

Natural Language Processing

Vivek Srikumar



Natural Language Processing

An attempt to replicate (in computers) a phenomenon *only* humans are known to exhibit.

Outline

1. Wouldn't it be great if computers can understand natural language!
 - *Ahem... not so fast...*
2. Statistics to the rescue
 - Or: The present state of NLP
3. Towards text understanding

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

Sarah Able = She - *Co-reference resolution*

8 X 2 Towns, Lives(She, X) , Lives(Martin, X)

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

Lives(Martin the Monkey, an oak tree)

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

How do we
know this?

Really?

Bananaville 2 Towns

Words are a very fantastical banquet,
just so many strange dishes

And yet, we seem to do fine

We can understand and generate language effortlessly.

Almost

Language is *ambiguous*

Which seal?

Tuna Recall Blamed on Seal

Mar 7, 2013 11:47am

By Katie Moisse
@katemoisse

A photograph showing a large stack of Bumble Bee Chunk Light Tuna cans. The cans are white with green and red accents, featuring the brand name "BUMBLE BEE" and "Chunk Light Tuna".

Bumble Bee has recalled some cans of tuna. (Image credit: Richard B. Levine via Newscom)

Tuna giant Bumble Bee Foods has [recalled cans](#) of its flaky fish because of a possible problem with the seal.

Seal

From Wikipedia, the free encyclopedia

Seal most commonly refers to:

- [Pinniped](#), a diverse group of semi-aquatic marine mammals, many of which are called seals
 - [Earless seal](#), or "true seal"
 - [Fur seal](#)
- [Seal \(emblem\)](#), a device to impress an emblem, used as a means of authentication
- [Seal \(mechanical\)](#), a device which helps prevent leakage, contain pressure or hold a closure
- [Seal \(surname\)](#), an Indian surname of Hindu Bengalis and Rajputs

Seal may also refer to:

Arts and entertainment [edit]

- [Seal \(musician\)](#) (born 1963), English singer and songwriter, or any of his three albums
 - [Seal \(1991 album\)](#)
 - [Seal \(1994 album\)](#), sometimes referred to as *Seal II*
 - [Seal \(2003 album\)](#), sometimes referred to as *Seal IV* (following his third studio album)
- [Seal Online](#), a 2003 massively multiplayer online role-playing game

Law [edit]

- [Record sealing](#)
- [Seal \(contract law\)](#), a legal formality for contracts and other instruments
- [Seal \(East Asia\)](#), a stamp used in East Asia as a form of a signature

Military [edit]

- [Fairey Seal](#), a 1930s British carrier-borne torpedo bomber aircraft
- [HMS Seal](#), two Royal Navy ships and one submarine
- [USS Seal](#), two American submarines

Special forces [edit]

- [United States Navy SEALs](#), the principal special operations force of the U.S. Navy
- [Thai Navy Seals](#), Thai Navy special operations force modelled on the U.S. Navy SEALs
- [Naval Special Operations Group](#), also known as the Philippine Navy SEALs
- [ROKN UDT/SEALs](#), a component of the [Republic of Korea Navy Special Warfare Command](#)

People [edit]

- [Seal \(musician\)](#) (born 1963), English singer and songwriter
- [Aditya Seal](#) (born 1988), Indian actor
- [Barry Seal](#) (1939–1986), American drug smuggler and DEA informant
- [Brajendra Nath Seal](#) (1864–1938), Indian humanist philosopher and educationist

Language is *ambiguous*

Olive oil = Oil made *from* olives

Baby oil = Oil made *for* babies

Diamond ring = Ring with a diamond

Gold ring = Ring made of gold

Nose ring = Ring for the nose

Ambiguity can take many forms: Lexical, syntactic, semantic

Language is *ambiguous*

I ate sushi with tuna.



I ate sushi with chopsticks.

I ate sushi with a friend.

I saw a man with a telescope.

Stolen painting found by tree.

