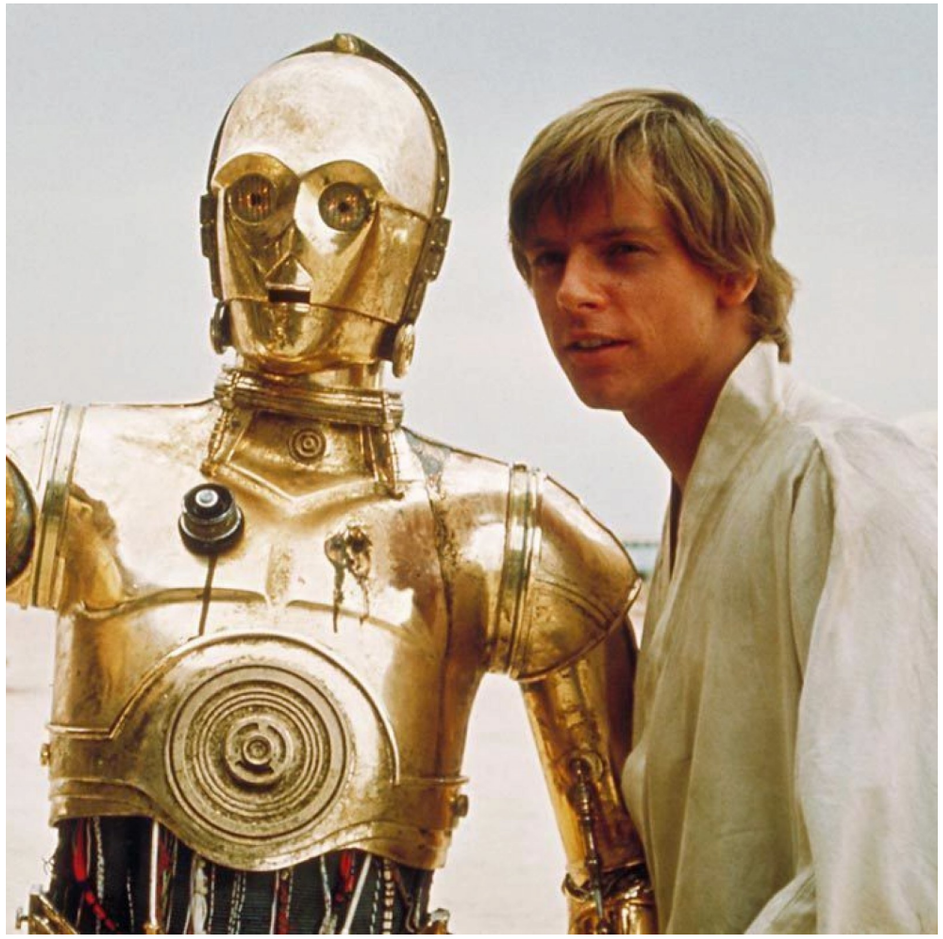
Natural Language Processing

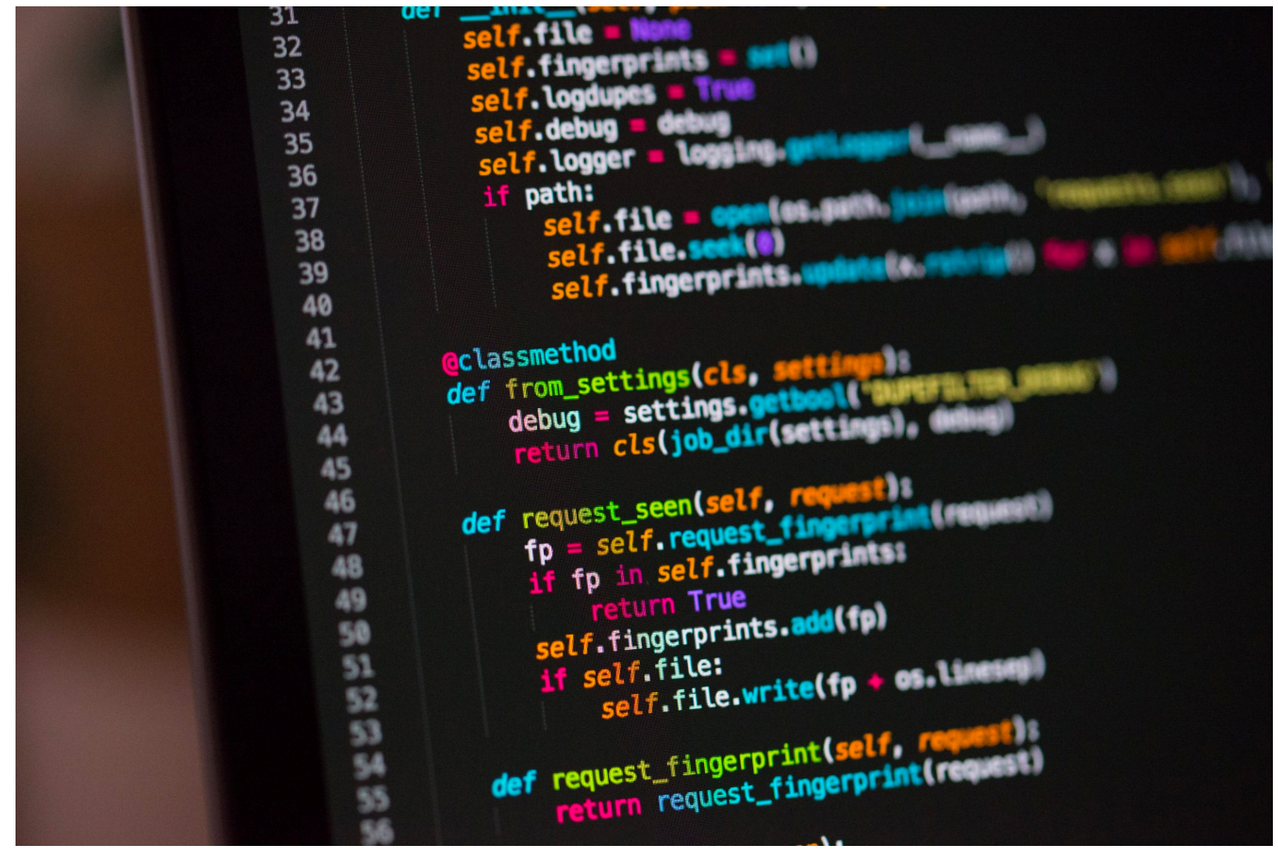
Yichu Zhou
University of Utah

