Modeling Linguistic Structure (And What We Can Do With One)

Vivek Srikumar
School of Computing

THE UNIVERSITY OF UTAH

Wanted

Programs that can learn to understand and reason about the world via language

Wanted

Programs that can learn to **understand** and **reason** about the world via language

How can we measure understanding or reasoning?

Are you smarter than a sixth grader?

Are you smarter than a sixth grader?

Martin The Monkey lives in an oak tree in Bananaville. He works for the Banana telephone company. He is the best employee they have because he can climb the telephone poles twice as fast as everyone else.

On Wednesday nights, Martin takes painting lessons with Sarah Able. She is a famous oil painter who lives in the same town as Martin. Martin is one of the only artists in the area that paints with his tail.

***Question*: What town does Sarah Able live in?** Bananaville

Really?

Bananaville ∈ Towns

Sarah Able = She — *Co-reference resolution*

∀ X ∈ Towns, Lives(She, X) ⇔ Lives(Martin, X) — *Semantic parsing* nanaville)

Martin = Martin the Monkey — *Co-reference resolution* (Martin, Bananaville)

Lives(Martin the Monkey, an oak tree) — *Semantic role labeling*

Located-In(an oak tree, Bananaville) → ⇒ Lives(Martin the Monkey, Bananaville)

How do we know this is correct?

Reasoning about the world requires:

(Or seems to require)

- Linguistic analysis
  - coreference, semantic parsing, semantic role labeling, etc
- Encoding world knowledge
  - Eg: If A lives in B and B is located in C, then we can say that A lives in C
- ***Structured*** inference

What are structures and why are they useful?

- For the purpose of this talk
  - A labeled (and possibly directed) graph
  - A discrete object
- Computer science knows how to deal with discrete objects
  - Databases, queries, graph algorithms, etc
- But, natural language is *unstructured*

Let's convert unstructured text into linguistically motivated structured representations

Jonathan Berant

Chris Manning

EMNLP 2014

Modeling Biological Processes for Reading Comprehension

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!

Enable

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!

Enable

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 $NADP^+$.

Cause

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

The punchline

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 (information extraction)
   - Parse question as query to this knowledge base (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]

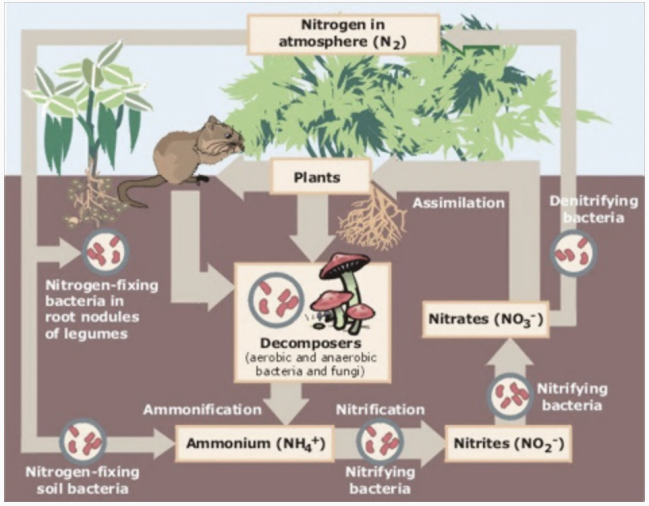
- Micro reading:
  - Single document
  - Requires reasoning
  - Non-factoid questions

  [Richardson et al., 2013, Kushman et al., 2014]

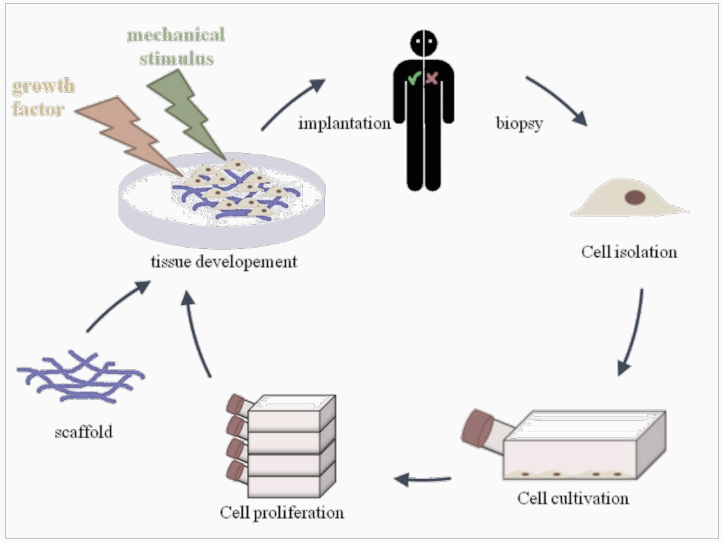
Chosen domain:

Biological process descriptions

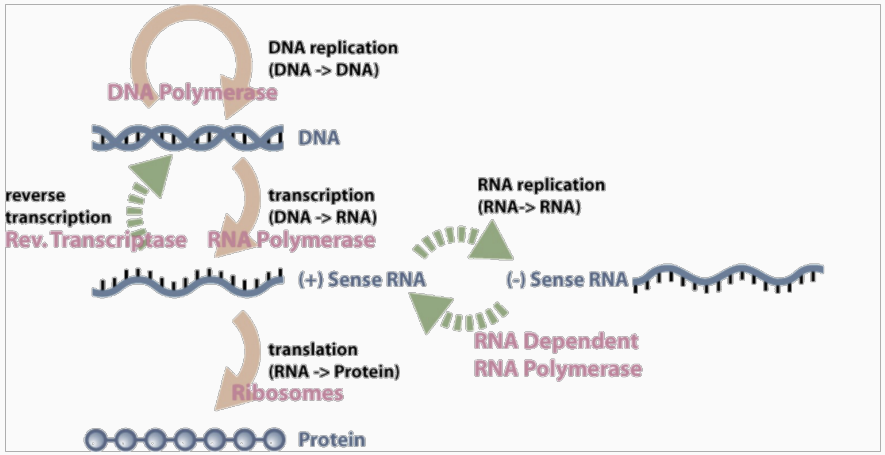
Processes abound in biology



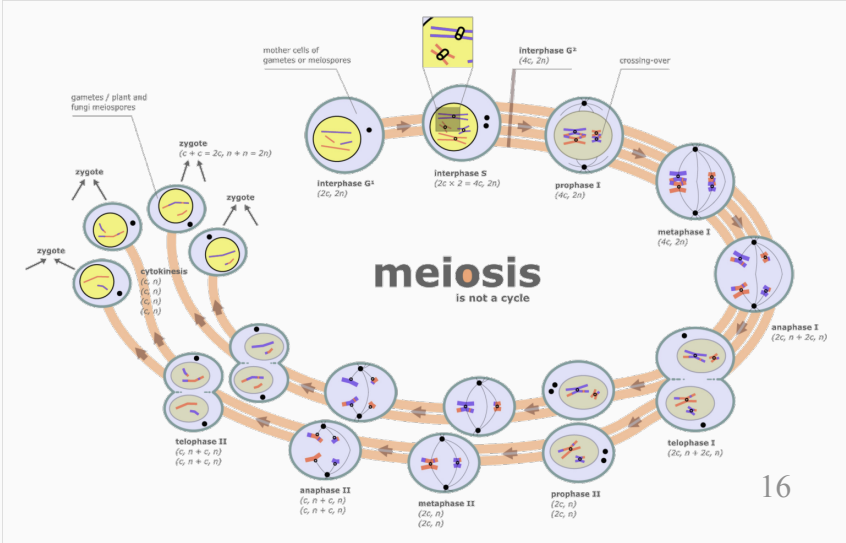
Nitrogen Cycle



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*

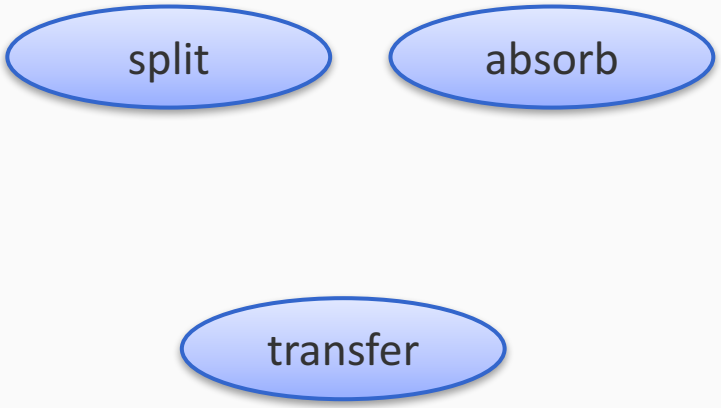
A: True
B: False

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

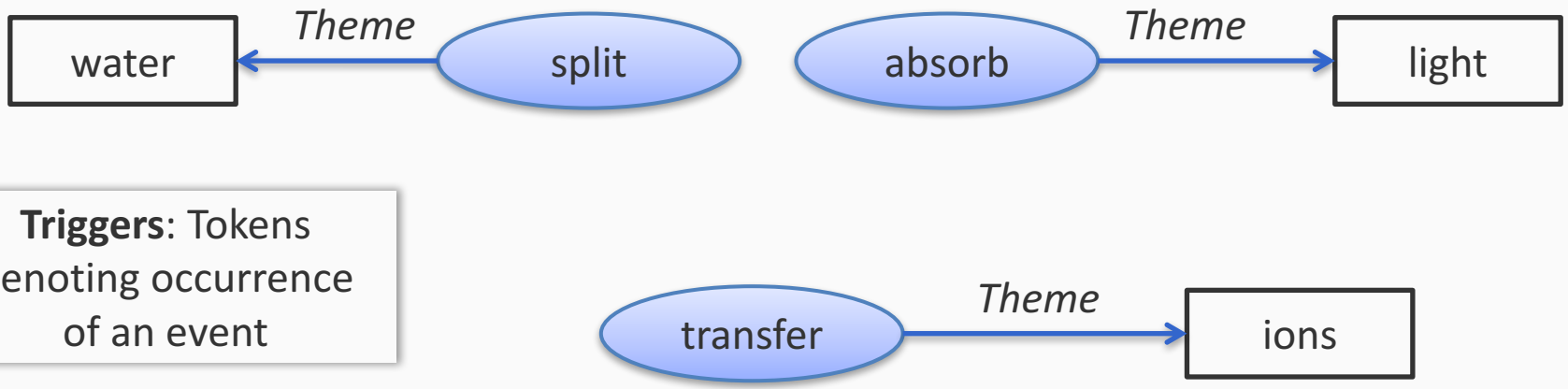
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.