Ambiguity can take many forms: Lexical, syntactic, semantic

Language has complex structure

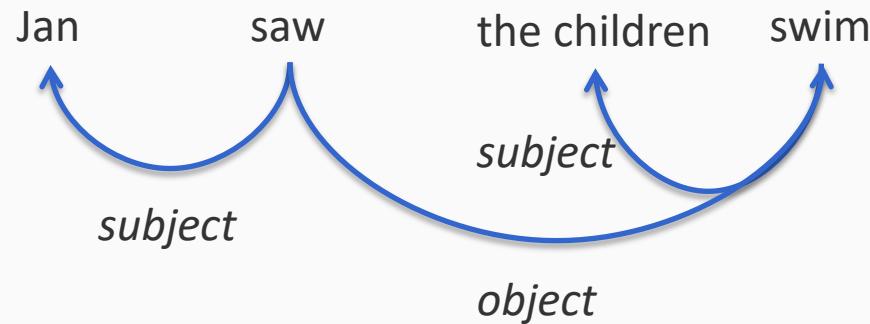
Mary saw a ring through the window and asked John for it.

(Why on earth did Mary ask for a window?!)

I am allergic to tomatoes. Also fish.

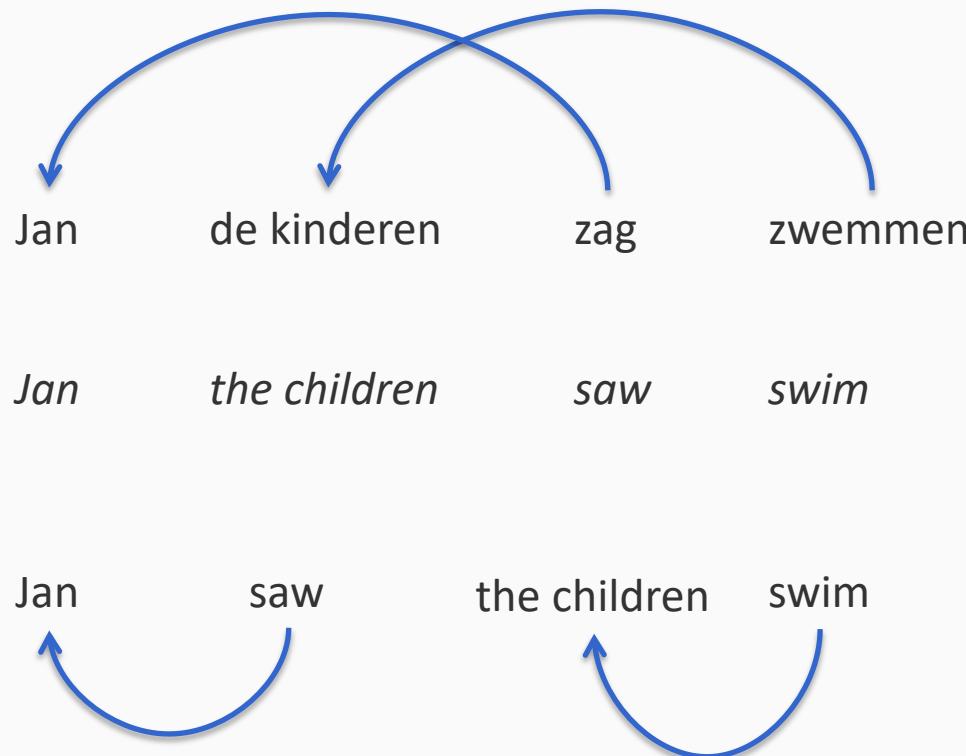
(Fish allergic to tomatoes?!)