How do we communicate with Computers?



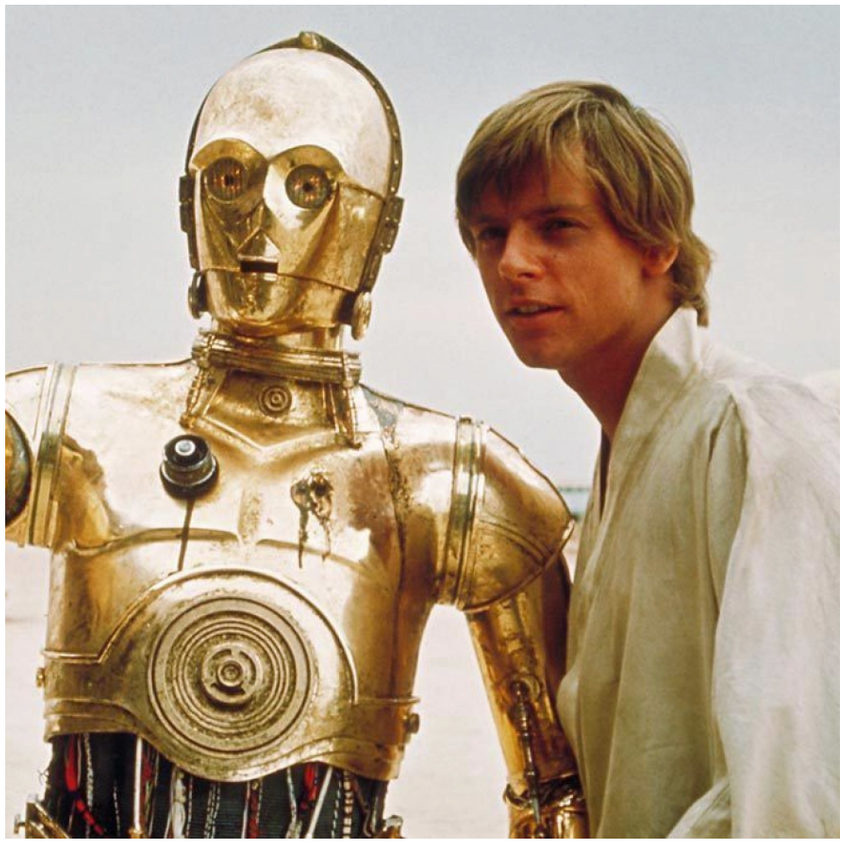
How we imagine talking to computers

How do we communicate with Computers?