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 $NADP^+$.

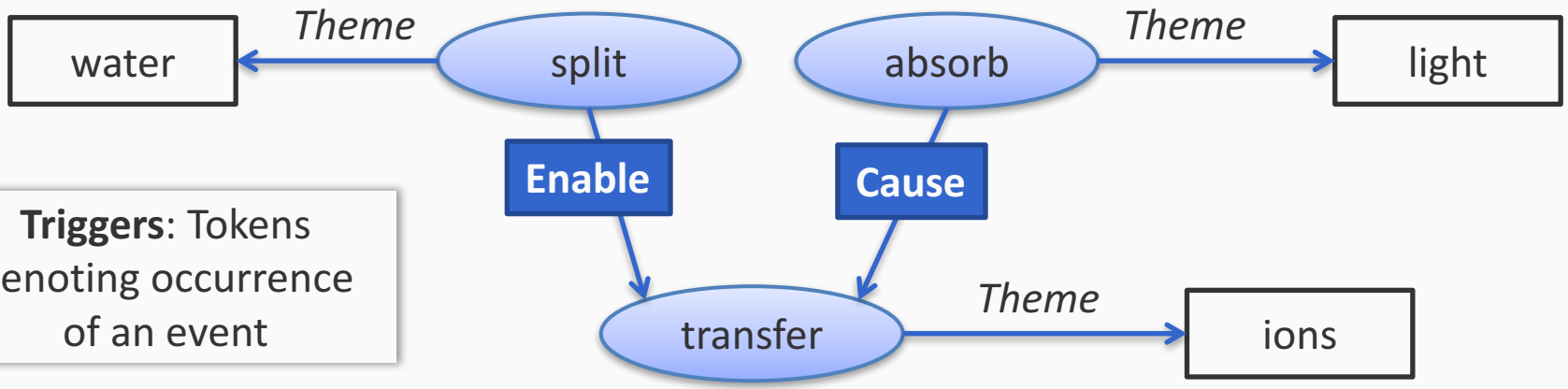


Triggers: Tokens denoting occurrence of an event

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*

Event-event 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
- There are more nuances to the annotation
  - 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
  1. Non-factoid reading comprehension questions
  2. 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

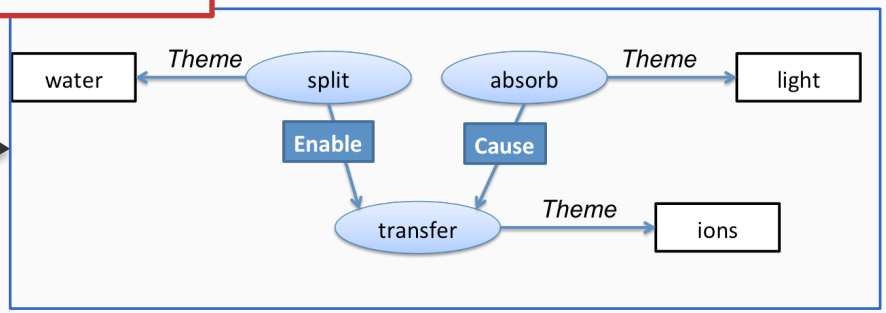
B: Transfer of ions

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

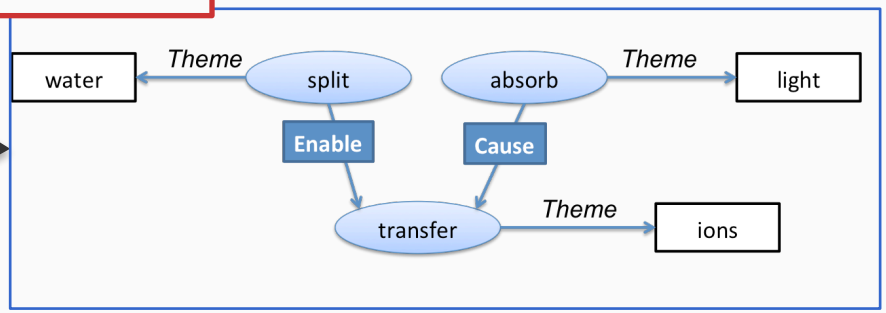
B: Transfer of ions

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



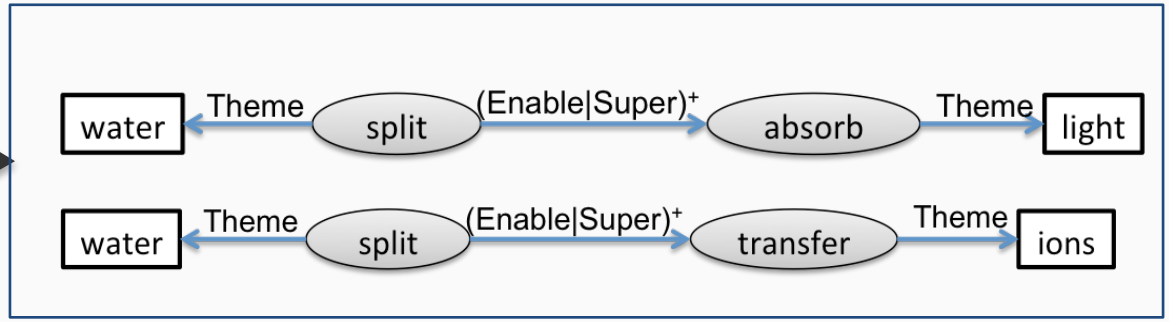
What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Step 2

Question Parsing

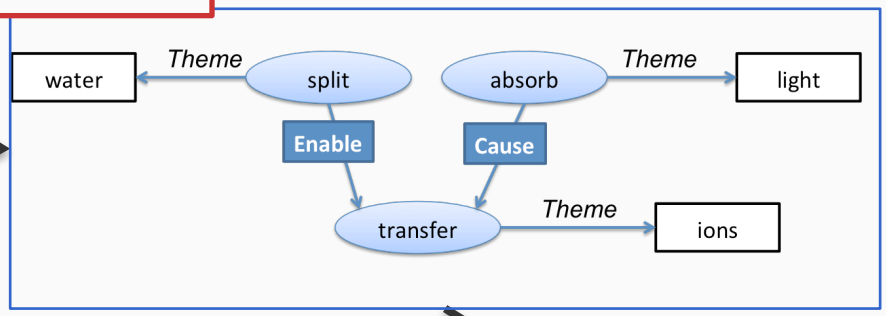


System in a nutshell

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



Answering Question

Step 3: Answer = **B**

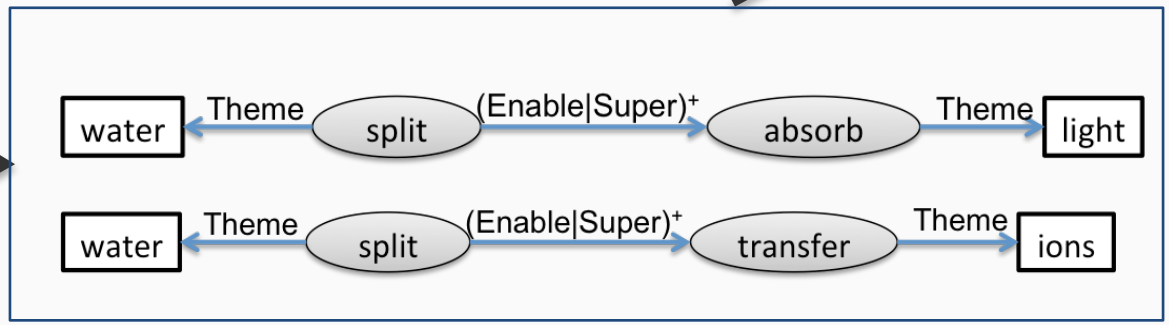
What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

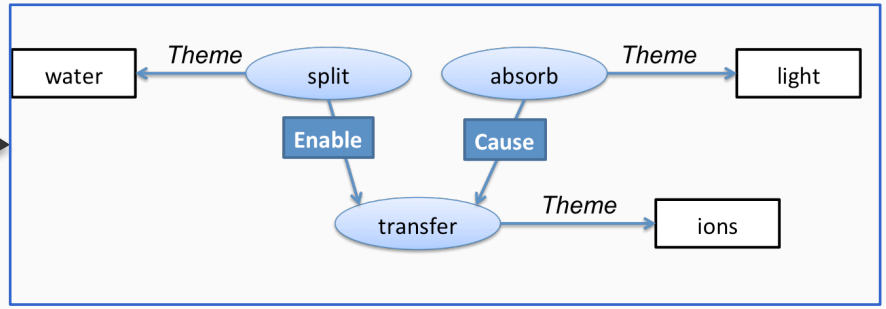
Step 2

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 event ***trigger identifier***

Logistic regression; features from words, lists

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 $NADP^+$.

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 $NADP^+$.

Process structure prediction

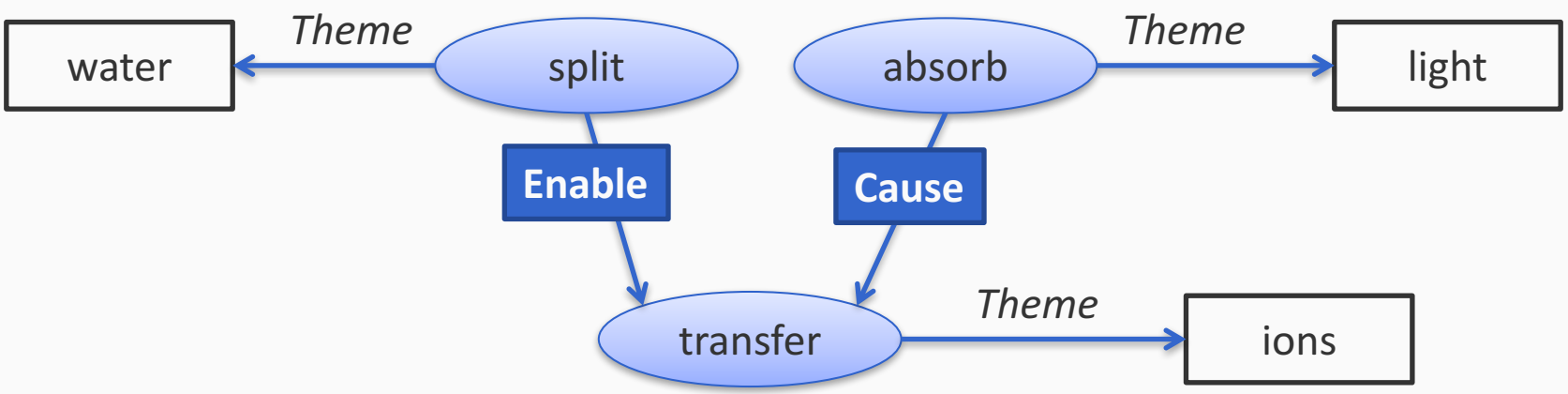
1. Train event trigger identifier

   Logistic regression; features from words, lists

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

Event-arguments and event-event relations

$$\max \boxed{\text{Total event-argument score}} + \boxed{\text{Total event-event relation score}}$$