Language has complex structure

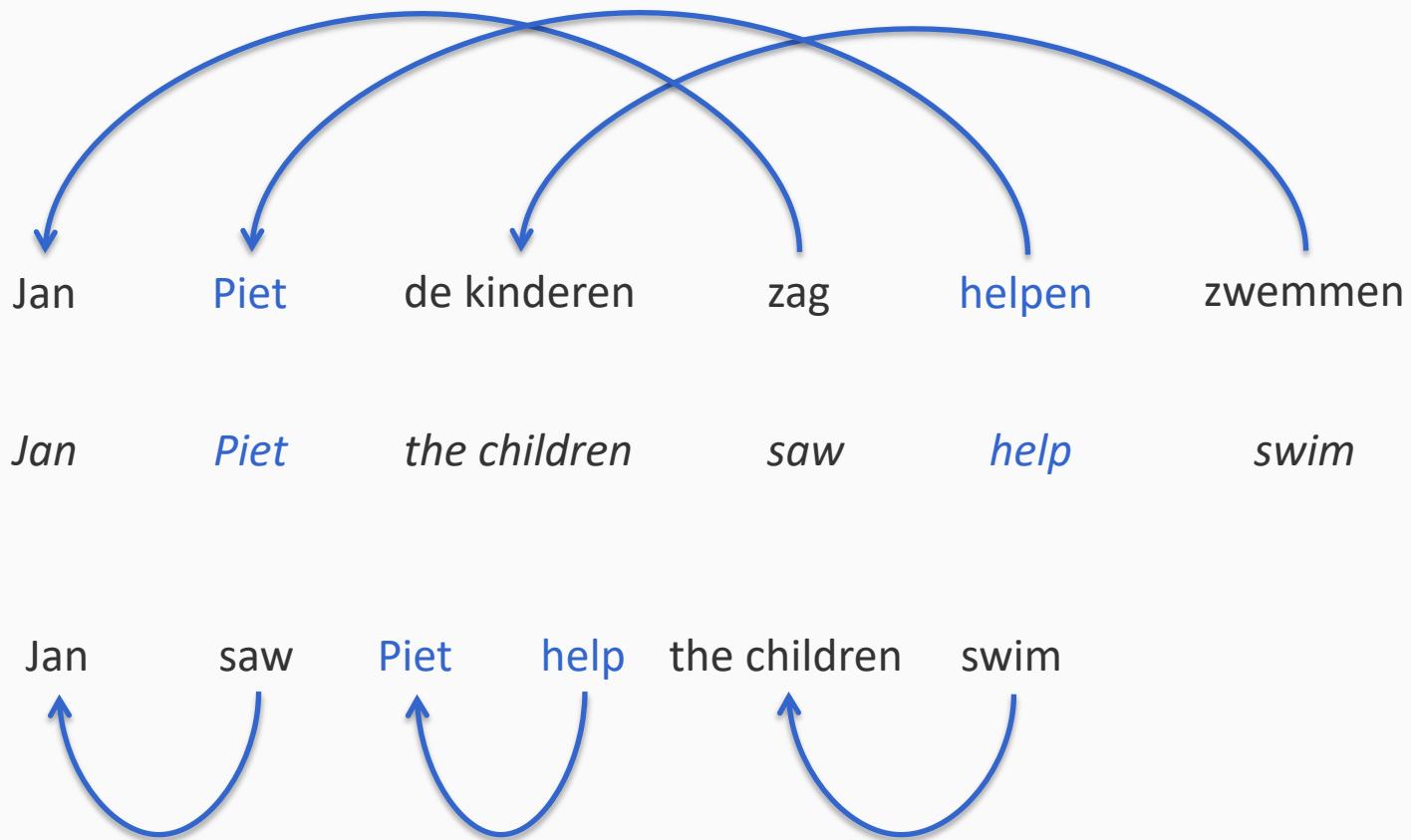


Parsing: Identifying the syntactic structure of sentences

Language has complex structure

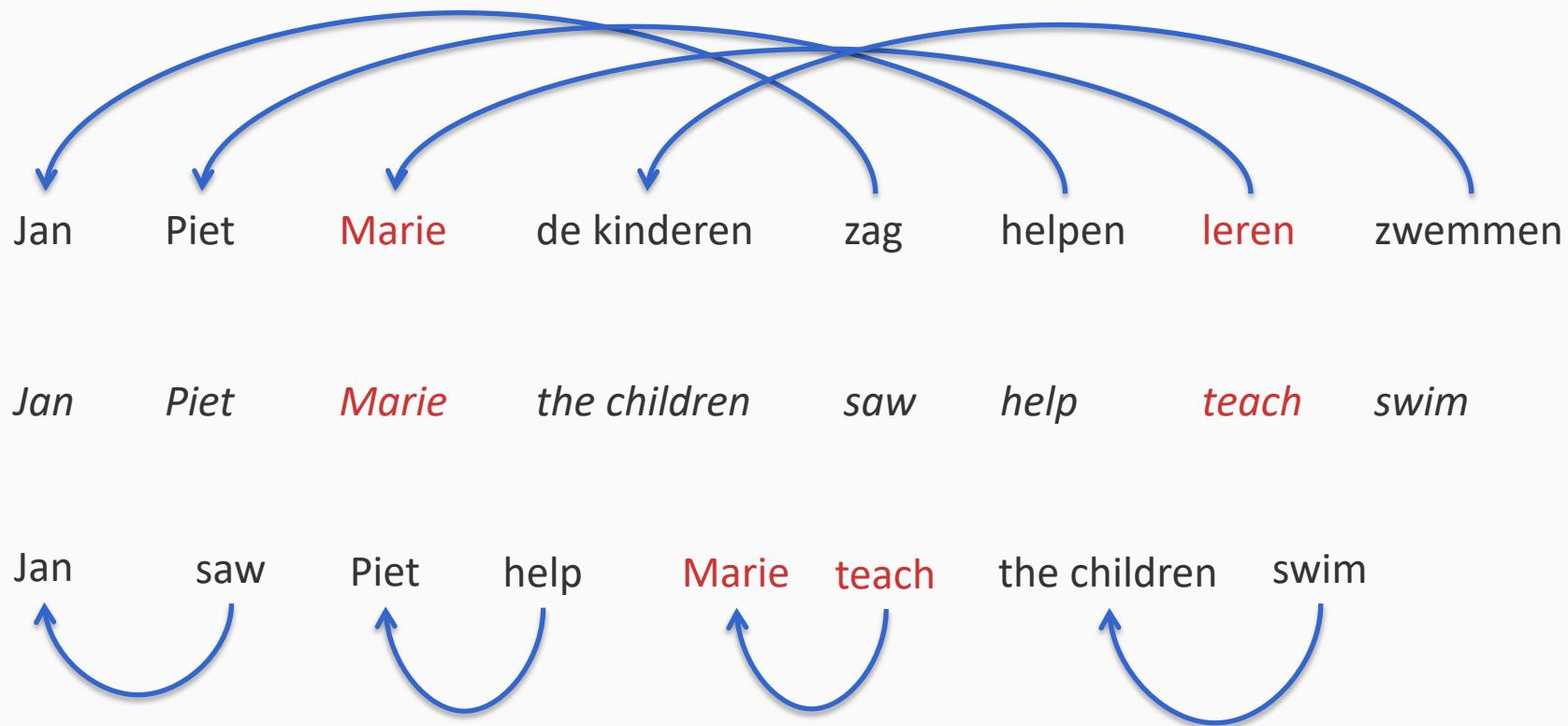


Language has complex structure



Language has complex structure

Natural language is not a context free language!



Many, many linguistic phenomena