How we actually talk to computers

Natural Language Processing



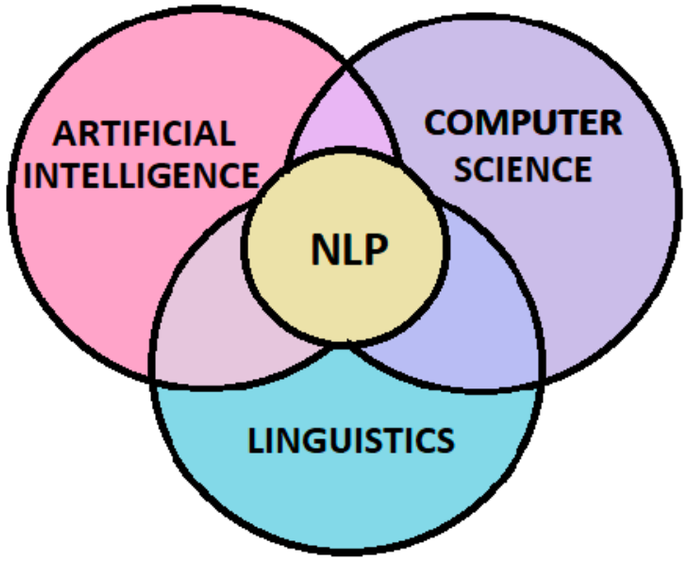
One of the ultimate goals

Overview

- What is NLP?
- What are the challenges?
- Techniques in NLP
- The Future

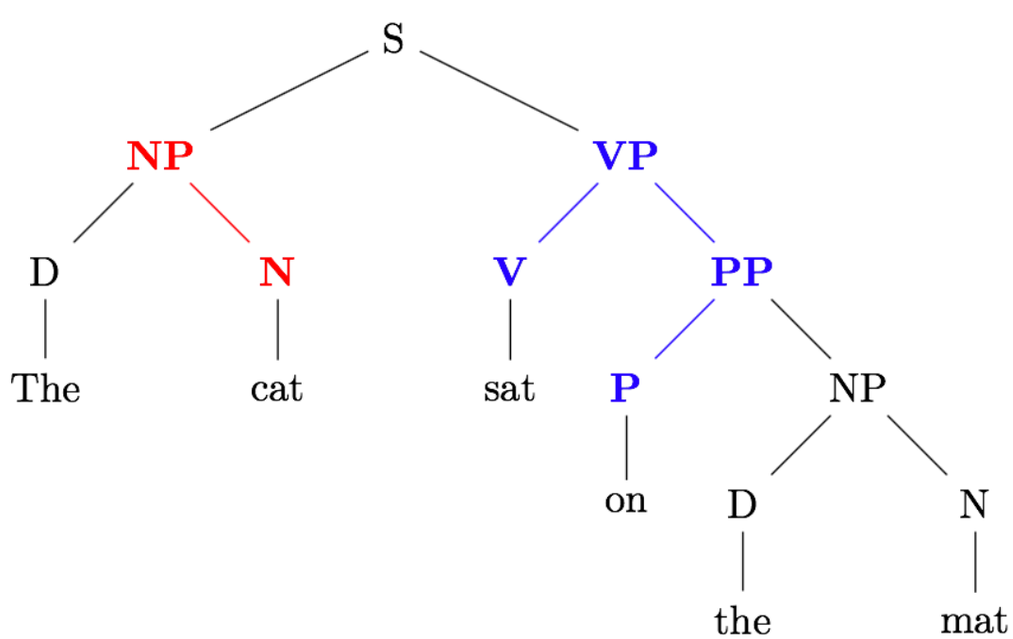
What is NLP

- Wiki: **Natural language processing (NLP)** is a subfield of **linguistics**, **computer science**, and **artificial intelligence** concerned with the interactions between computers and human language.

