Natural Language Processing in the Wild: Opportunities & Challenges

Vivek Srikumar

THE UNIVERSITY OF UTAH

This talk includes collaborative work with many people, including Zac Imel, Michael Tanana, and Nathan Schneider and the adposition gang.

The Wright Flyer

The pinnacle of aviation in 1905

We have traveled far in 100 years

Freud’s therapy room

The pinnacle of mental health in 1905

How far have we traveled in 100 years?

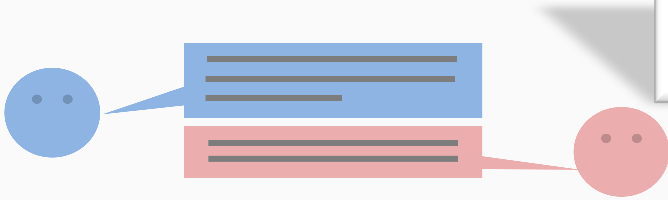


Photo: Jeff Hitchcock - Own work, CC BY 2.0

This talk

Natural language processing can revolutionize many fields

…but we need to overcome several challenges.

- Empathy and the Machine: A case study of NLP and Mental Health

- Challenge 1: The importance of representations

- Challenge 2: Scaling NLP for everyone

Michael Tanana

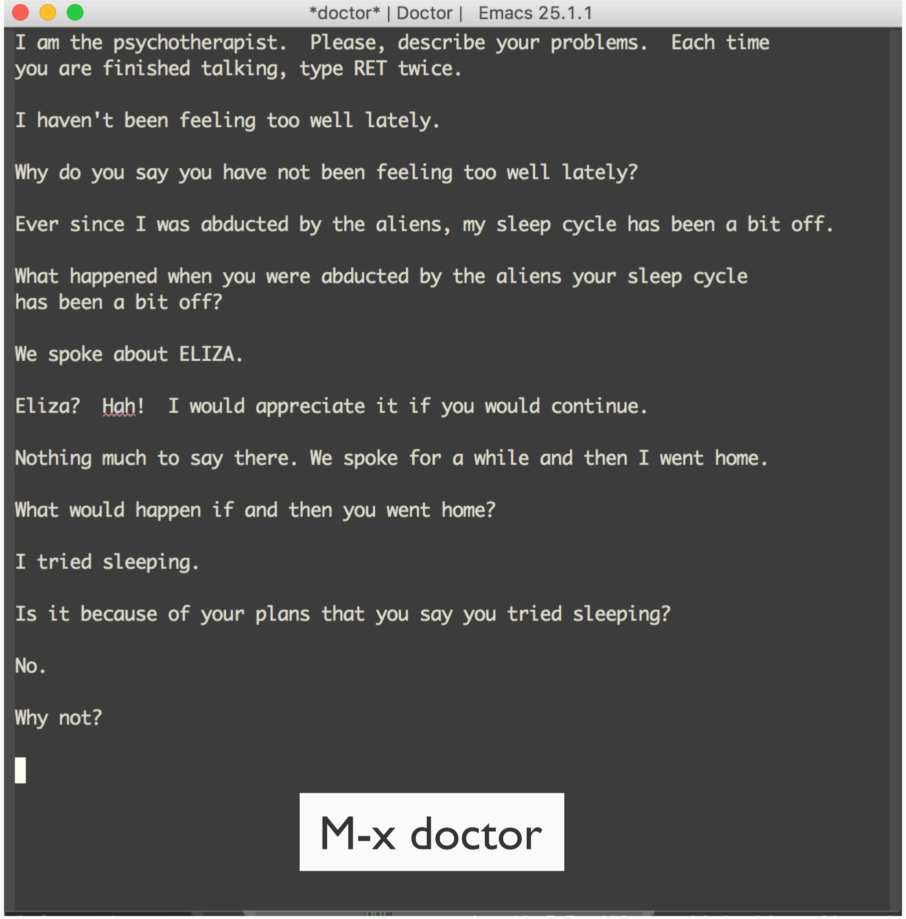
Zac Imel

Empathy and the Machine

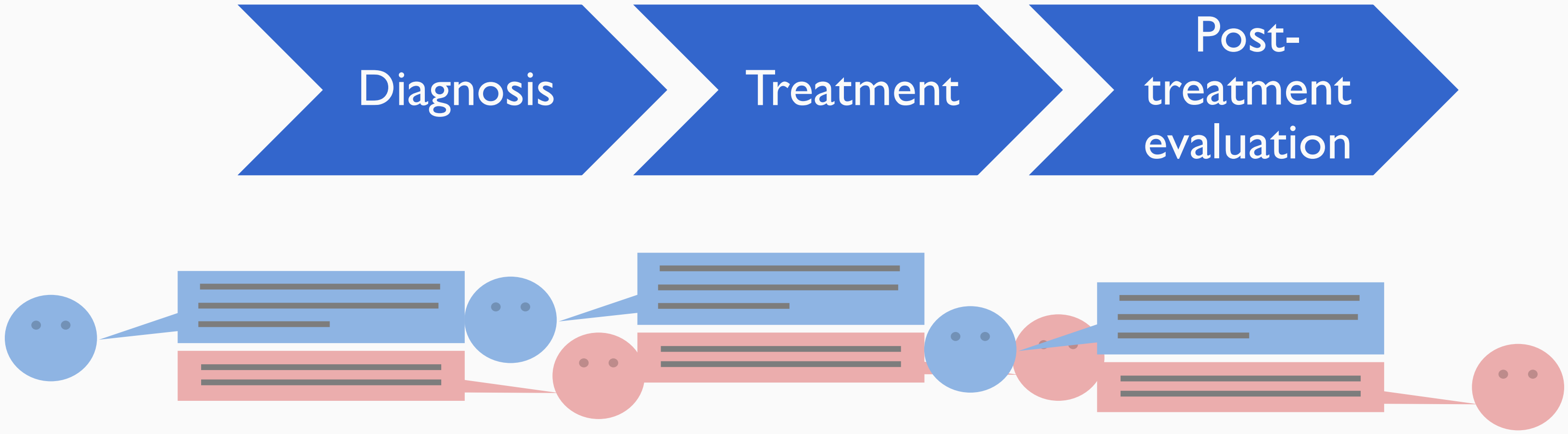
How can NLP help mental health?

A long history of therapy bots in NLP

ELIZA



A trip to a therapist…



Mental health therapy

Essentially a dialogue between two people

- Some treatments better than others?
- Some counselors better than others?
- Can we measure patient improvements?
- How does therapy lead to changed behavior?

A collection of NLP problems

Making sense of the hours of unstructured, often emotional dialogue

- Analyze, improve and study the effect of psychotherapy

Several opportunities Joint work across Utah, UW, UCI,…

– Assessing therapy sessions [Tanana et al 2016]
– Tracking sentiment [Tanana et al 2015]
– Helping counselors *during* therapy
– Helping train better therapists
– Many more…

A therapy session transcript

Counselor: How do you feel about your progress so far?

Patient: Everyone's getting on me about my drinking

Cou

Is the therapist being helpful?

Patie

Counselor: You're not sure you can finish treatment.

Patient: I drank a couple of times this week when I was with my brother.
I want to quit so badly,
but I don't think I can do it.

Is a therapist being helpful?

Therapist can make a `reflection` what the patient said, ask an `open question`, etc.

Patient can talk about `sustaining` or `changing` their behavior or say something `neutral`.

These labels form the basis of a set called the MISC (Motivational Interviewing Skill Codes). [Houck et al 2012]

MISC coded session

[Houck et al 2012]

Counselor: How do you feel about your progress so far? `[Open Question]`

Patient: Everyone's getting on me about my drinking `[Follow-Neutral]`

Counselor: Kind of like a bunch of crows pecking at you? `[Complex Reflection]`

Patient: I'm not sure I can finish treatment. `[Sustain talk]`

Counselor: You're not sure you can finish treatment. `[Simple Reflection]`

Patient: I drank a couple of times this week when I was with my brother. `[Sustain Talk]`
I want to quit so badly, `[Change Talk]`
but I don't think I can do it. `[Sustain Talk]`

How therapists get feedback today

1. Patient and doctor talk to each other
2. Transcribe the conversation
3. Label every utterance of the transcript as being in line with the counseling style
4. Generate aggregate statistics from the labels
5. Give feedback to the therapist

Does not scale

Can we automatically label utterances?

