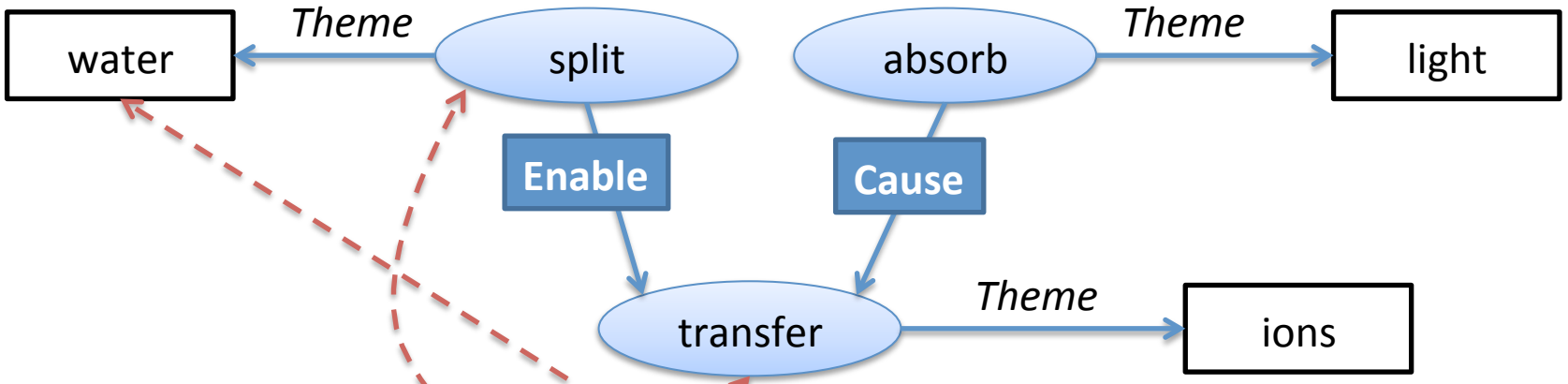
Event-arguments and event-event relations

$$\max_{\mathbf{y},\mathbf{z}} \sum_{t,a,L} y_{t,a,L} b_{t,a,L} + \sum_{t_1,t_2,R} z_{t_1,t_2,R} c_{t_1,t_2,R}$$

- $\mathbf{y}$, $\mathbf{z}$: decision variables for inference
- $\mathbf{b}$, $\mathbf{c}$: scores for corresponding decision
  - Dot products of weights and features

- Features
  - Event-argument features based on semantic role labeling features
  - Event-event features based on [Scaria et al, 2013]
  - Dependency paths, connectives, context, linear distance, clustering features,…

Aligning question-answers to structure



What does the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Source and target identification

What does the splitting of water lead to?

- *"lead to"* = *"cause"* (WordNet)
- Question word (*"what"*) is an indirect object
- Sentence is in active voice

Answer should be the target of query path

Selecting query regular expression

What does the splitting of water lead to?
A: Light absorption
B: Transfer of ions

- "splitting" is an event
- "transfer" and "absorption" are events
- Looking for event $\rightarrow$ event path
- "lead to" = "cause"

$$\texttt{Same}^{*}\texttt{(Enable}\,|\,\texttt{Super)}^{+}\texttt{Same}^{*}$$

Structure Prediction Performance

	Precision	Recall	F1
Triggers	75.8	73.9	74.8
Arg. Iden.	57.3	45.6	50.8
SRL	42.5	33.9	37.7
Relations	26.5	21.8	23.9

Nuanced differences in structures don't always matter