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

Event-arguments and event-event relations

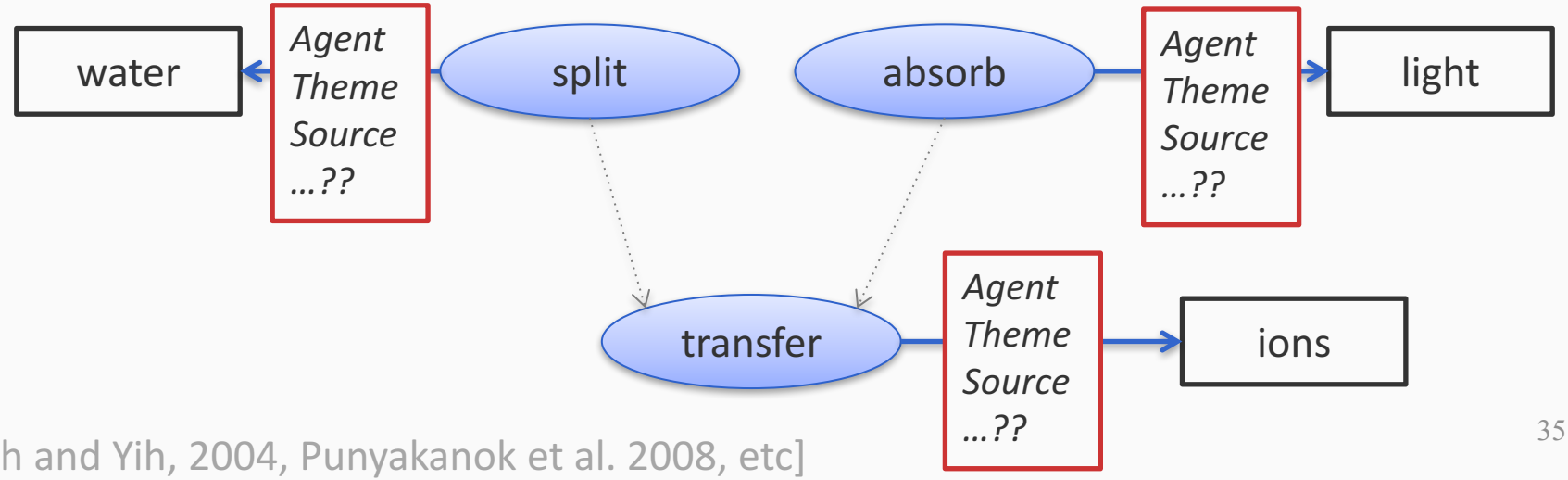
Total event-argument score

$$\max \sum_{t,a,L} \text{Score that arg-candidate has label L} + \text{Total event-event relation score}$$

Trigger t

Argument candidate a

Argument label L



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

Event-arguments and event-event relations

Total event-event relation score

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

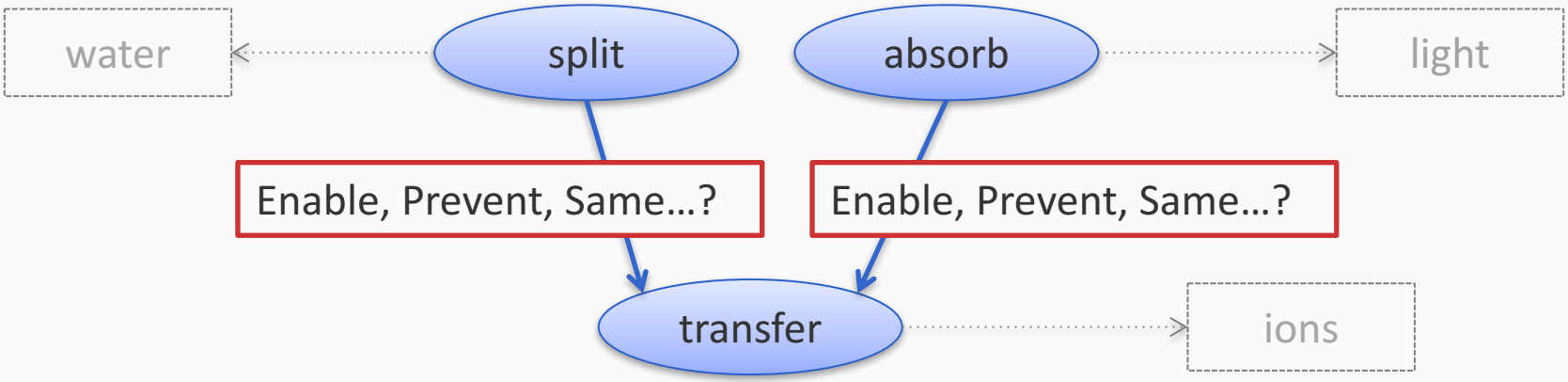
Trigger *t*

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]

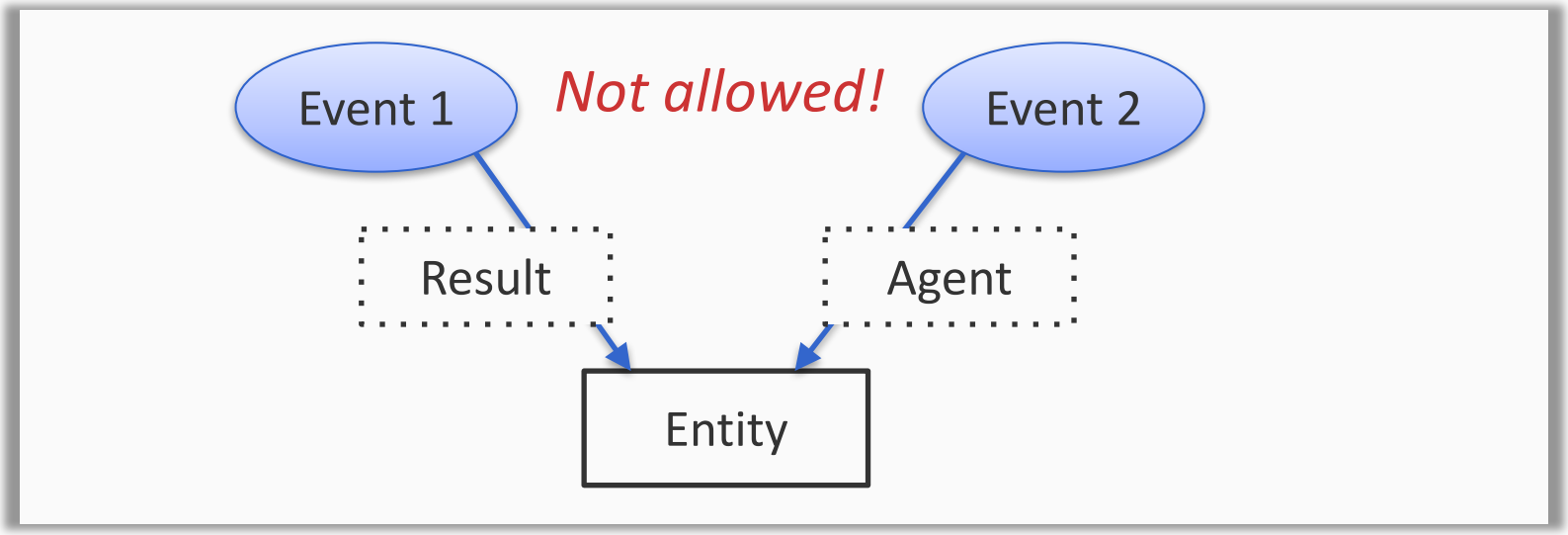
Joint inference with constraints

1. No overlapping arguments
2. Maximum number of arguments per event
3. Maximum number of events 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.