[Atkins, et al 2014, ..., Tanana et al 2016, JSAT]

The hope

To provide automatic feedback to therapists after every session

To train better therapists

To conduct aggregate studies about the state of mental health treatment

The therapy dataset

- 341 therapy sessions
  - approximately 1.7 million words,
  - 175,000 utterances

- Focus on substance abuse
  - Affects 21 million Americans (as of 2014)

We trained two sequence models

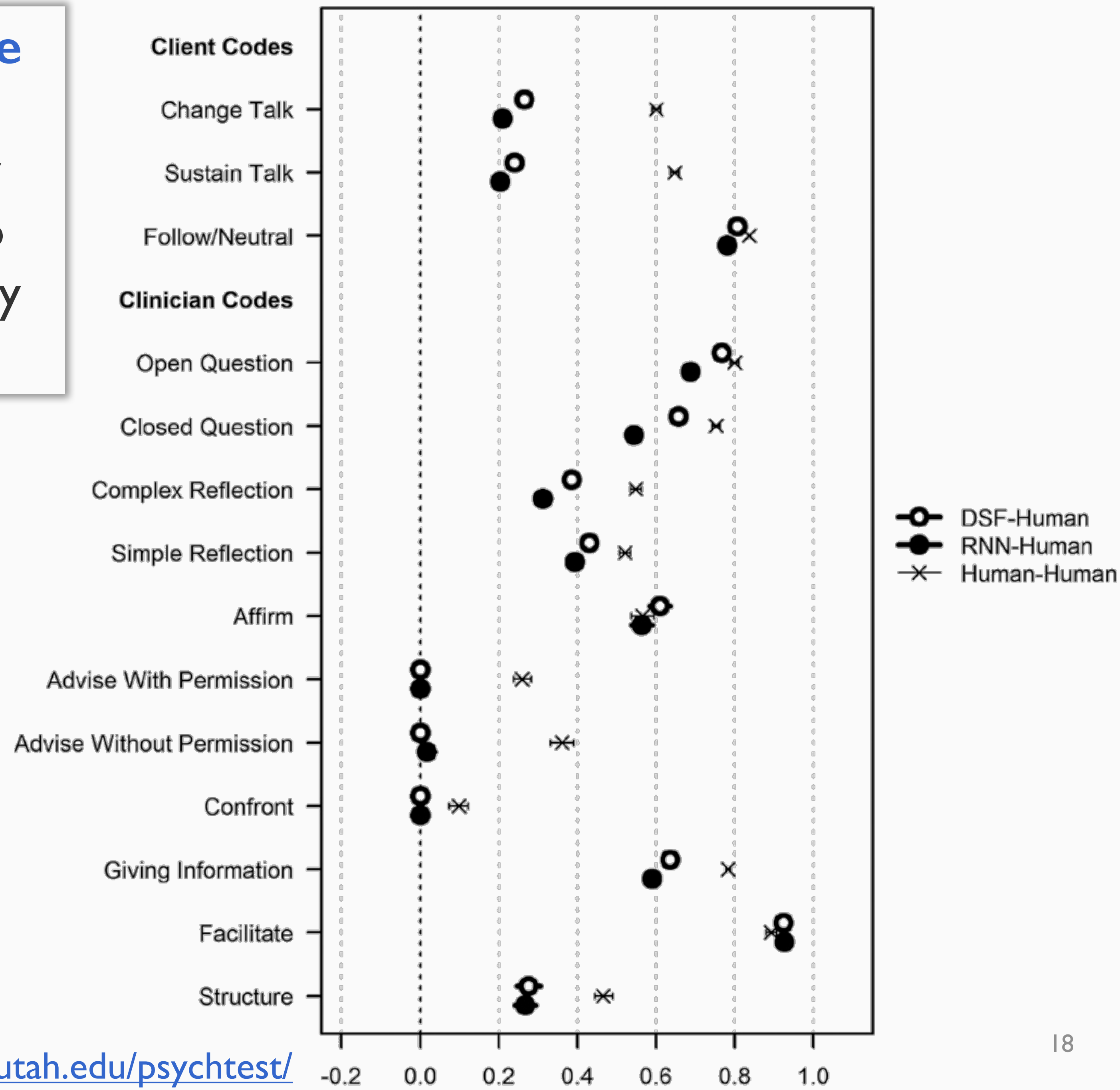
Both models label utterances in the session with one of the MISC labels

- Differ in their feature representation of each utterance

- Model 1:
  - A traditional feature based model
  - Based on words and dependency features
- Model 2:
  - A recursive neural network for each sentence
  - Operates on top of the dependency tree

The punchline

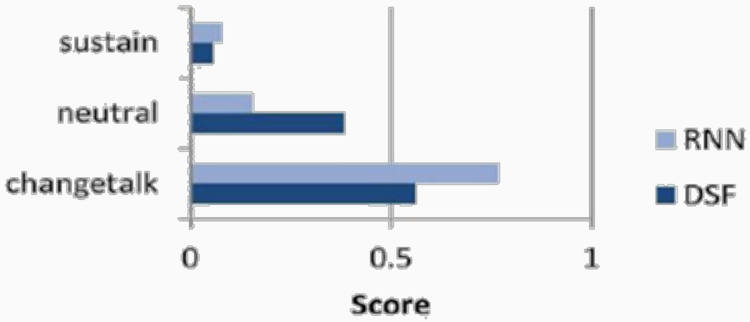
Trained models predicted many labels similar to human reliability on the test set.



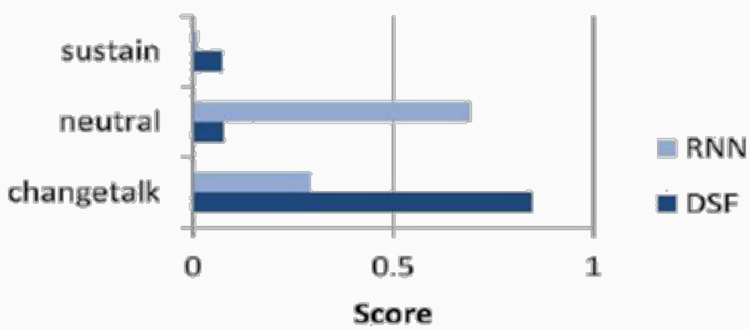
Live demo: http://sri.utah.edu/psychtest/

Client

(a) "and just sometime i just get tired of being high so" (change talk)

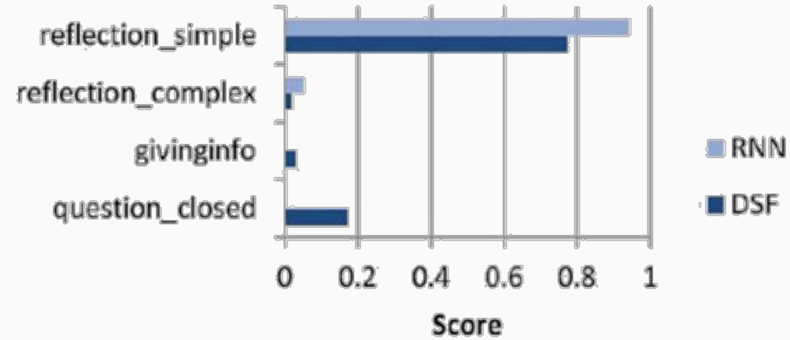


(b) "but something happened in that motel where i just said i just can't do this anymore" (change talk)

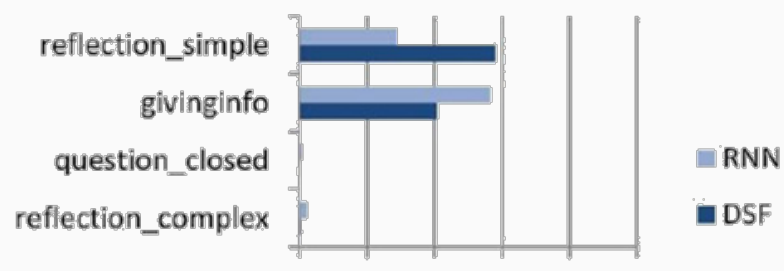


Clinician

(c) "you said that you graduated and that you did accomplish some of your goals it sounds like" (reflection: simple)



(d) "it seems like you were drinking seven days a week about four drinks at a time" (giving info)



Wanted

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

Natural Language Processing can transform many fields. But…

Still a long way to go

Challenges abound

Challenges

Representations

What is a **good** representation of language?

Eg: How can we best represent words, sentences and utterances to reason about complex text like the therapy transcripts?

Data

How do we get around the need for annotated data?

Annotated data is a precious resource

Efficiency

How can we scale predictive models (eg: to every doctor and patient in the world)?

How can we avoid unnecessary computation?

Challenges

Representations

What is a **good** representation of language?

Eg: How can we best represent words, sentences, and utterances to reason about complex text like the therapy transcripts?

Data

How do we get … data?

Annotated d…

This talk

Efficiency

How can we scale predictive models (eg: to every doctor and patient in the world)?

How can we avoid unnecessary computation?

Jonathan Berant

Chris Manning

What is a good semantic representation?

Modeling Biological Processes for Reading Comprehension

EMNLP 2014

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 $NADP^+$.

Cause

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 accept

Does such a representation help reading comprehension?