Metaphor

- *makes my blood boil, apple of my eye, etc.*

Metonymy

- *The White House said today that ...*

A very long list...

And, we make up things all the time

If not actually disgruntled, he was far from being .

Twas , and the

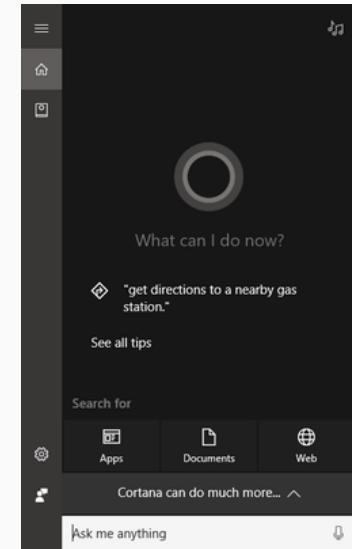
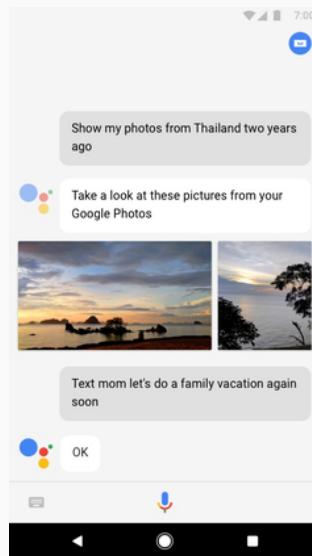
Did and in the :