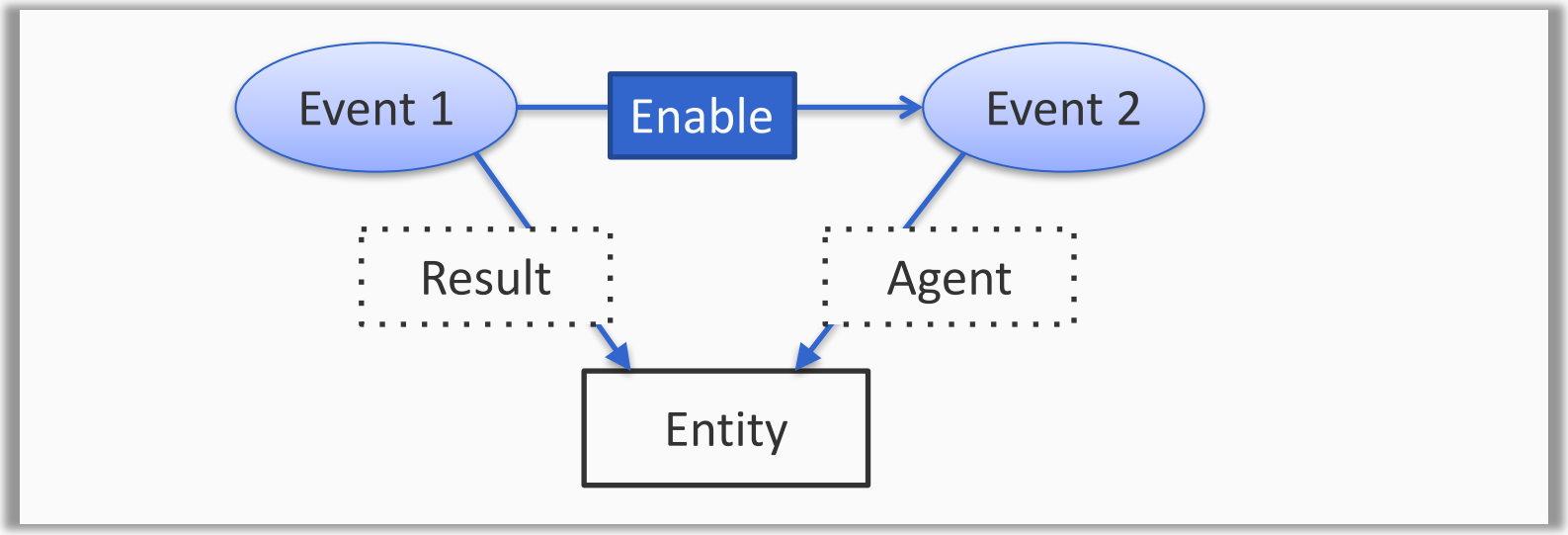


5. ***Events that share arguments must be related***

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

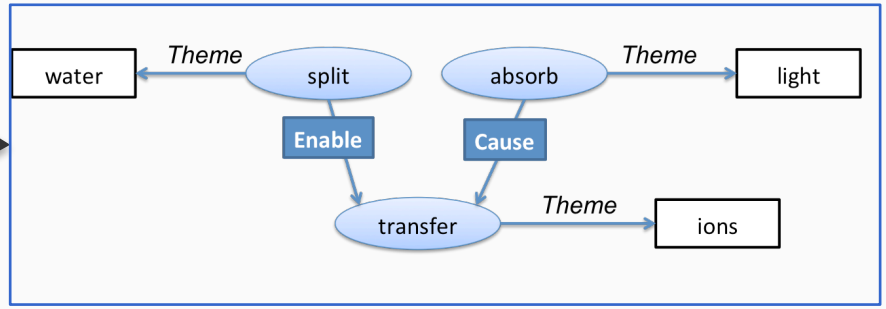
- Linear model to score argument labels and event-event relations
  - Related: Semantic role labeling, information extraction
- Structured averaged perceptron
- Gurobi ILP solver (exact solution)

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



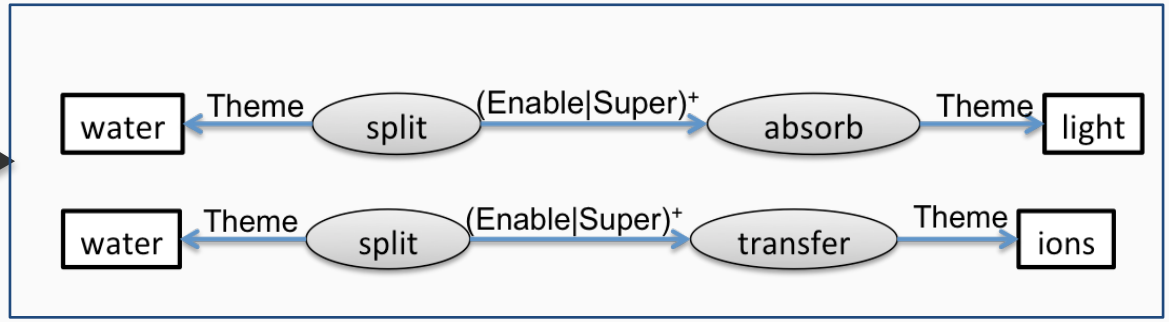
What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Step 2

Question Parsing

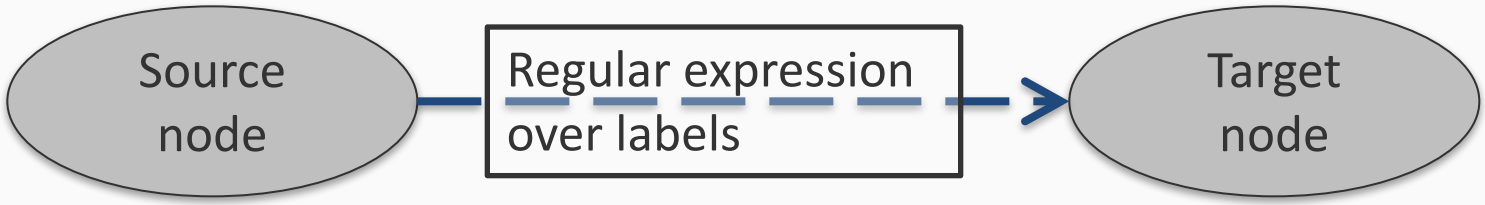


What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions

Question parsing

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

- Query structure

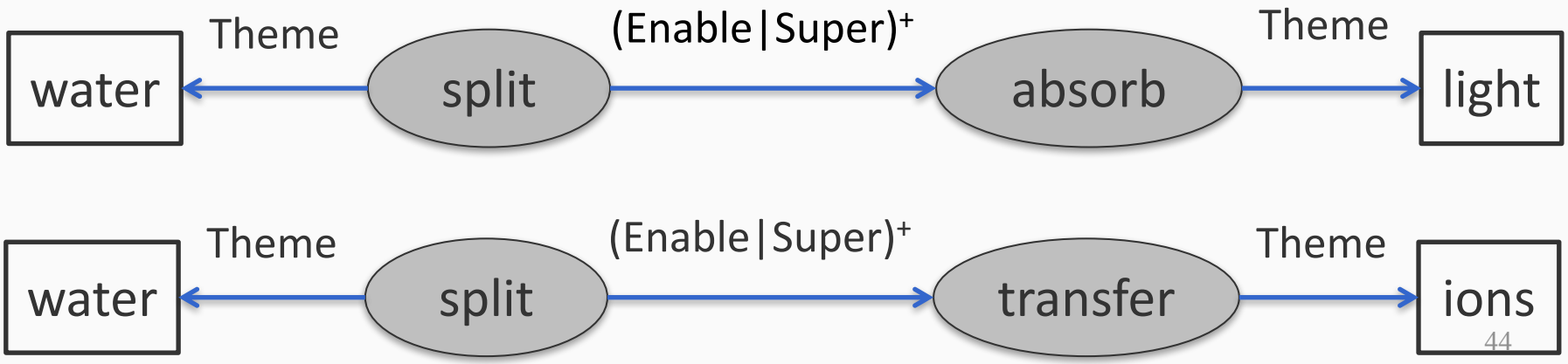


Parsing question to produce formal queries

What does the splitting of water lead to?

A: Light absorption

B: Transfer of ions



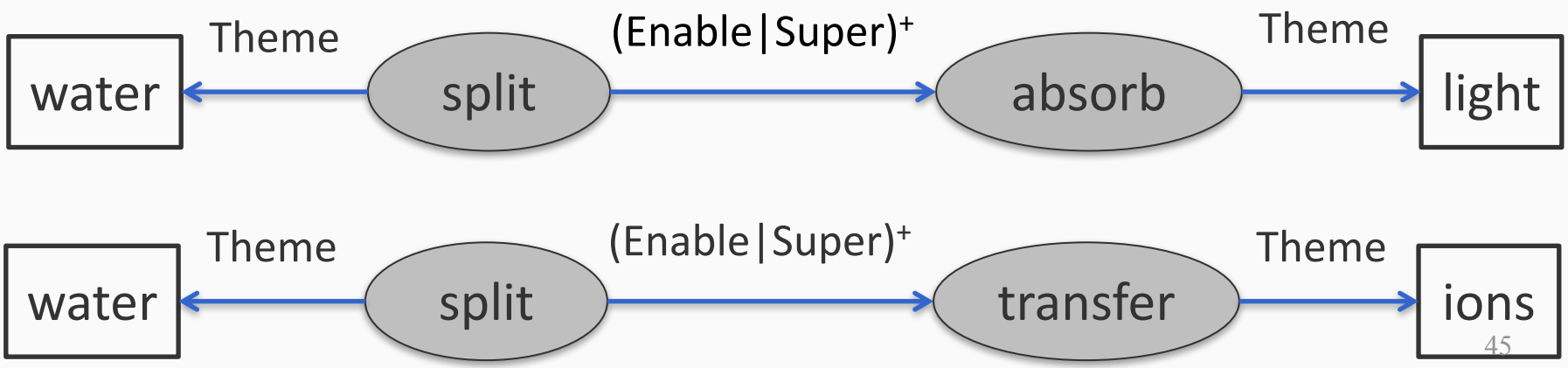
Parsing question to produce formal queries

What does the splitting of water lead to?

A: Light absorption

B: Transfer of ions

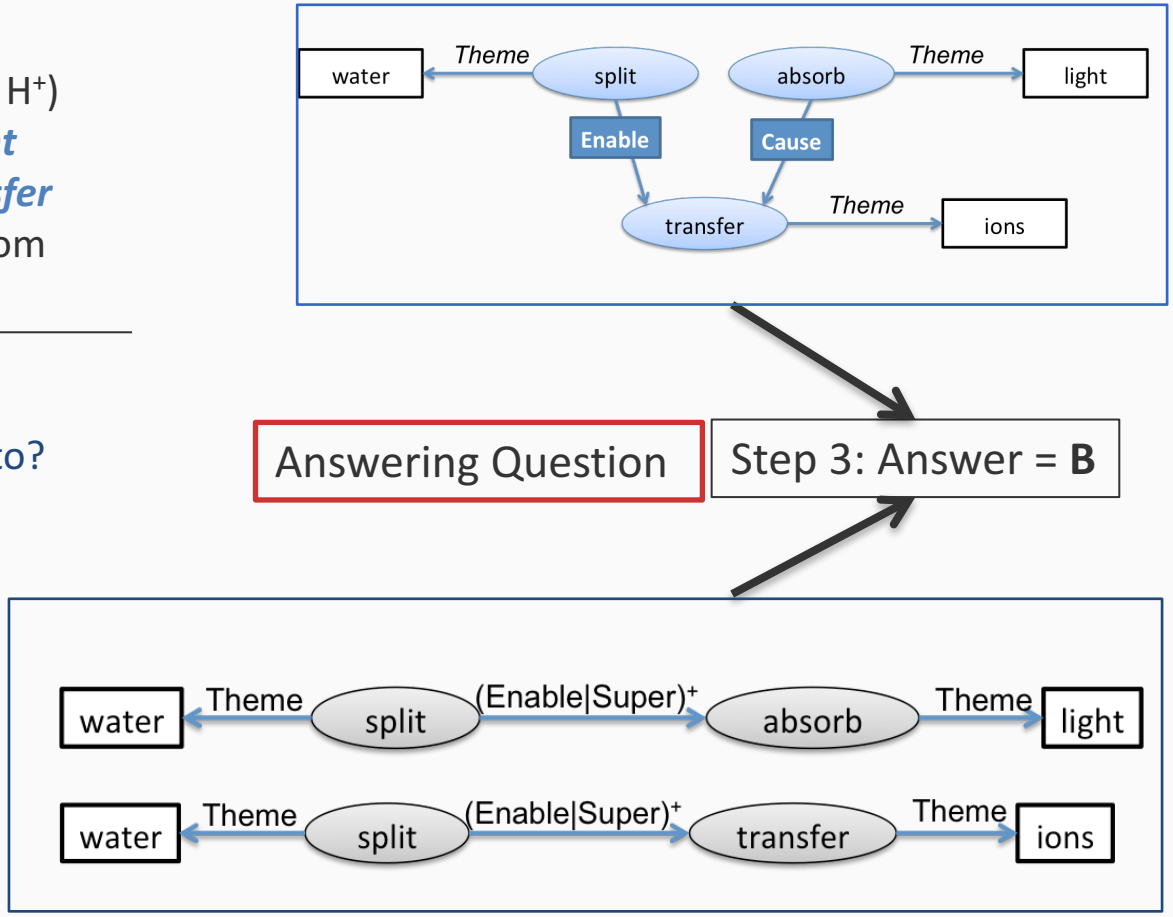
1. Align Q&A triggers and arguments to structure
2. Identify source and target
3. Identify regular expressions
   From small set (~10)



Where are we?