B: Transfer of ions

A difficult reading comprehension task

200 paragraphs from the textbook *Biology*

[Campbell & Reese, 2005]

Desiderata

1. Test understanding of inter-relations between events and entities
2. Answers should have similar lexical overlap
   - Shallow approaches will fail

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

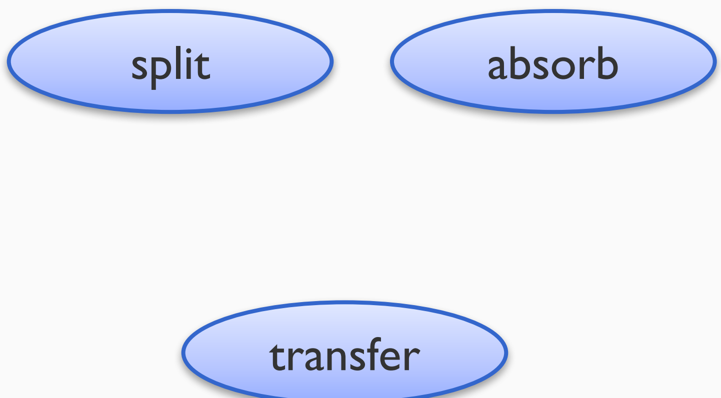
B: False

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

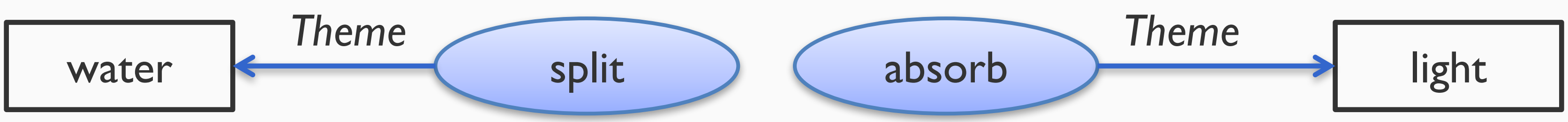
Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.



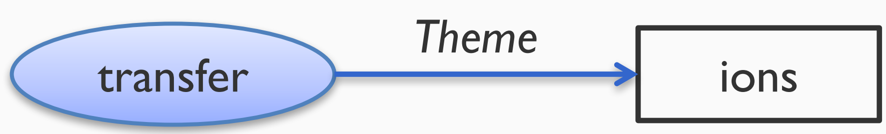
Triggers: Tokens denoting occurrence of an event

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 $NADP^+$.



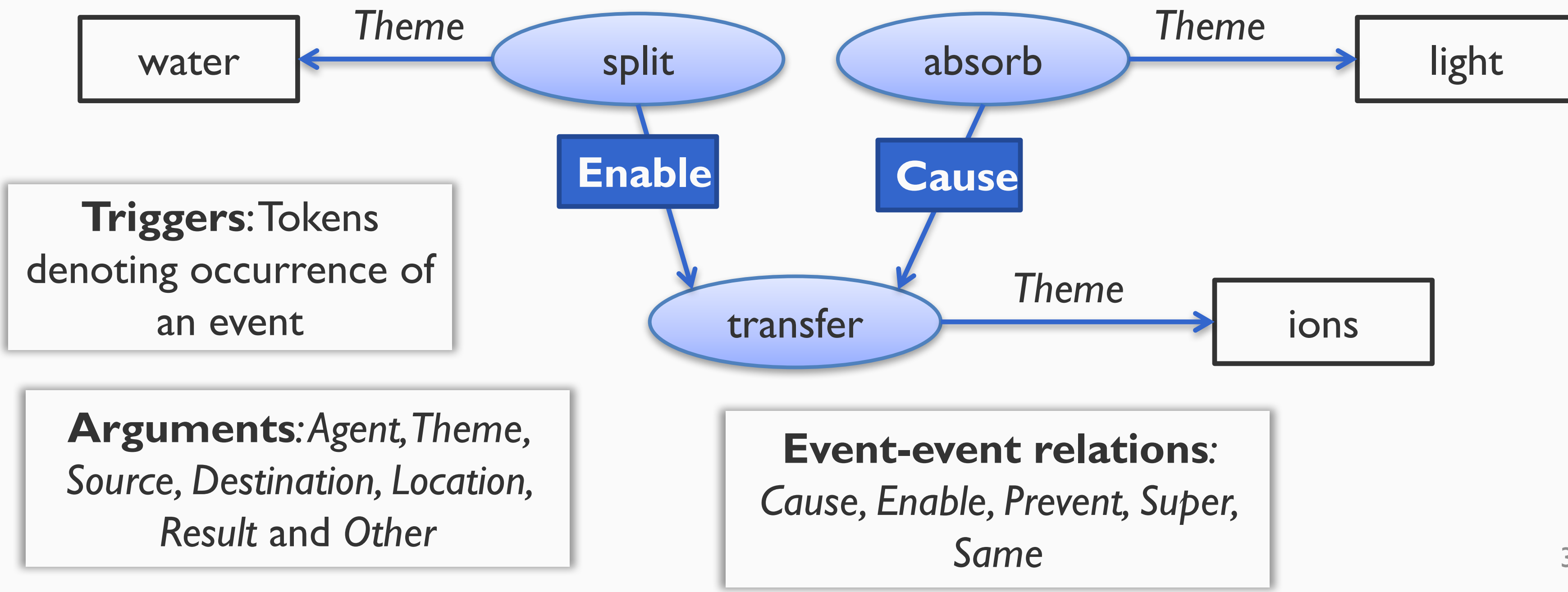
Triggers: Tokens denoting occurrence of an event



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 $NADP^+$.



ProcessBank

- 200 paragraphs from the textbook *Biology*
  - Manually chosen to represent biological processes

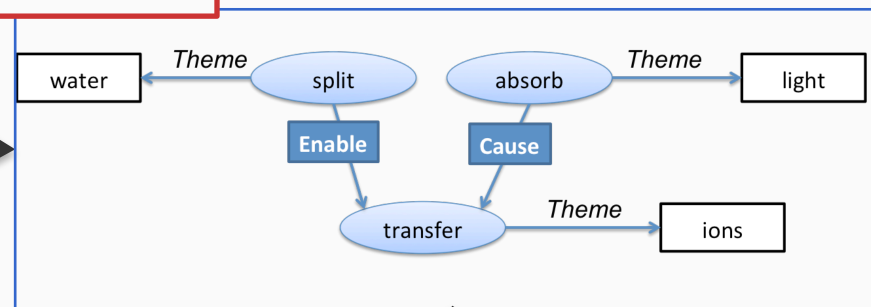
- Each paragraph annotated with
  1. Non-factoid reading comprehension questions
  2. Process structures

Answering questions: Overview

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

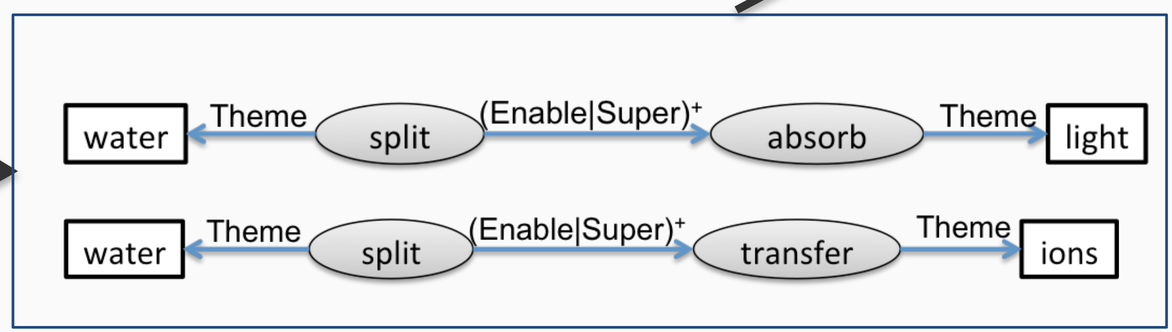
B: Transfer of ions

Answering Question

Step 3: Answer = **B**

Step 2

Question Parsing

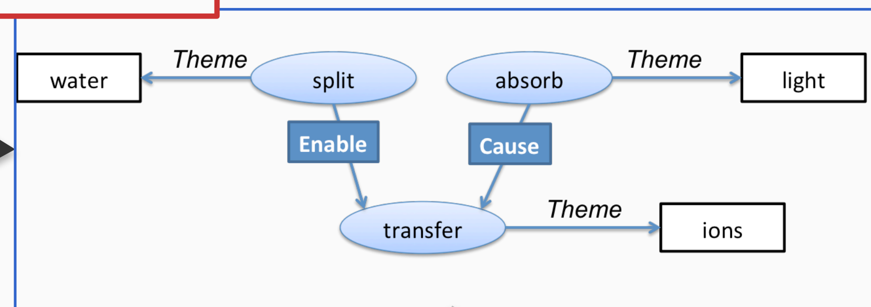


Answering questions: Overview

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



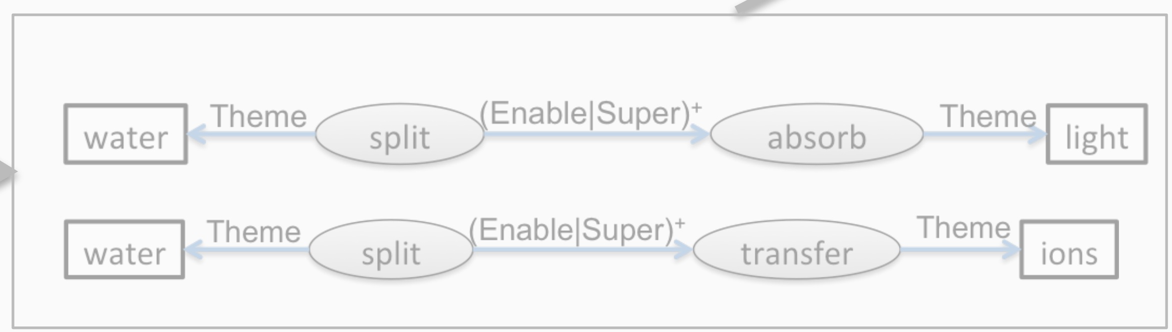
What can the splitting of water lead to?
A: Light absorption
B: Transfer of ions