All were the ,

And the .

But the situation is far from hopeless!

Virtual assistants



Web search

what is the capital of utah?

All News Images Videos More ▾ Search tools

About 51,000,000 results (1.48 seconds)

Utah / Capital



Salt Lake City

What color are bananas?

All Images Shopping News Videos More Settings Tools

About 43,600,000 results (0.75 seconds)

yellow

Sweet-banana skins are most commonly known to be **yellow**, but ripe banana skins can also be red, pink, purple, and black. Aug 10, 2017



Machine Translation

Translate

The screenshot shows a machine translation interface. On the left, a text input field contains the Dutch sentence "Jan de kinderen zag zwemmen". On the right, the English translation "John saw the kids swimming" is displayed. The interface includes language selection dropdowns at the top, with "Dutch - detected" and "Arabic" selected respectively. A "Translate" button is also visible. Below the text boxes are small icons for audio playback.

How did we get here?

Early NLP systems

- Some successful examples
 - ELIZA (Weizenbaum 1966): A chatbot that simulated a psychotherapist
 - PARRY (Colby 1972): A chatbot that simulated a paranoid schizophrenic
 - SHRDLU (Winograd 1969): A program that accepts text input and reasons about the blocks world
 - PUNDIT (Palmer 1989): An information extraction program
- Rule based, hand crafted
- Rich representations of the semantics of text

Remember this one. We will get back to this point about representation

The problem with GOFAI for NLP

Good Old Fashioned Artificial Intelligence

- Worked well, but **only** for a very small set of examples
- Took several years of human effort to create
- And still did not generalize well
 - Because too many hand-crafted rules

Hand-crafted rules do not really reflect how language works in practice!

The answer: *Power to the data*

Statistical methods for NLP

a.k.a

Machine learning to the rescue

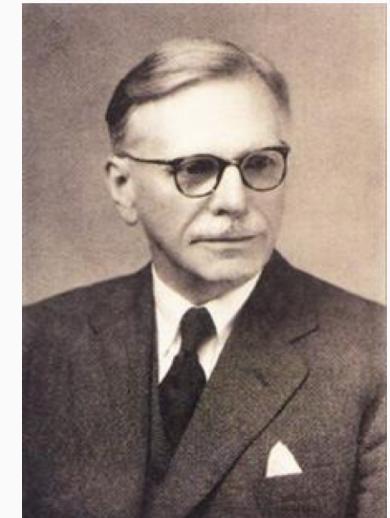


Another old idea

“You shall know a word by the company it keeps.”

- J. R. Firth (1957)

...that is, if you want to understand language, you need to know how it has been used before



The company words keep

I would like to eat a _____ of cake

peace or piece?

An idea

- Train a *binary classifier* to make this decision
- Use indicators for neighboring words as features
 - Spelling/grammar checkers

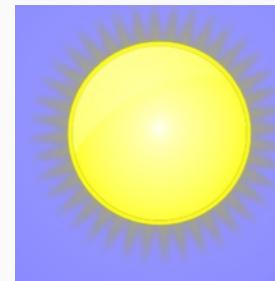
Works surprisingly well!

Data + features + learning algorithm = Profit!

What do words mean?

How do we represent meaning in a computationally convenient way?

bunny and *sunny* are only one letter apart but very far in meaning



bunny and *rabbit* are very close in meaning, but look very different

And can we *learn* their meaning from data?