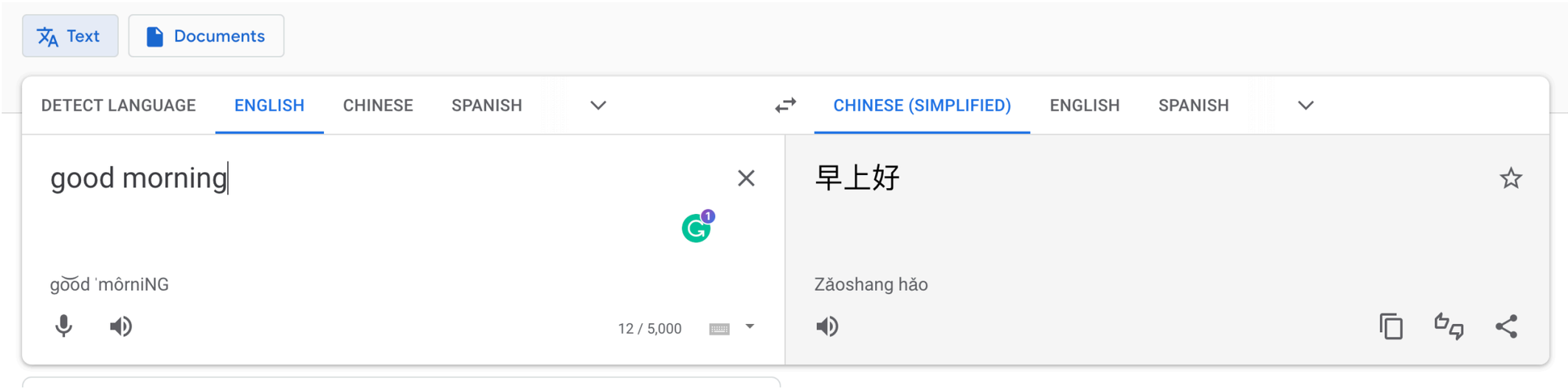


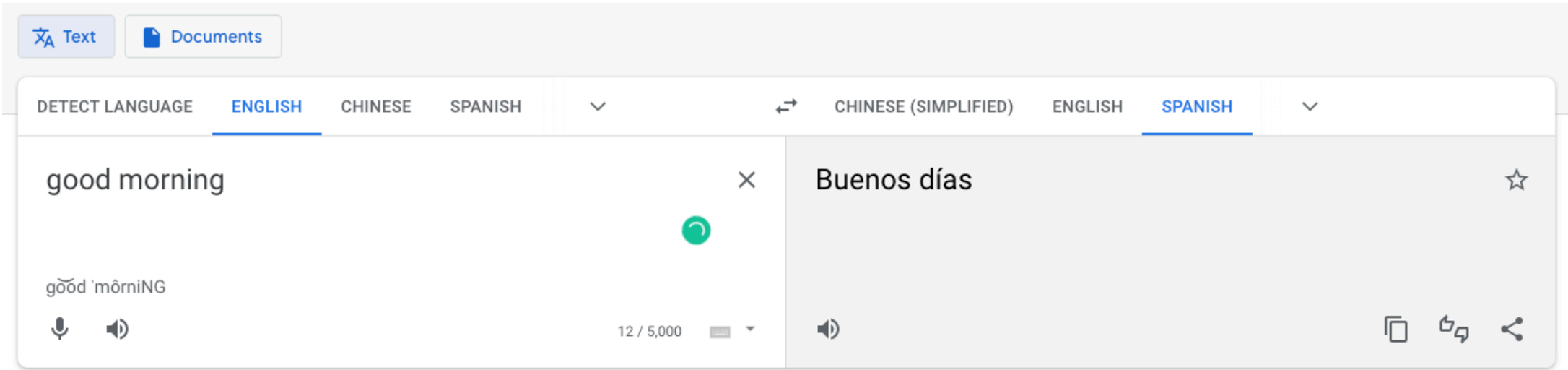
What is NLP

- Identify the structure and meaning of words, sentencs, texts and conversations.
- Deep understanding of broad languages