Answering Question

Step 3: Answer = **B**

Step 2

Question Parsing



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

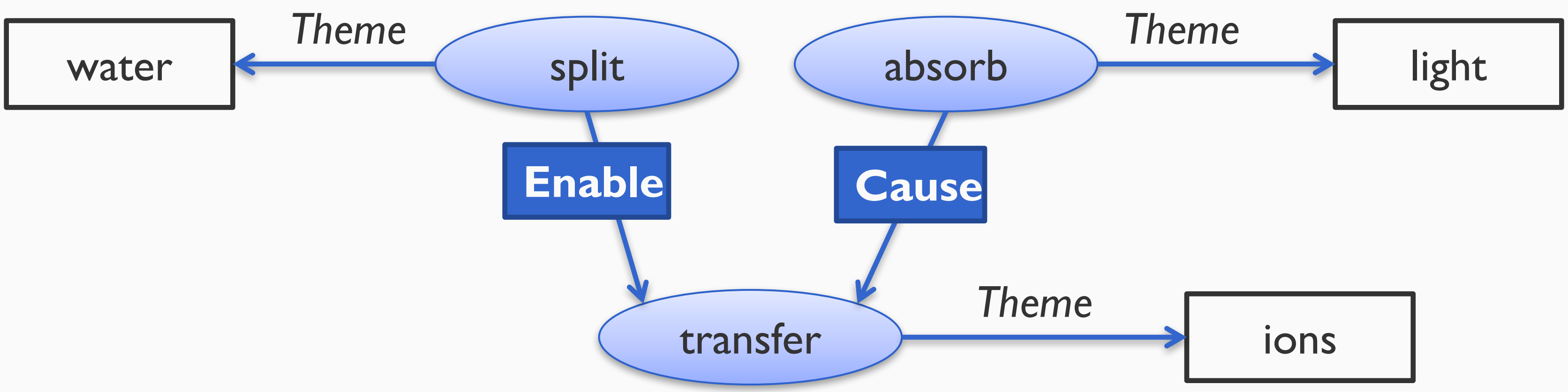
1. Train event trigger identifier

   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]

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

Learning and Inference

- Linear model to score argument labels and event-event relations
  - Related: Semantic role labeling, information extraction