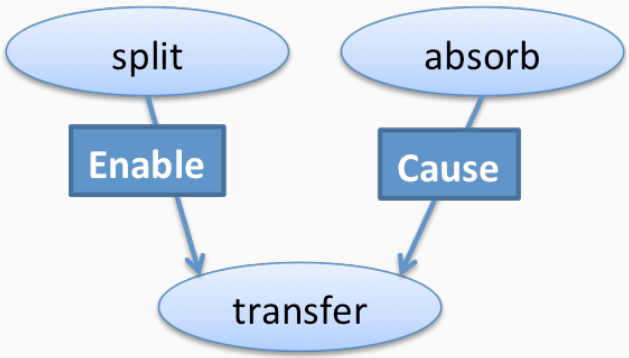
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

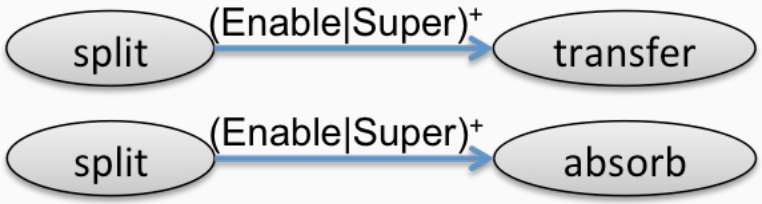
What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions



Step 3: Answering questions

- Given
  - Process structure
  - Two queries

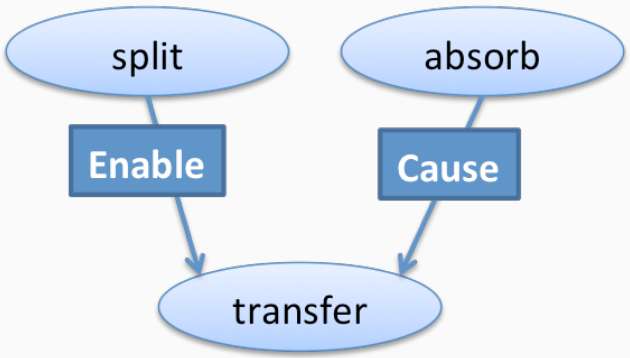


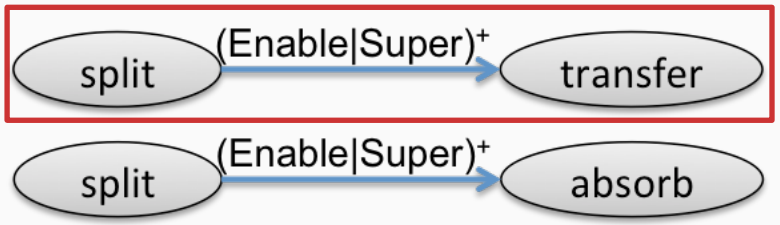


- Answering algorithm

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

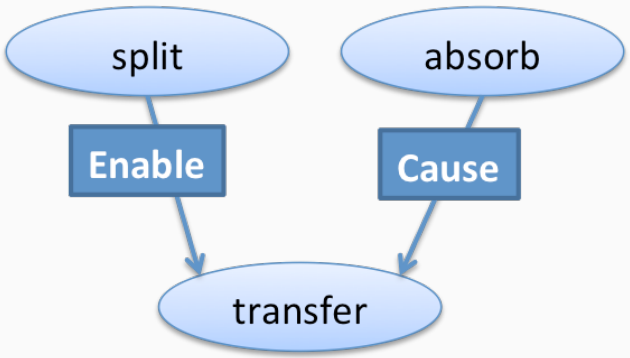


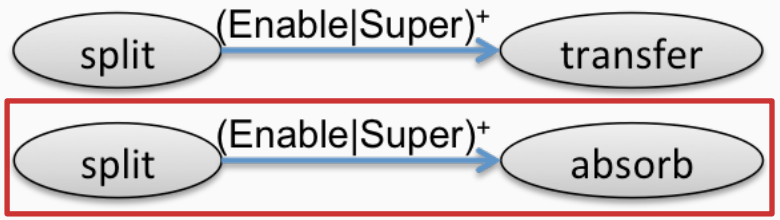


- Answering algorithm
  1. Find matching path (valid proof)

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

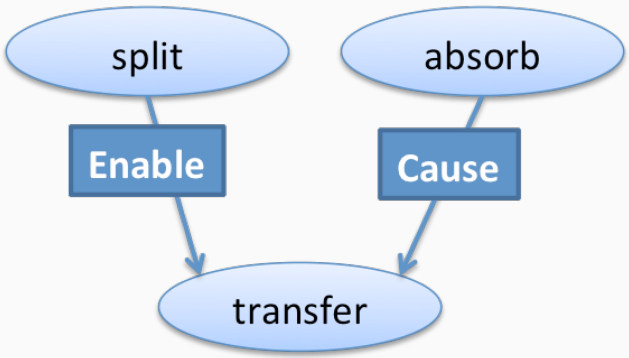


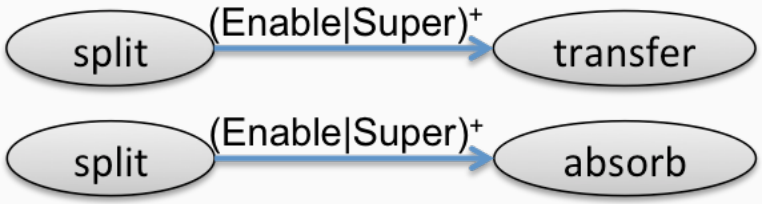


- Answering algorithm
  1. Find matching path (valid proof)
  2. Else, find contradiction of causality (refutation)