Words = vectors

A very old idea: Represent every word by a very high dimensional vector

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

A newer idea: Represent words by much lower dense vectors

- *Word embeddings*

Example 2: Part of speech tagging

Identify the part-of-speech of all words in a sentence



The standard solution: Train a [sequence model](#)

- Many flavors exist – we want to assign a probability to the entire sequence
- Train on what data: Penn Treebank

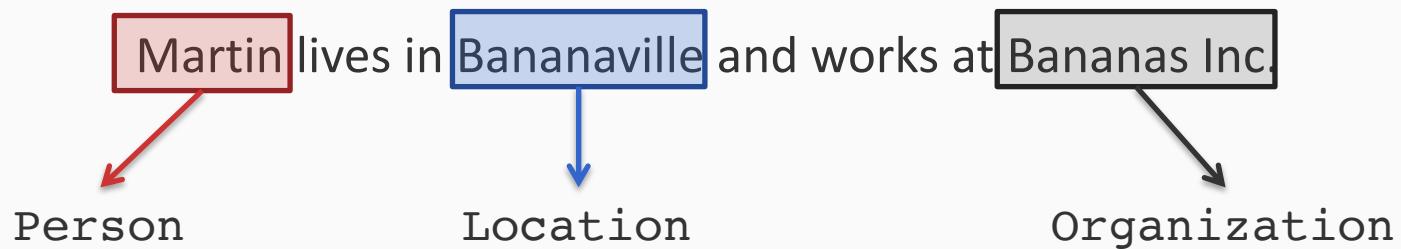


Example 3: Information Extraction

Build a database about the entities, events, etc in text

One task: Named entity recognition

- Identify all named entities in text and their types



Example 4: Sentiment Analysis

Predict whether some text has positive or negative polarity

The movie was so awesome that I want to watch it again.



The movie was terrible.



The movie feels like they finished filming before completing the script.



This movie was long and convoluted but I liked it anyway



Can be framed as a classification problem

Powered by probabilistic models

Example 5: Semantic Parsing

Input: “A python function that takes a name and prints the string Hello followed by the name and exits.”

Output:

```
def hello(name):  
    print("Hello , " + name)
```

Input: “Find the largest state in the US.”

Output:

```
SELECT name  
FROM us_states  
WHERE size = (SELECT MAX(size) FROM us_states)
```

What have we seen so far...

- Language is **ambiguous, complex & diverse** from a lexical, structural and semantic point of view
- Data-driven methods excel in the face of ambiguity
- To build a text understanding task, the right **representation** is important
- Understanding how language works (**the linguistics**) and where it is used (**the application**) can help model it

Recent research

Let's go back to reading comprehension

Reading comprehension is hard!

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

Reading comprehension is hard!

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

Reading comprehension is hard!

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**

Reading comprehension is hard!

Enable

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

Reading comprehension is hard!

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⁺.

The diagram features two black-bordered boxes at the top. The left box contains the word 'Enable' and has a curved blue arrow pointing down to the word 'split' in the text. The right box contains the word 'Cause' and has a curved blue arrow pointing down to the phrase 'transfer of the electrons and hydrogen ions'.

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 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⁺.

The diagram consists of two horizontal lines. The top line has a box labeled 'Enable' at its right end, with a curved arrow pointing to the word 'split' in the text. The bottom line has a box labeled 'Cause' at its right end, with a curved arrow pointing to the phrase 'transfer of the electrons and hydrogen ions'.

What can the splitting of water lead to?