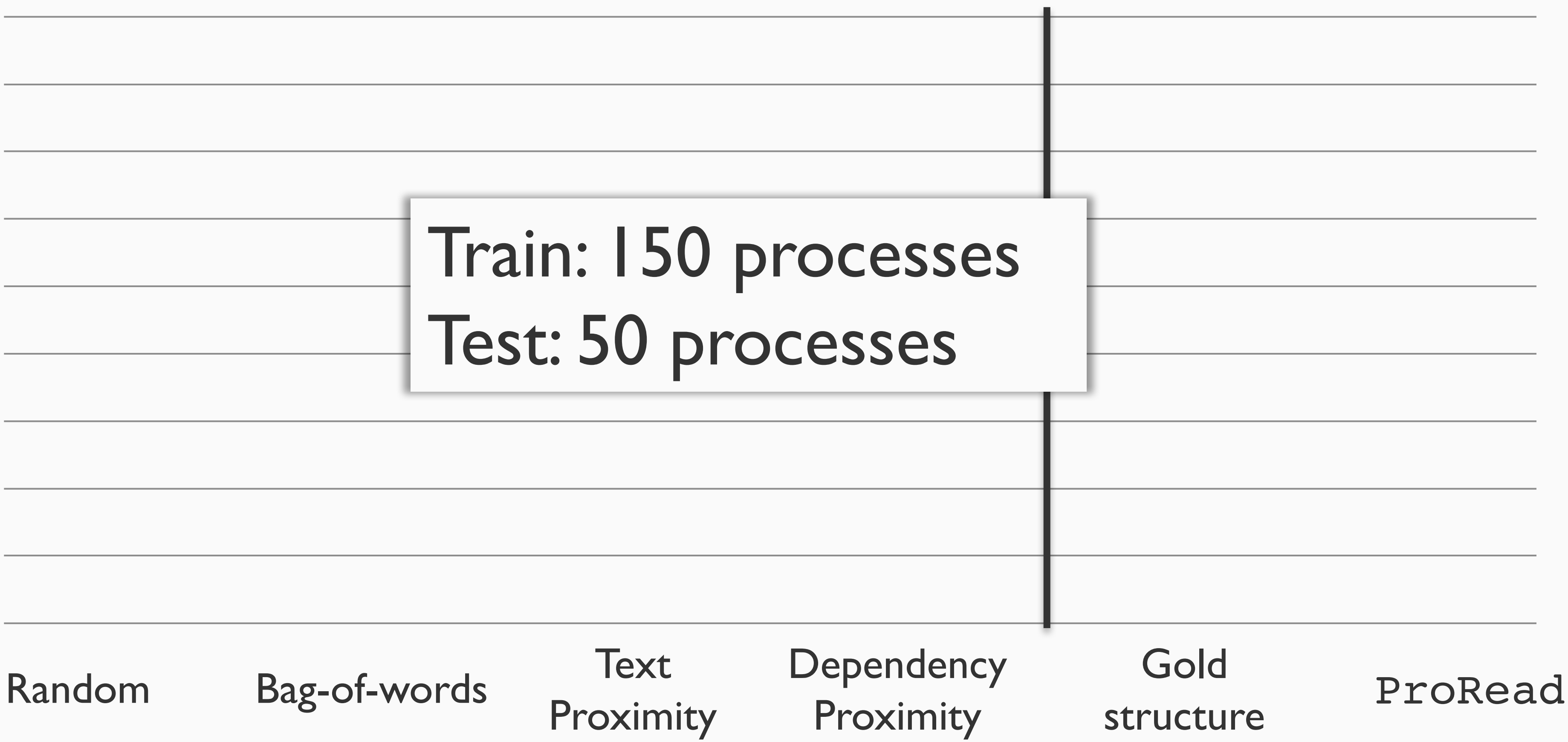
- Structured averaged perceptron

- Integer linear program solver for inference

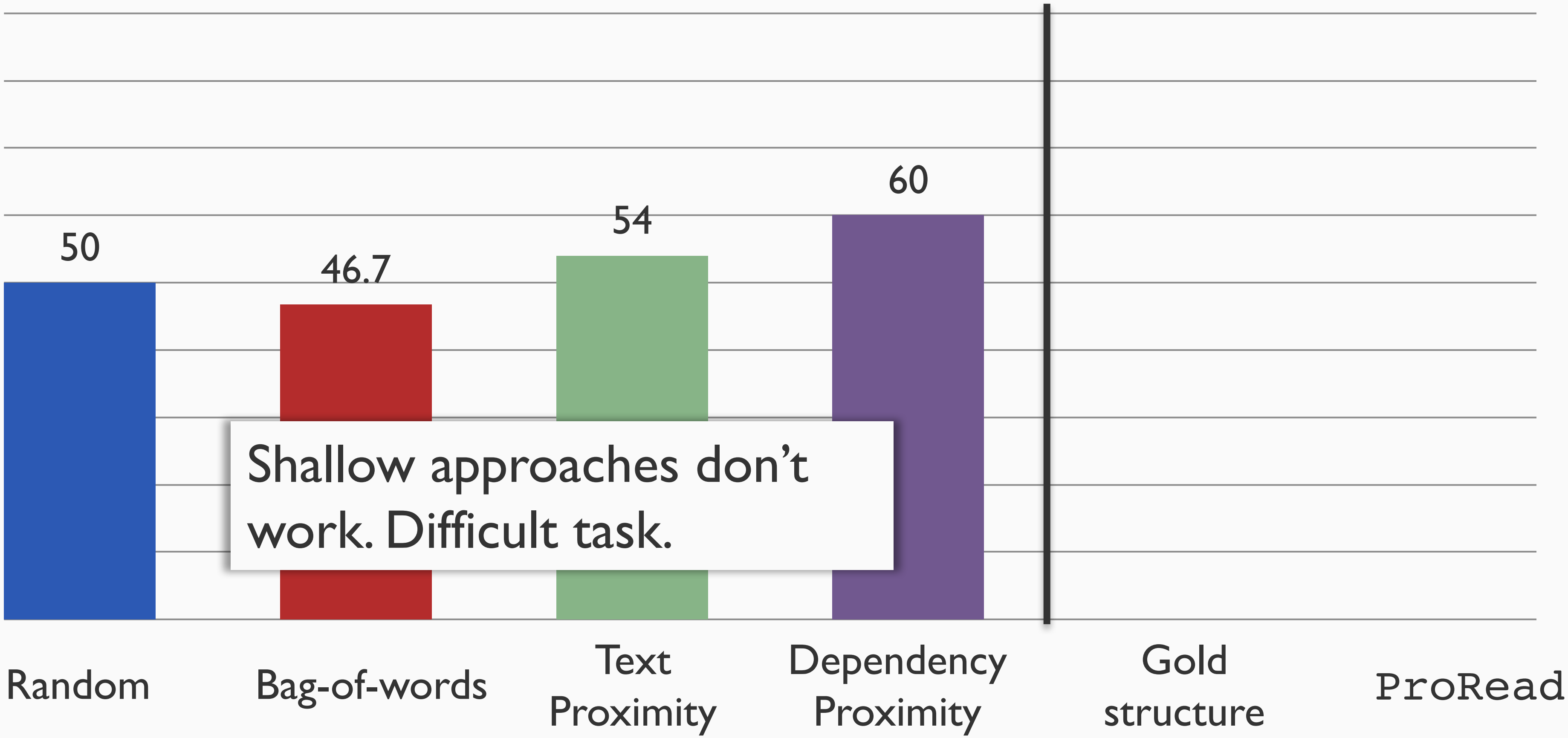
Question Answering Accuracy

Train: 150 processes

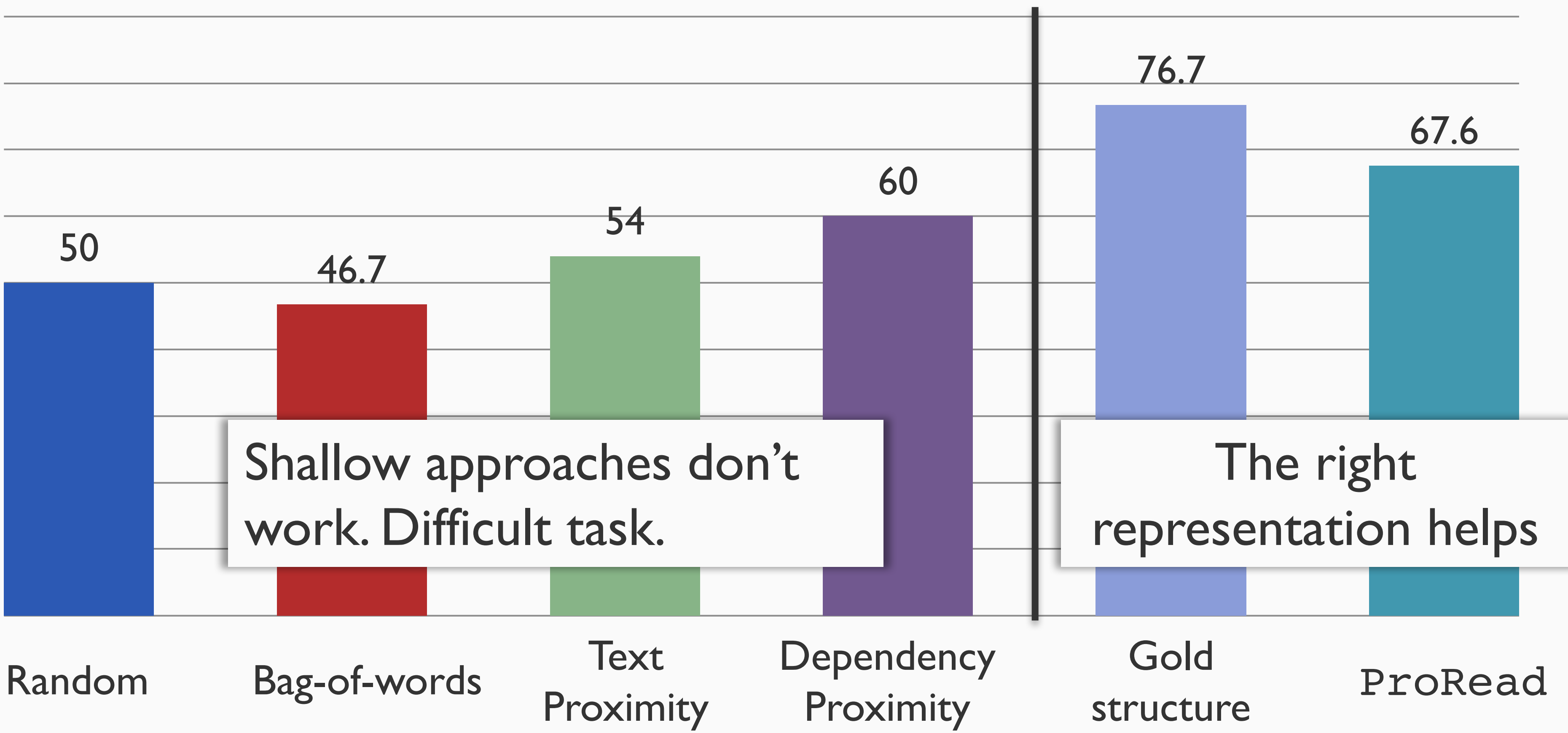
Test: 50 processes



Question Answering Accuracy

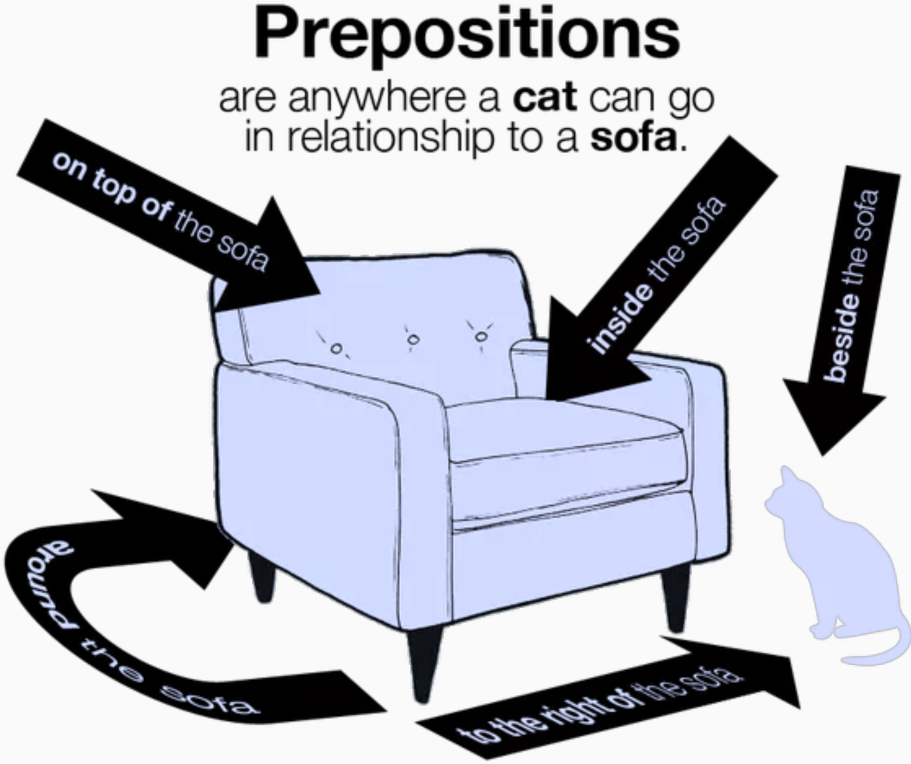


Question Answering Accuracy



Key lessons

- Linguistic intuitions can help develop useful *structured* representations for performing reasoning about text
- But, both designing and predicting such representations can be difficult



sofa image via The Guardian (http://goo.gl/SgEvJf)