Step 3: Answering questions

- Given
  - Process structure
  - Two queries

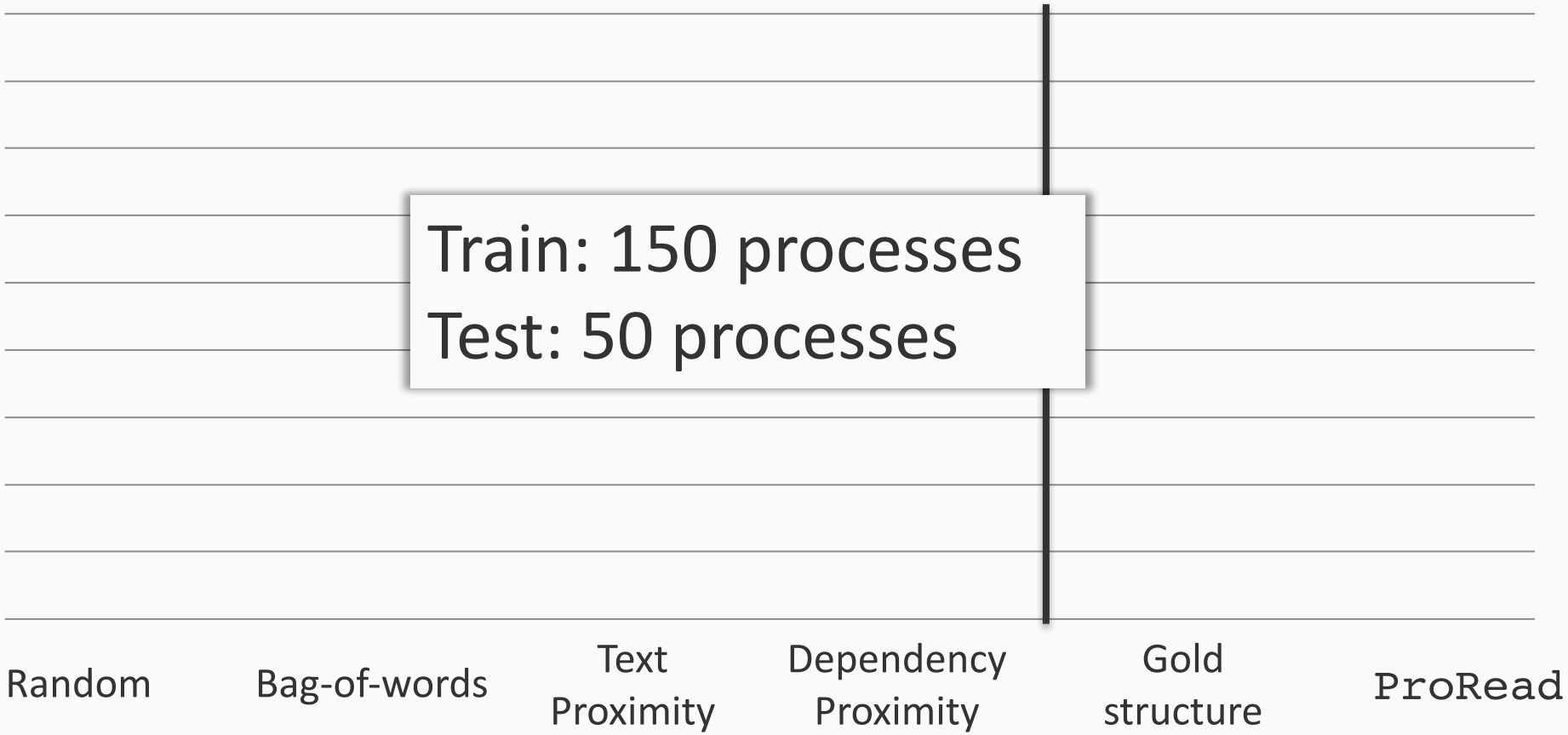




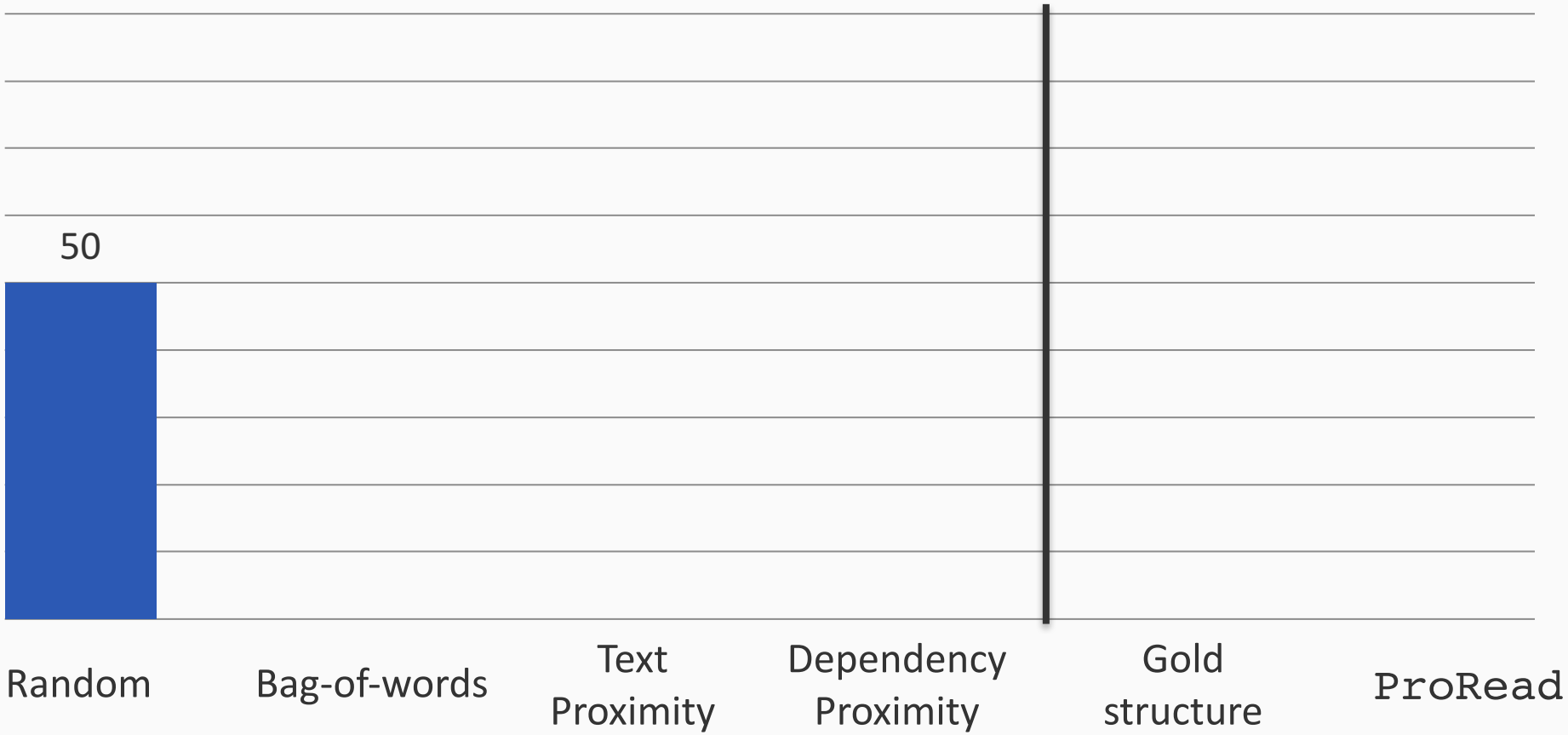
- Answering algorithm
  1. Find matching path (valid proof)
  2. Else, find contradiction of causality (refutation)
  3. Back off to baseline