- A: Light absorption
- B: Transfer of ions

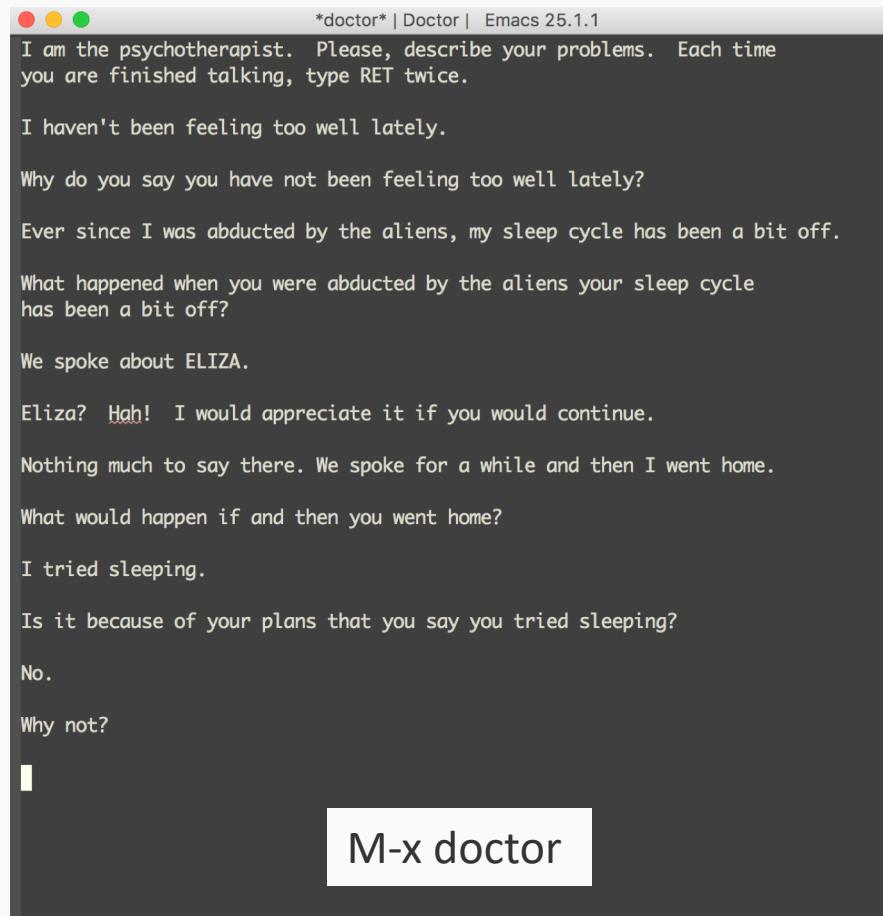
Gets ~70 percent answers correct

Recent research

Empathy and the machine

NLP and mental health treatment

Long history of therapist bots. Eg, ELIZA



The screenshot shows an Emacs window with the title bar "*doctor* | Doctor | Emacs 25.1.1". The buffer contains a simulated conversation:

```
I am the psychotherapist. Please, describe your problems. Each time
you are finished talking, type RET twice.

I haven't been feeling too well lately.

Why do you say you have not been feeling too well lately?

Ever since I was abducted by the aliens, my sleep cycle has been a bit off.

What happened when you were abducted by the aliens your sleep cycle
has been a bit off?

We spoke about ELIZA.

Eliza? Hah! I would appreciate it if you would continue.

Nothing much to say there. We spoke for a while and then I went home.

What would happen if and then you went home?

I tried sleeping.

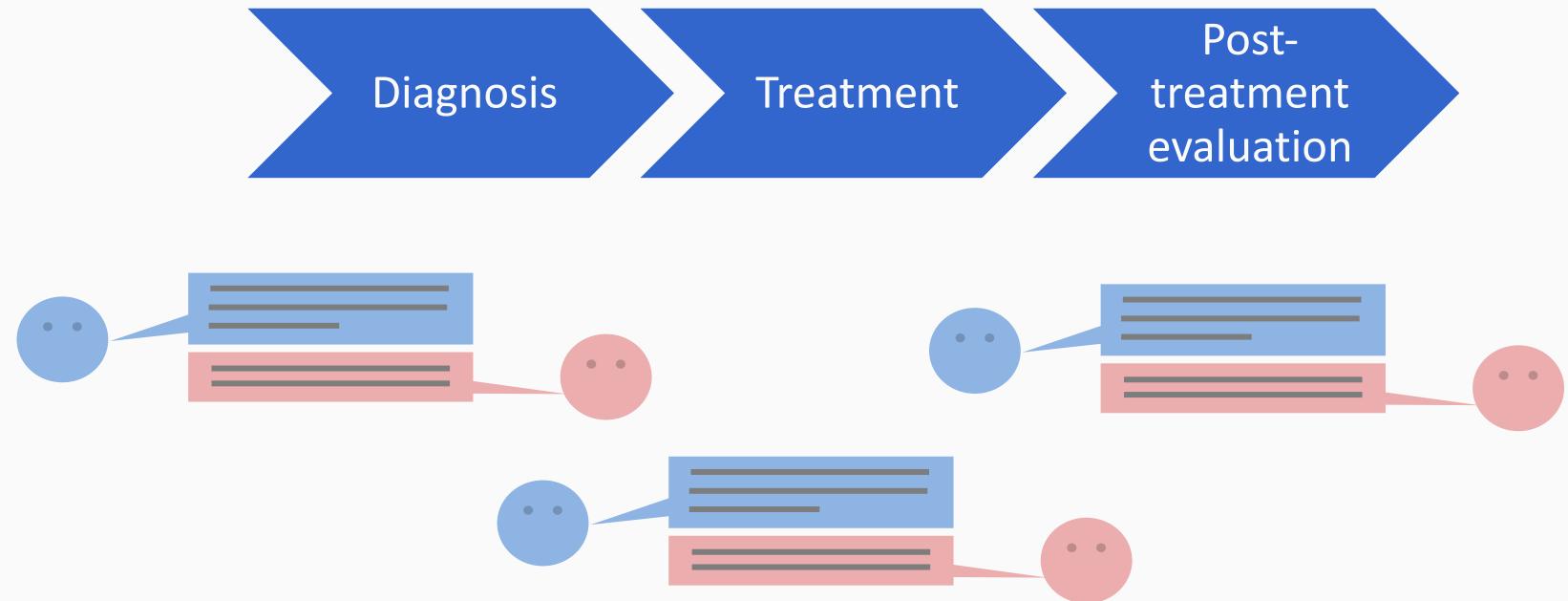
Is it because of your plans that you say you tried sleeping?

No.

Why not?

M-x doctor
```

A trip to a therapist...



Mental health therapy

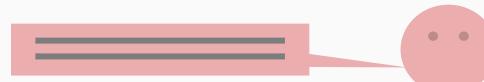
Essentially a dialogue between two people

- Are some treatments better than others?
- Are some counselors better than others?
- Can we measure patient improvements?
- How does the therapy process 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



TherapyBot [Trained on PatientBOT taken Eliza@17]
For training therapists

Therapist assistant

Assessing therapy sessions [Tanana et al 2016, Cao et al, 2019]