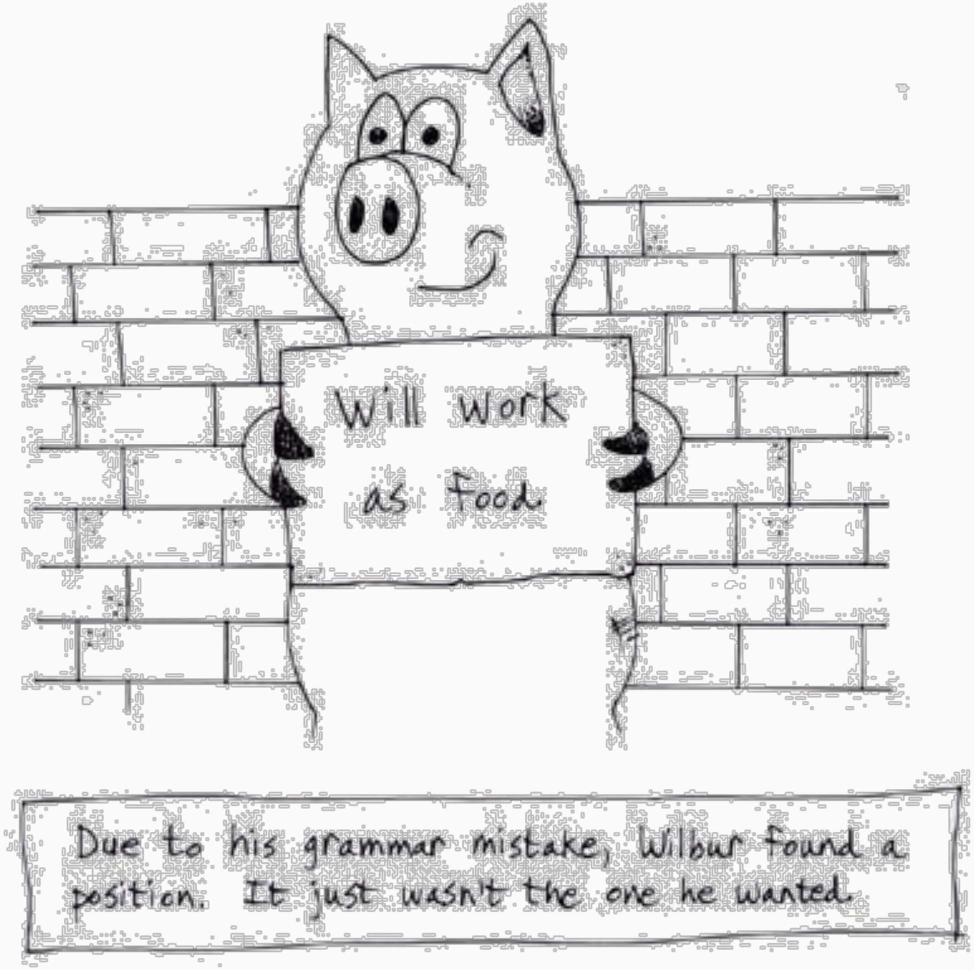
Nathan Schneider

Jena Hwang

What is the right representation for disambiguating prepositions?

Joint work with
Nathan Schneider, Jena Hwang, Archna Bhatia, Na-Rae Han, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, Austin Blodgett, Jakob Prange, Omri Abend, Sarah Moeller, Aviram Stern, Adi Bitan, Dan Roth, Martha Palmer

Prepositions trigger semantic relations



[Srikumar & Roth 2013],
[Schneider et al 2015, 2016, 2018],
[Hwang et al 2017]

Preposition Sense Disambiguation

Merriam-Webster **over**

Fine grained labels
Like `Over:15(6)-1`

Full Definition of OVER

1 —used as a function word to indicate motion or situation in a position higher than or above another <towered *over* his mother> <flew *over* the lake> <rode *over* the old Roman road>

2 **a** —used as a function word to indicate the possession of authority, power, or jurisdiction in regard to some thing or person <respected those *over* him>

b —used as a function word to indicate superiority, advantage, or preference <a big lead *over* the others>

c —used as a function word to indicate one that is overcome, circumvented, or disregarded <passed *over* the governor's veto>

3 **a :** more than <cost *over* $5>

b : ABOVE 4

4 **a** —used as a function word to indicate position upon or movement down upon <laid a blanket *over* the child> <hit

Preposition *Super*sense Disambiguation

Cross-lexical classes

Interpretable names like `Topic`

Supersenses = shared functions

They ran **to** the roof **for** a quick escape.

`Destination` `Purpose`

They ran **for** the roof **to** escape the cops.

CARMLS:

Case and Adposition Representation for Multi-Lingual Semantics

"*Comprehensive Supersense Disambiguation of English Prepositions and Possessives*".
Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan and Omri Abend. *ACL 2018.*

- New high quality dataset with **all** prepositions and possessives (types+tokens) semantically annotated
- Feature based and neural disambiguation system

Challenges

- Representations

  Linguistic intuitions can help develop useful *structured* representations for reasoning about text

- Data
  - Annotated data is a precious resource
  - How do we get around the need for annotated data?

- Efficiency
  - Predictive models need to scale (eg: to every doctor and patient in the world)
  - How can we avoid unnecessary computation?

Whence Data?

Most successes of data-driven computing made possible by supervised methods.

Data is today's most precious commodity.

Tao Li
[EMNLP 2016]

We may be able to exploit easier-to-obtain signals

Many problems where we want to predict linguistic structure, but

- We don't have (much) data for that representation
- We have data for a related phenomenon

Challenges

- Representations

  Linguistic intuitions can help develop useful *structured* representations for performing reasoning about text

- Data

  Can exploit the need for consistency between different kinds of linguistic predictions to act as a surrogate for training data.

- Efficiency
  - Predictive models need to scale (eg: to every doctor and patient in the world)
  - How can we avoid unnecessary computation?

Challenge:
Scaling Human Language Technology

Natural language understanding needs to scale to every device, book, document, utterance out there

The Long Room of the Old Library at Trinity College Dublin. Source: Wikipedia

A ground up rethinking

- Faster inference [Srikumar et al 2012, Kundu et al 2013]
  – No need to solve certain inference problems if they are similar to ones we have already solved
- Faster feature extraction [Srikumar 2017]
  – Automatically refactor feature extraction to reduce redundant computation
- Faster dot products [Shafiee et al 2016]
  – Better machine level support, novel hardware

A ground up rethinking

- Faster inference [Srikumar et al 2012, Kundu et al 2013]
  - No need to solve certain inference problems if they are similar to ones we have already solved
- Faster feature extraction [Srikumar 2017]
  - Automatically refactor feature extraction to reduce redundant computation
- Faster dot products [Shafiee et al 2016]
  - Better machine level support, novel hardware

Faster feature extraction

[Srikumar 2017]

Feature extraction has always been "somebody else's problem"

Can be time consuming and adds up over a classifier's lifetime

Can we make feature extraction faster?

Let us examine feature extraction

A closer look at feature extraction

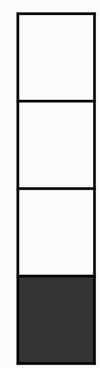
Example: Classifying words

The cat in a hat sat on the mat.

Some "typical" NLP features may include:

`word`: The surface form of a word $\xrightarrow{\text{produces}}$ {word=hat}

Really shorthand for the sparse vector that is zero everywhere except for one element whose basis corresponds to the string "word=hat"



"word=hat"

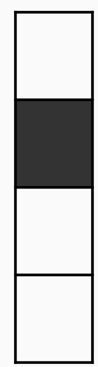
Example: Classifying words

The cat in a hat sat on the mat.

Some "typical" NLP features may include:

`pos`: Part of speech of a word $\xrightarrow{\text{produces}}$ {pos=Noun}

Really shorthand for the sparse vector that is zero everywhere except for one element whose basis corresponds to the string "pos=Noun"



"pos=Noun"

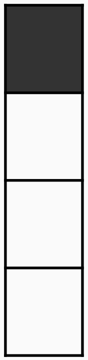
Example: Classifying words

The cat in a hat sat on the mat.

Some "typical" NLP features may include:

`dep_path`: Dependency path to root $\xrightarrow{\text{produces}}$ {dep_path=cat↑hat↑sat}

Really shorthand for the sparse vector that is zero everywhere except for one element whose basis corresponds to the string "dep_path=cat↑hat↑sat"



"dep_path=cat↑hat↑sat"

Feature extractors are functions

$\boldsymbol{x}$ = The cat in a hat sat on the mat.

`word` ($\boldsymbol{x}$) : The surface form of a word → {word=hat}

`pos` ($\boldsymbol{x}$) : Part of speech of a word → {pos=Noun}

`dep_path` ($\boldsymbol{x}$) : Dependency path to root → {dep_path=cat↑hat↑sat}

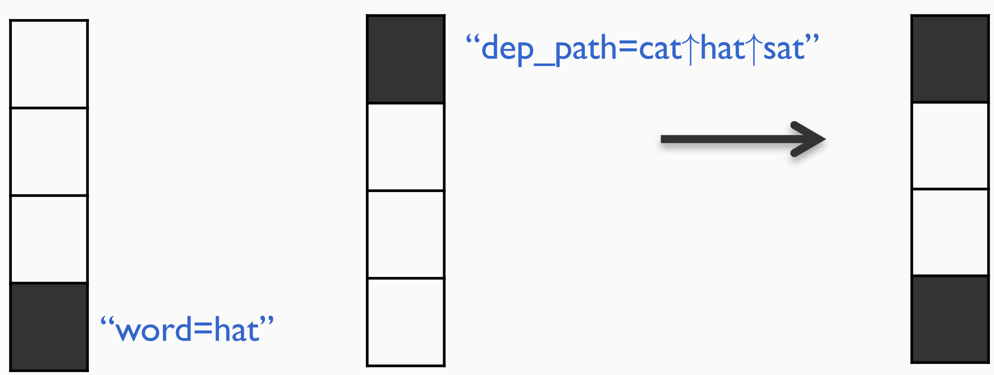
They map any objects (such as words, images) to a vector space (possibly infinite dimensional)

1. Feature addition

This is a function that maps inputs to a vector space i.e, it is a feature extractor

$\boldsymbol{x}$ = The cat in a hat sat on the mat.