Experiments and Results

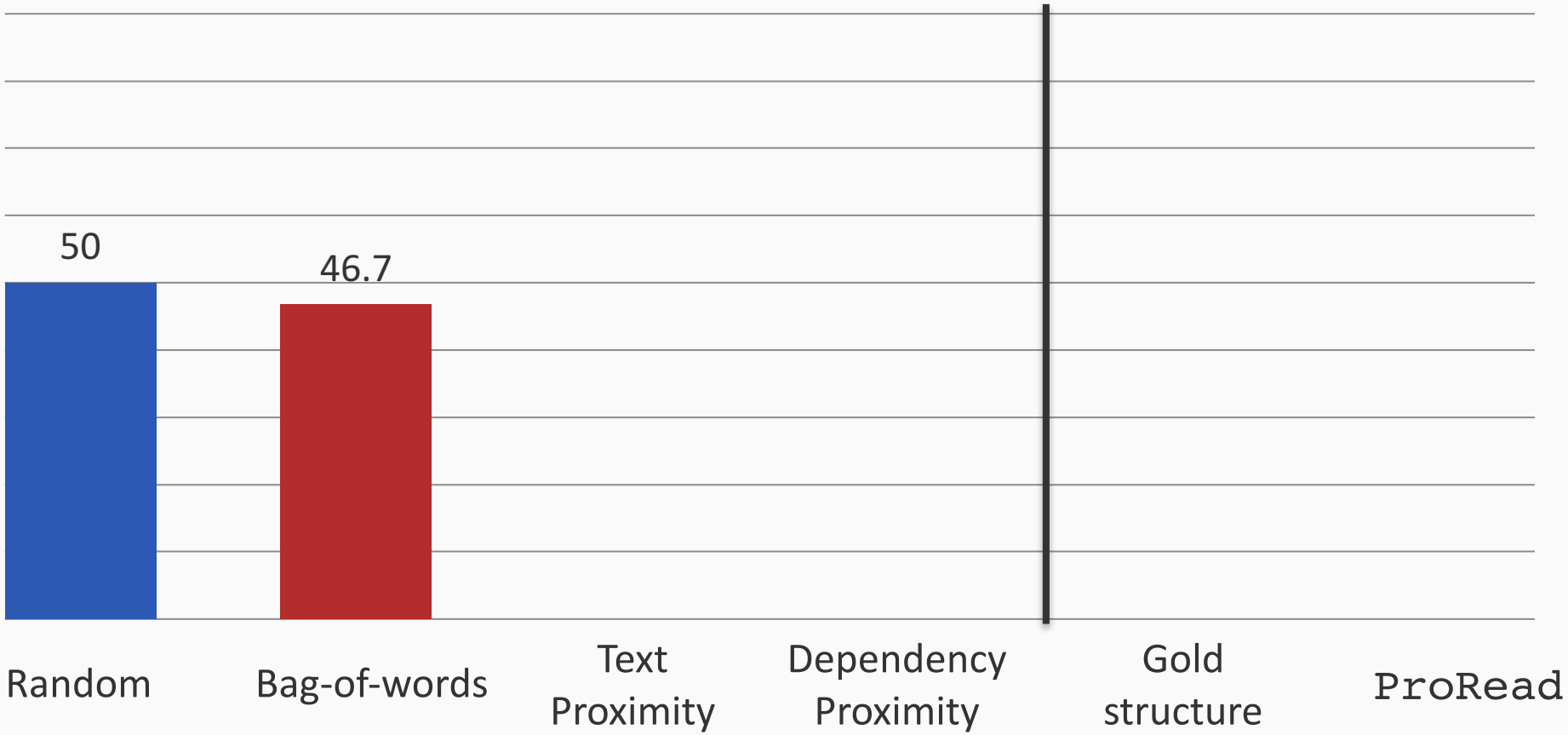
Question Answering Accuracy



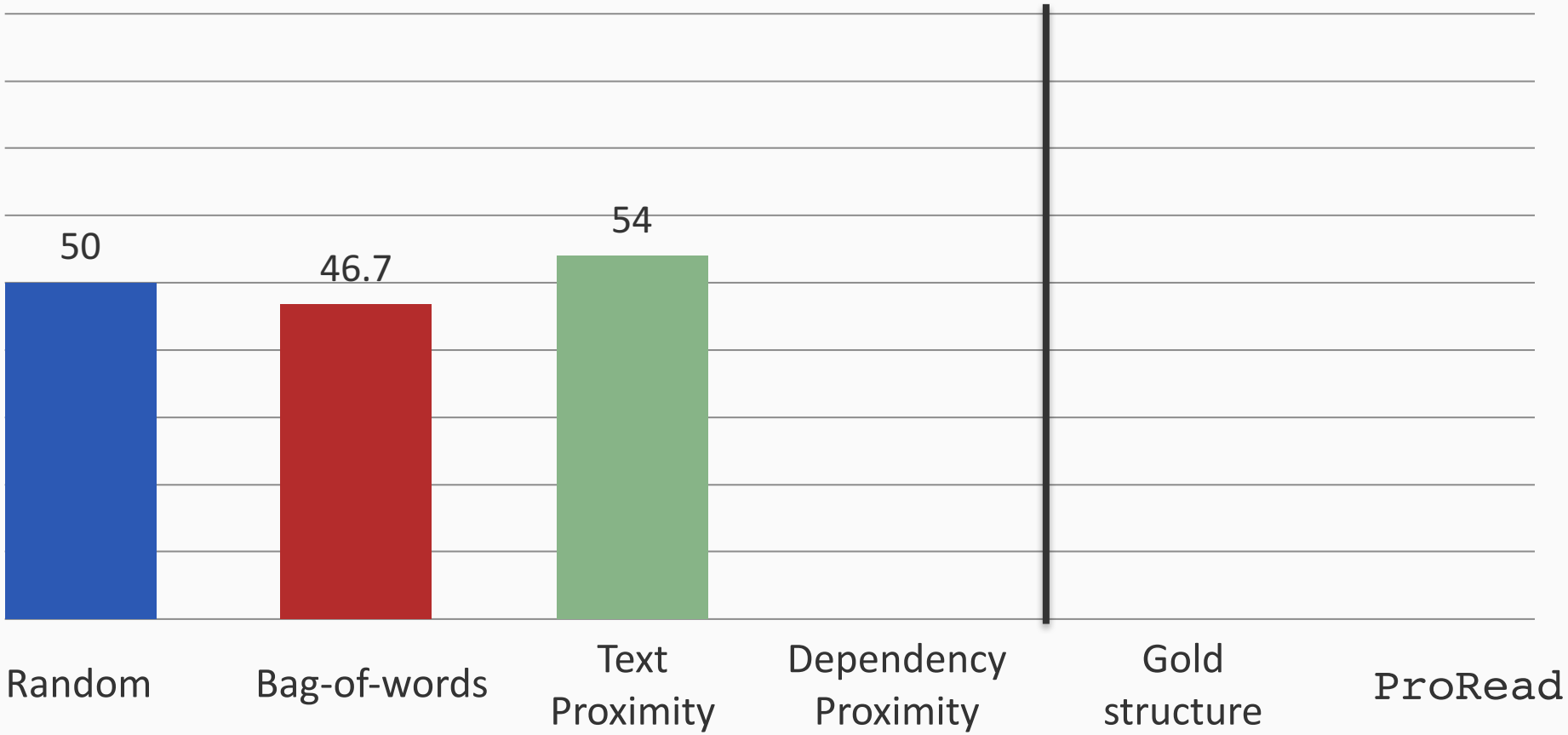
Question Answering Accuracy



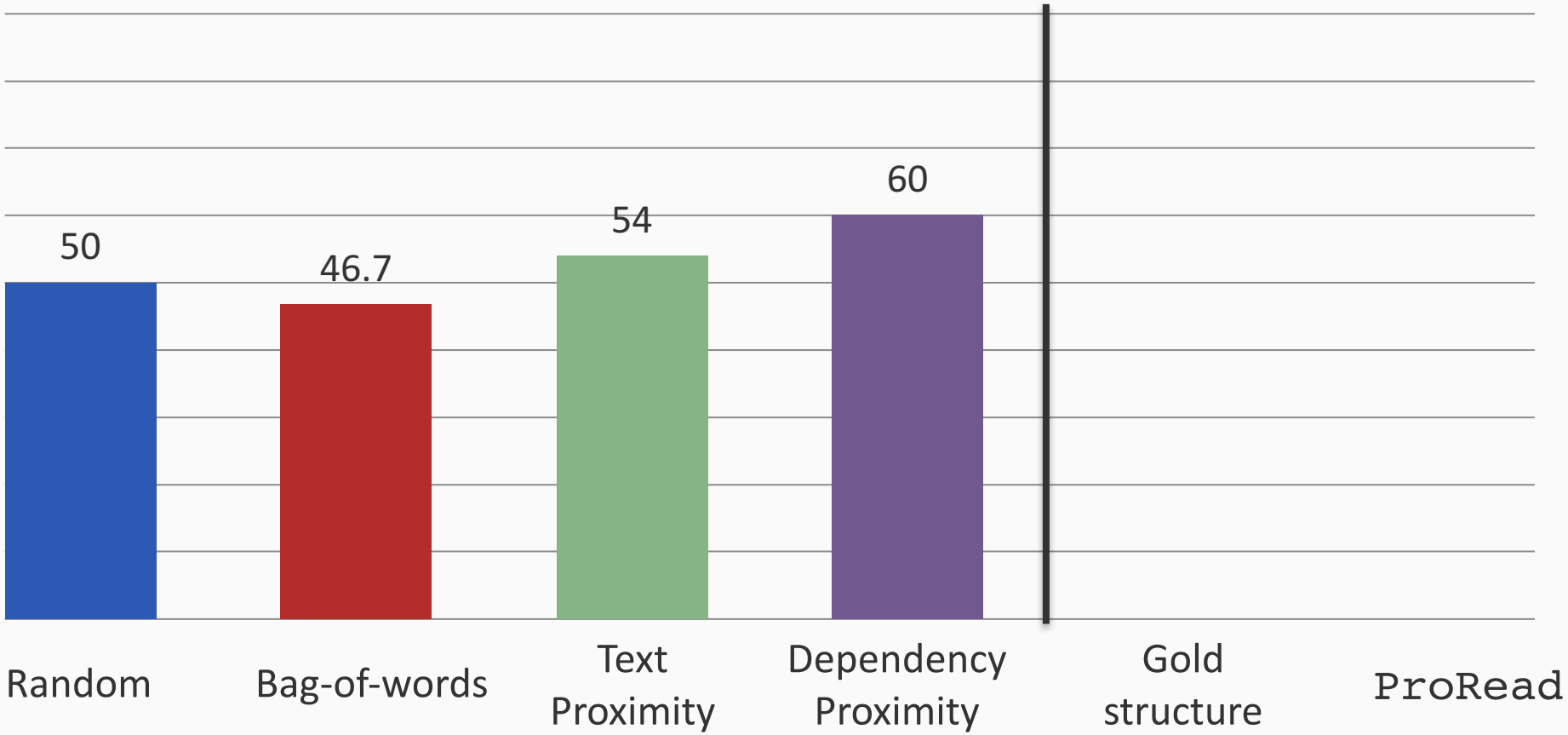
Question Answering Accuracy



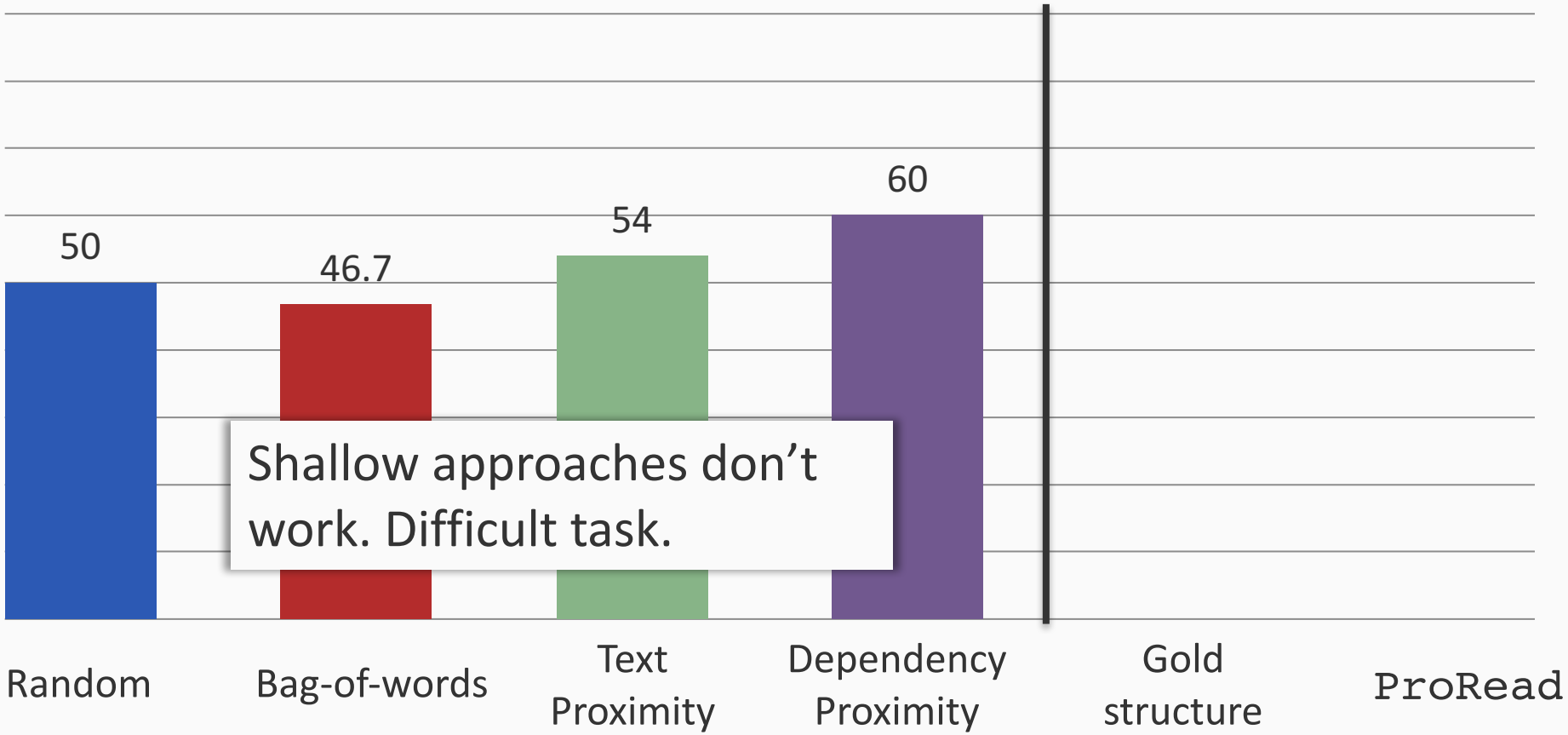
Question Answering Accuracy



Question Answering Accuracy

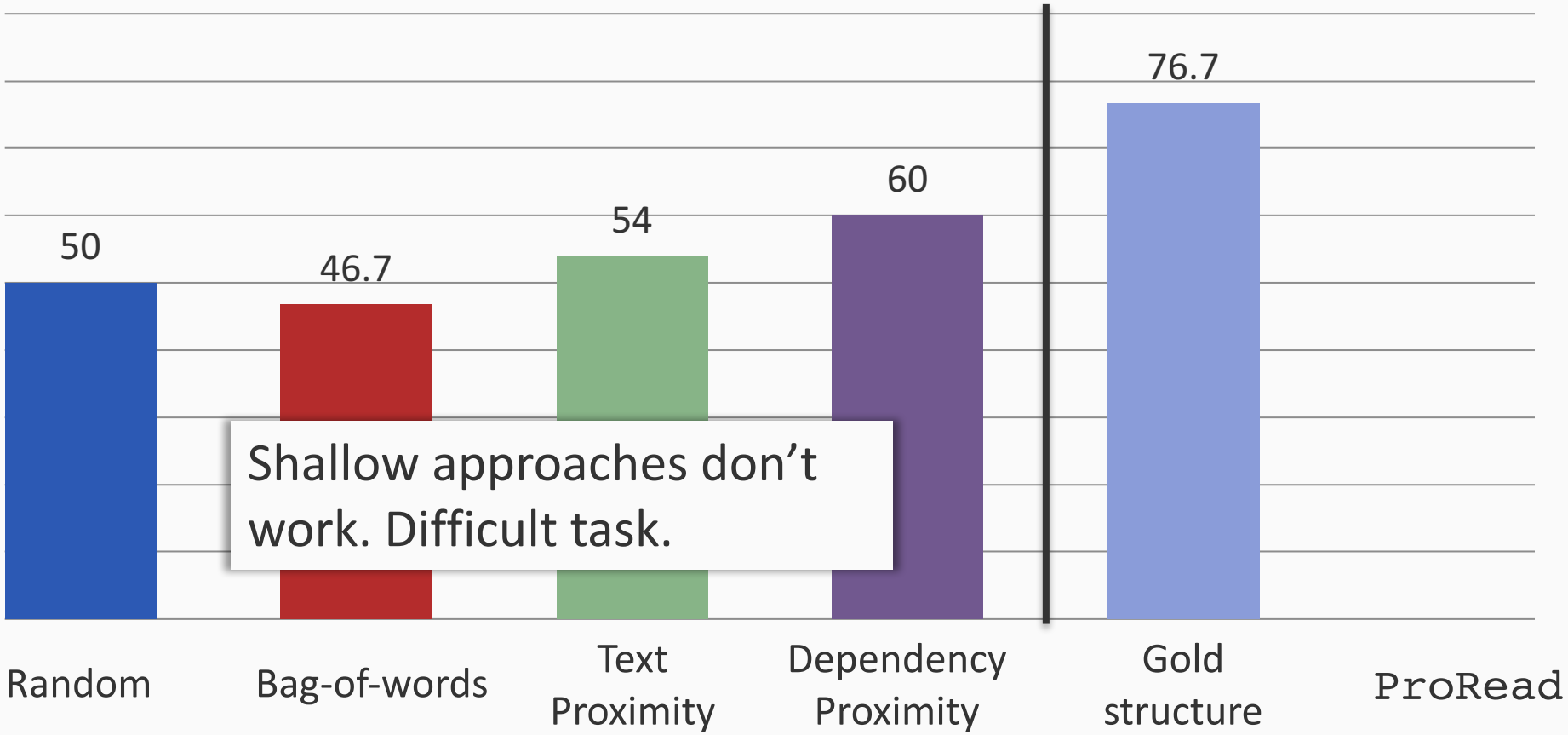


Question Answering Accuracy



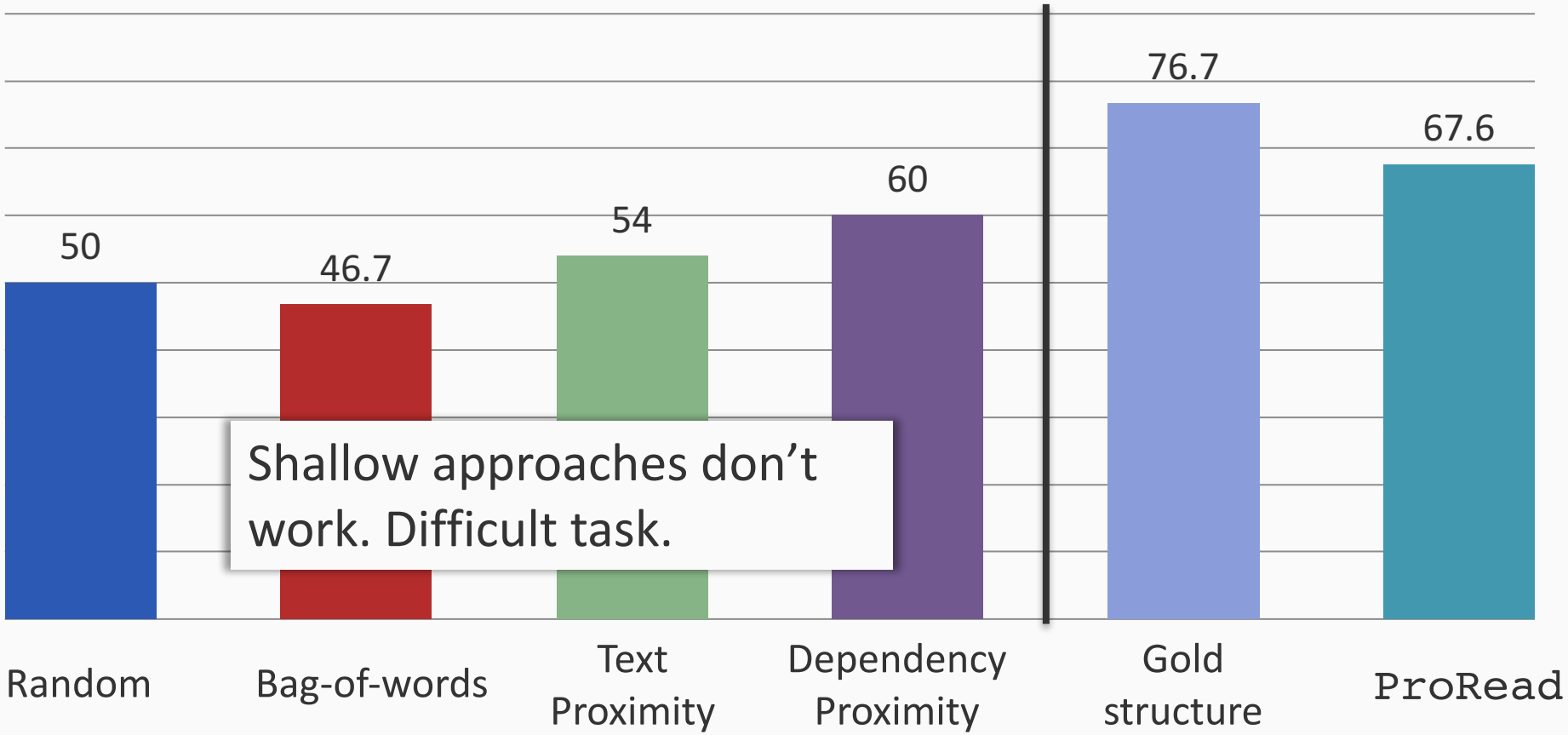
Shallow approaches don't work. Difficult task.

Question Answering Accuracy



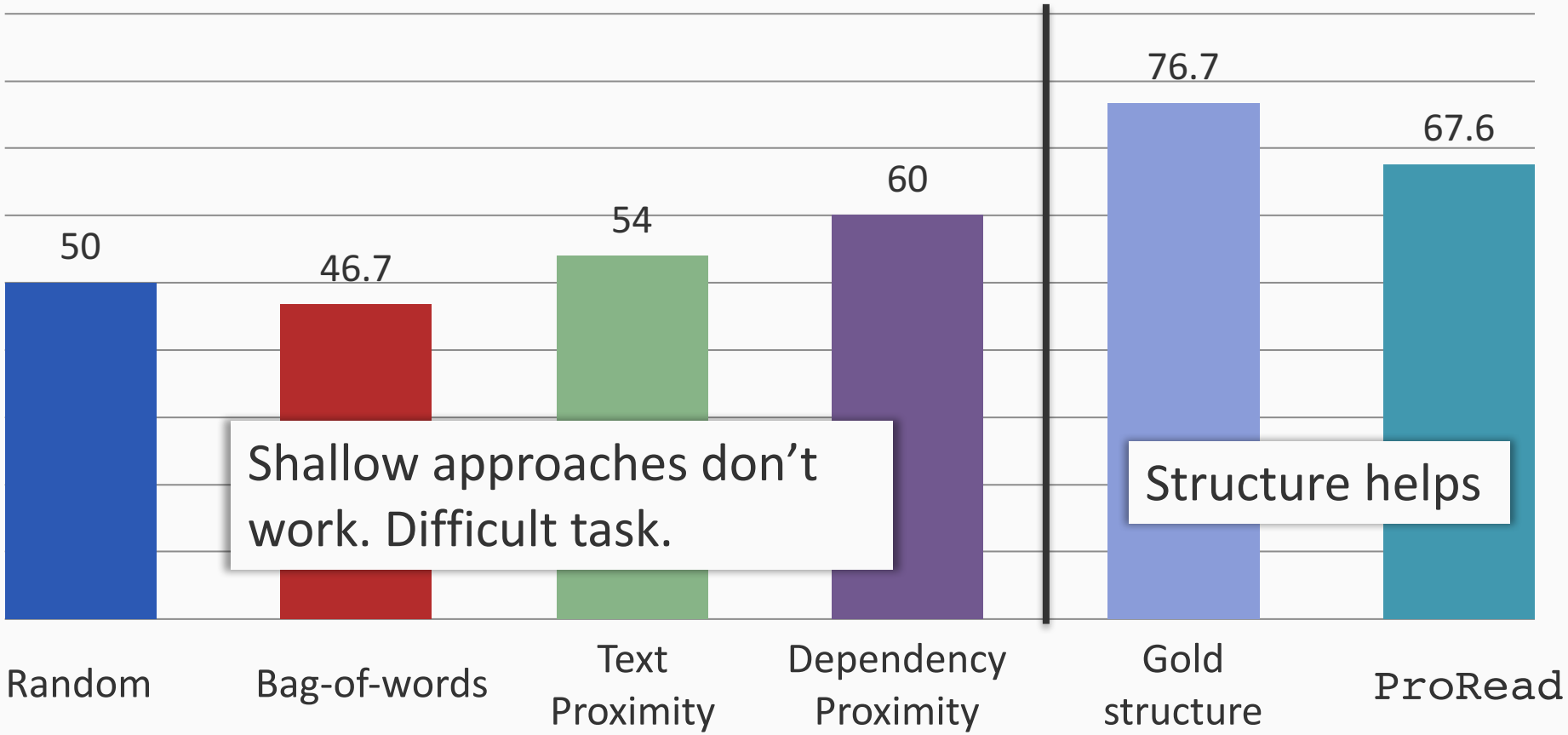
Shallow approaches don't work. Difficult task.

Question Answering Accuracy



Shallow approaches don't work. Difficult task.

Question Answering Accuracy



Summary

- 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

Looking ahead

Structures are useful abstractions

- Other projects that involve linguistic structures
  - [With Tao Li] Structures can explain judgments about semantic similarity or entailment between sentences
  - [With Nathan Schneider, Jena Hwang and Martha Palmer] Semantic relationships triggered by prepositions

Some open research questions

- Experts are expensive
  - Can we use non-experts to provide expert-level annotation and use learning to fill in the gap?
- Deep learning: we can learn good feature representations
  - Can we integrate the representational benefits of deep learning with structured inference?
- How do we make prediction fast?
  - Structured inference can be slow. Can we learn to make faster predictions?
- What is the right representation for language?
  - Is there a “right representation”?

Broader research concerns

Questions?

Various interconnected threads

- The NLP question: How to represent the semantics of text?
  - What is a good representation?
    - A graph? Or something in a real valued?
- The machine learning question: How to learn to predict this representation?
  - Learning and inference algorithms
  - Using world knowledge
  - Typically techniques transfer to non-NLP domains too!
- The AI question: Can a program reason about the state of the world?