Tracking sentiment [Tanana et al 2015]

Omniscient helper for counselors *during* therapy

**NLP can transform mental
health care. But...**

Still a long way to go.

Challenges abound

Challenges

- **Representations**
 - *What is a **good** representation of language?*
- **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)?
- **Reasoning with language**
 - How do we build systems that reason about the world using statistical machinery?

The bigger picture

Wanted

Programs that read, understand and reason about text

- Some lessons from the field
 - Representation of text matters for building application
 - (GOFAI was right)
 - Statistical machine learning is here to stay
 - (Firth was right)
 - Knowledge about the application domain helps
 - Good data helps train a better model

Natural language can be fun but weird



<https://xkcd.com/1443/>

The future: Computers and human language

- Machine reading
 - A program that accumulates knowledge by reading
 - There is a lot of text out there....
- Question answering
 - Can a computer answer my questions? (requires reasoning)
 - “Would it be easier for me to fly from Phoenix to SLC or to drive on Sunday?”
- Semantic parsing in context
 - Can a computer *do* things based on what I instruct?
 - “Computer, change the first fifteen slides of my presentation to use the new style”
- Low resource languages
 - A lot of statistical NLP is driven by expensive to obtain datasets
 - Can we do all the above for Swahili/Zulu/Xhosa/Balochi/....?
- Understanding language
 - What does language mean anyway?
 - And, what does understanding mean anyway?

Ask me anything