$\texttt{word}(\boldsymbol{x})\ +\ \texttt{dep\_path}(\boldsymbol{x})$ $\qquad$ $\texttt{word+dep\_path}(\boldsymbol{x})$



$\mathbf{F}$ = Set of all feature extractors

$+ : F \times F \rightarrow F$

Operators over feature extractors

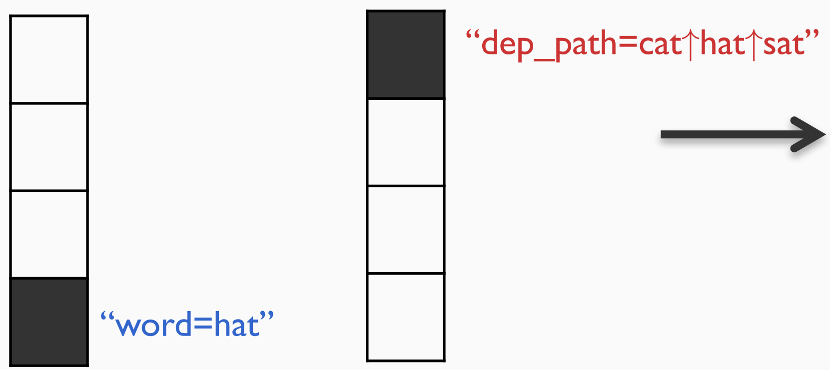
2. Feature conjunction

This is a function that maps inputs to a vector space
i.e, it is a feature extractor

$\boldsymbol{x}$ = The cat in a hat sat on the mat.

`word(`$\boldsymbol{x}$`)` **&** `dep_path(`$\boldsymbol{x}$`)`

`word&dep_path(`$\boldsymbol{x}$`)`



"word=hat & dep_path=cat↑hat↑sat"

"dep_path=cat↑hat↑sat & word=hat"?

Boolean conjunctions are symmetric

$\mathbf{F}$ = Set of all feature extractors

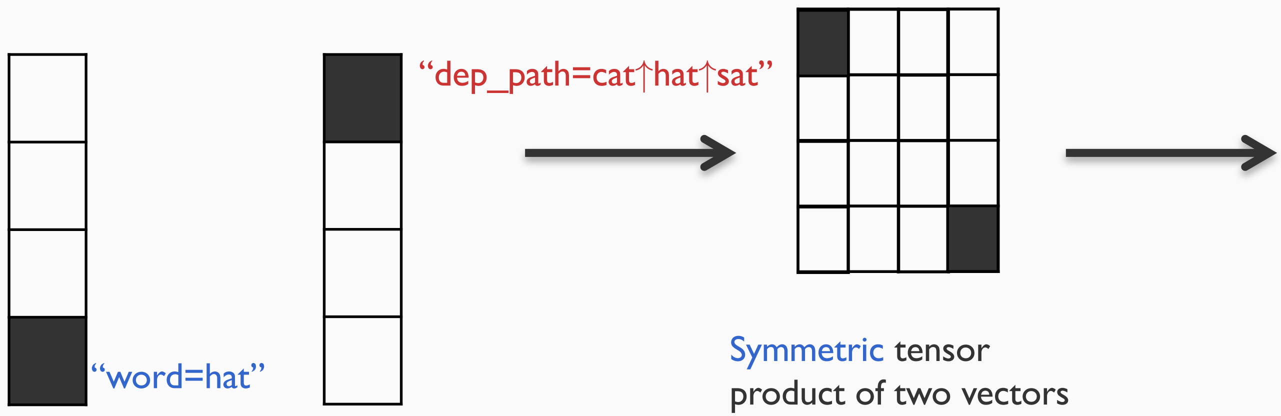
$\boldsymbol{\&} : F \times F \rightarrow F$

2. Feature conjunction

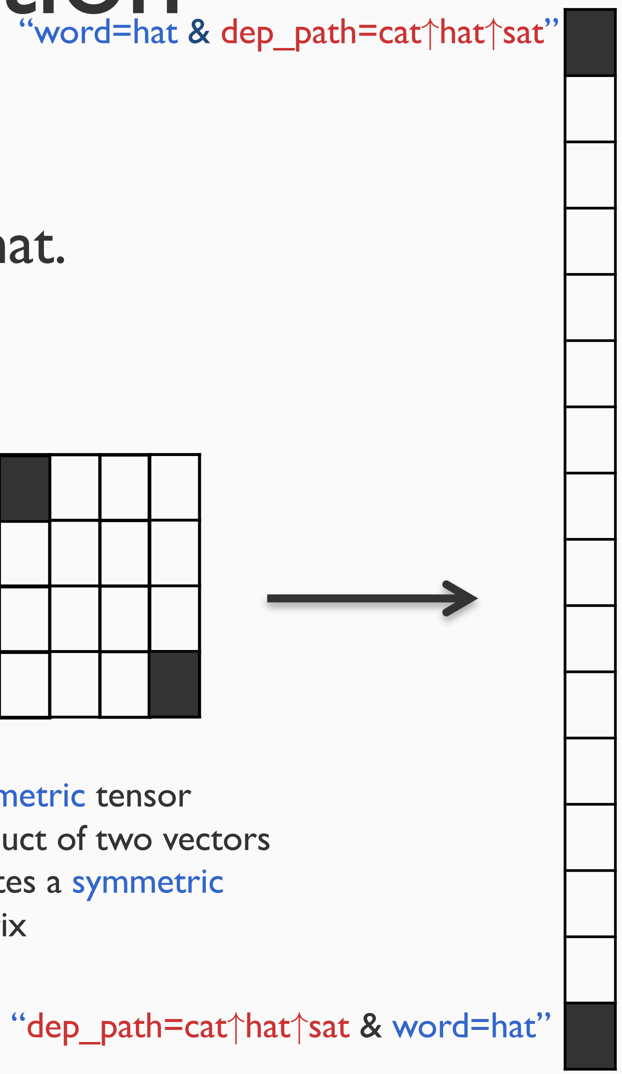
`word&dep_path(`$\boldsymbol{x}$`)`

$\boldsymbol{x}$ = The cat in a hat sat on the mat.

`word(`$\boldsymbol{x}$`)` **&** `dep_path(`$\boldsymbol{x}$`)`



Symmetric tensor product of two vectors creates a symmetric matrix



Formalizing Feature Extraction

Feature extractors Functions from any objects to a vector space. Examples: `word`, `pos`, `dep_path`, etc.

Special feature extractors

- $\mathbb{0}$: The **`zero`** feature extractor, maps inputs to the zero vector.
- $\mathbb{1}$: The **`bias`** feature extractor, maps inputs to a fixed bias feature vector.

Operators on feature extractors Compose feature functions to produce new feature extractors.

Feature addition:
(eg: `word+pos`)

Feature conjunction:
(eg: `word&pos`)

These definitions align with the intuitive interpretations of these operators.

An algebra for feature extraction

Theorem: The set of feature extractors, with feature addition and conjunction, forms a **commutative semiring**.

1. Feature addition is
   - Associative
     `(a+b)+c = a+(b+c)`
   - Commutative `a+b = b+a`
   - **`zero`** is the identity element

2. Feature conjunction is
   - Associative
     `(a&b)&c = a&(b&c)`
   - Commutative `a&b = b&a`
   - **`bias`** is the identity element

3. Multiplication distributes over addition
   `a&(b+c) = a&b+a&c`

4. Conjoining with the zero feature extractor gives back zero

An algebra for feature extraction

Theorem: The set of feature extractors, with feature addition and conjunction, forms a **commutative semiring**.

1. Feature addition is
   - Associative
     `(a+b)+c =`
   - Commutative
   - `zero` is the ic

2. Feature conjunction is

   `&c)`

   `b&a`

   element

So what?

3. Multiplication distributes over addition
   `a&(b+c) = a&b+a&c`

4. Conjoining with the zero feature extractor gives back zero

An opportunity for speedup

Apply distributive property to refactor feature extractors

`a&b+a&c`

Two conjunctions
One addition

Looks trivial. But saves computation in two ways:

1. Fewer conjunctions
2. Can automatically move expensive feature extractors outside

And this is done at a symbolic level before any feature extractor is ever applied

Can we do this systematically?

Algebra begets an algorithm

Commutative semirings admit the use of the **Generalized Distributive Law (GDL) algorithm** to calculate sums of products faster.

A message passing algorithm that takes familiar forms for various semirings.

Eg: Belief propagation, Baum-Welch, Viterbi, etc.

Aji, Srinivas M., and Robert J. McEliece. "The generalized distributive law." *IEEE transactions on Information Theory* 46.2 (2000): 325-343.

Making Feature Extraction Faster

$$f = w + p + d + w\&d + p\&d$$

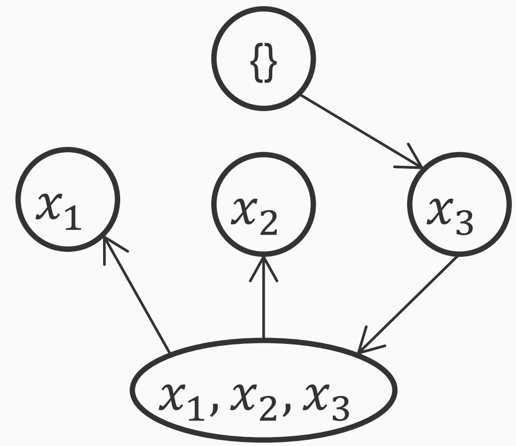
1. Convert any feature extractor into a canonical sum of products form

   Now the goal is compute this sum of products efficiently.

2. Construct a junction tree and assign local potential functions to each node

3. Run message passing from leaves to root (using the semiring operators)

$$f = w\&\mathbb{1}\&\mathbb{1} + \mathbb{1}\&p\&\mathbb{1} + \mathbb{1}\&\mathbb{1}\&d + w\&\mathbb{1}\&d + \mathbb{1}\&p\&d$$



Message at root:

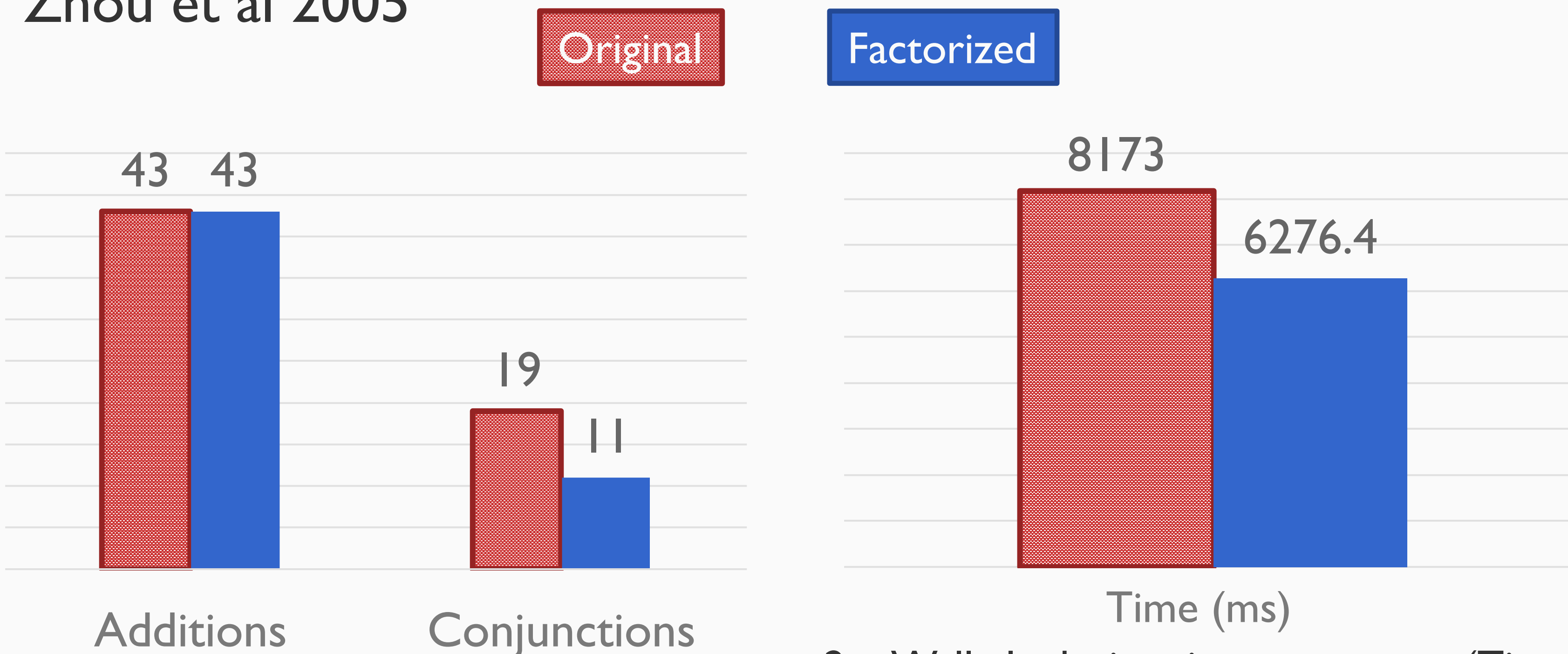
$$w + p + d\&(\mathbb{1}+w+p)$$

Functionally equivalent factorized feature extractor. One less conjunction, and dependency path is computed only once.

Does this work?

1. ACE relation extraction

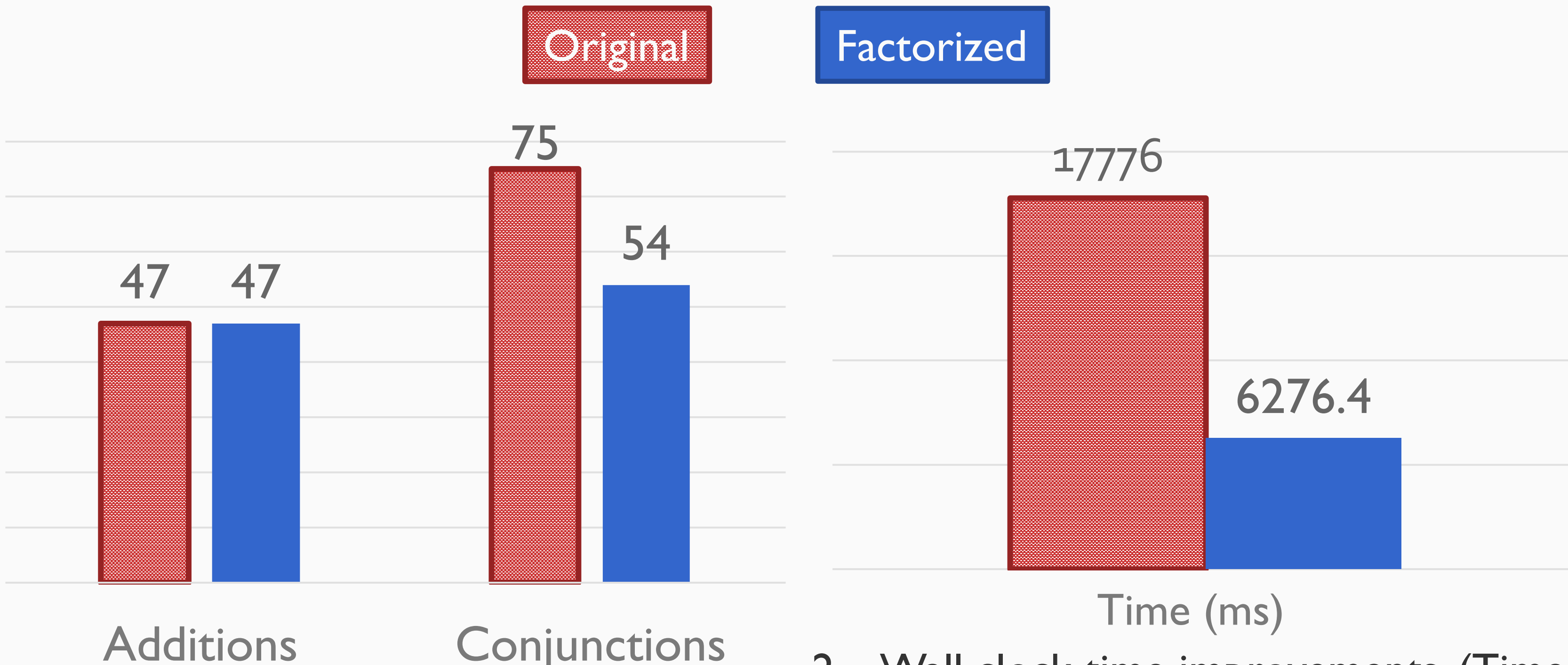
Assigning a relation label to a pair of entities. Feature set from Zhou et al 2005



1. Refactoring decreases number of conjunctions (at template level)
2. Wall clock time improvements. (Time in ms averaged multiple runs over the dataset.)

2. Text Chunking

CoNLL text chunking task. Feature set from Martins et al 2011



1. Refactoring decreases number of conjunctions (at template level)
2. Wall clock time improvements. (Time in ms averaged multiple runs over the dataset.)

An algebra for feature extraction

A formalization of feature extraction → An algebra of feature extractors → A message passing algorithm that makes feature extraction faster

The punchline: A way to automatically refactor feature extractors to be faster

A ground up rethinking

- Faster inference [Srikumar et al 2012, Kundu et al 2013]
  – No need to solve certain inference problems if they are similar to ones we have already solved

- Faster feature extraction [Srikumar 2017]
  – Automatically refactor feature extraction to reduce redundant computation

- Faster dot products [Shafiee et al 2016]
  – Better machine level support, novel hardware

Challenges

- Representations

  Linguistic intuitions can help develop useful *structured* representations for performing reasoning about text

- Data

  Exploit the need for consistency between different kinds of linguistic predictions to act as a surrogate for training data

- Efficiency

  Do not perform any redundant computation by automatically discovering regularities

This talk

Natural language processing can revolutionize many fields

…but we need to overcome several challenges.

- Empathy and the Machine: A case study of NLP and Mental Health
- Challenge 1: The importance of representations
- Challenge 2: Scaling NLP for everyone

Questions?