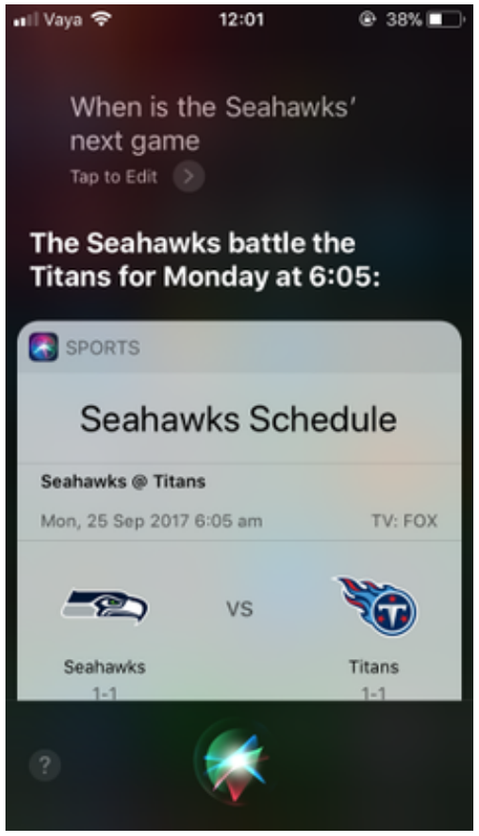
Machine Translation

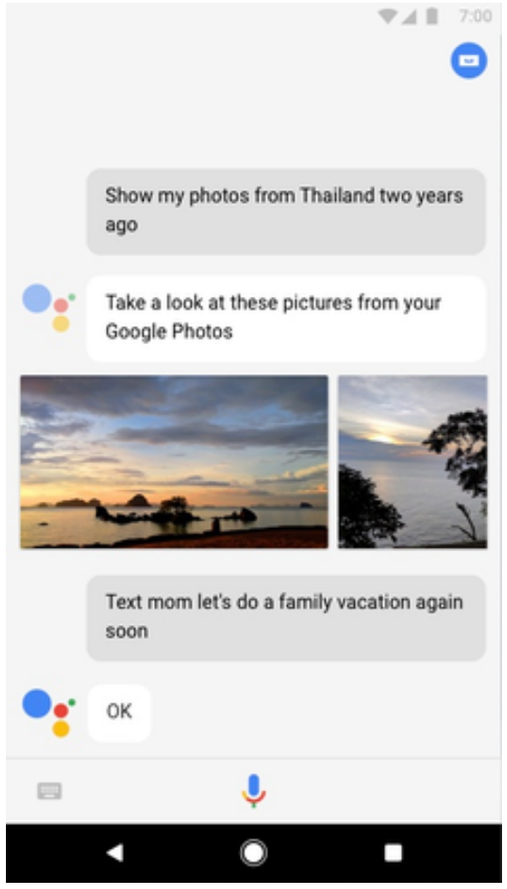


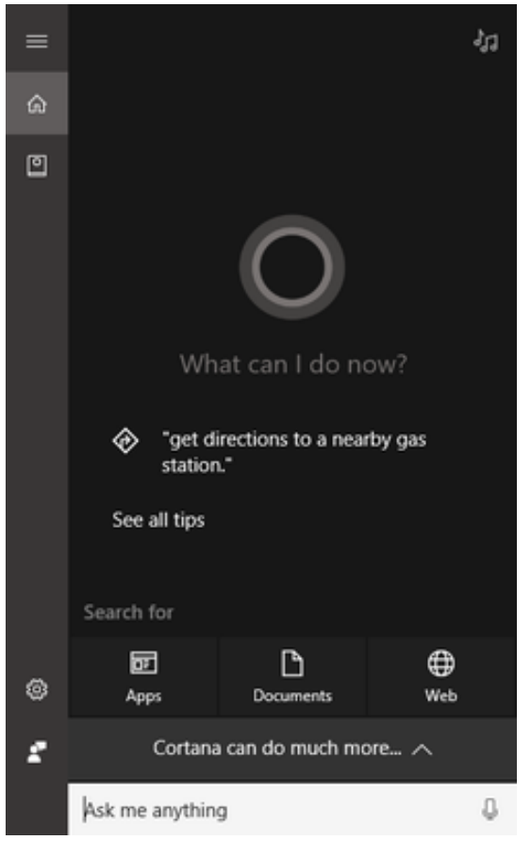


Visual Assistants

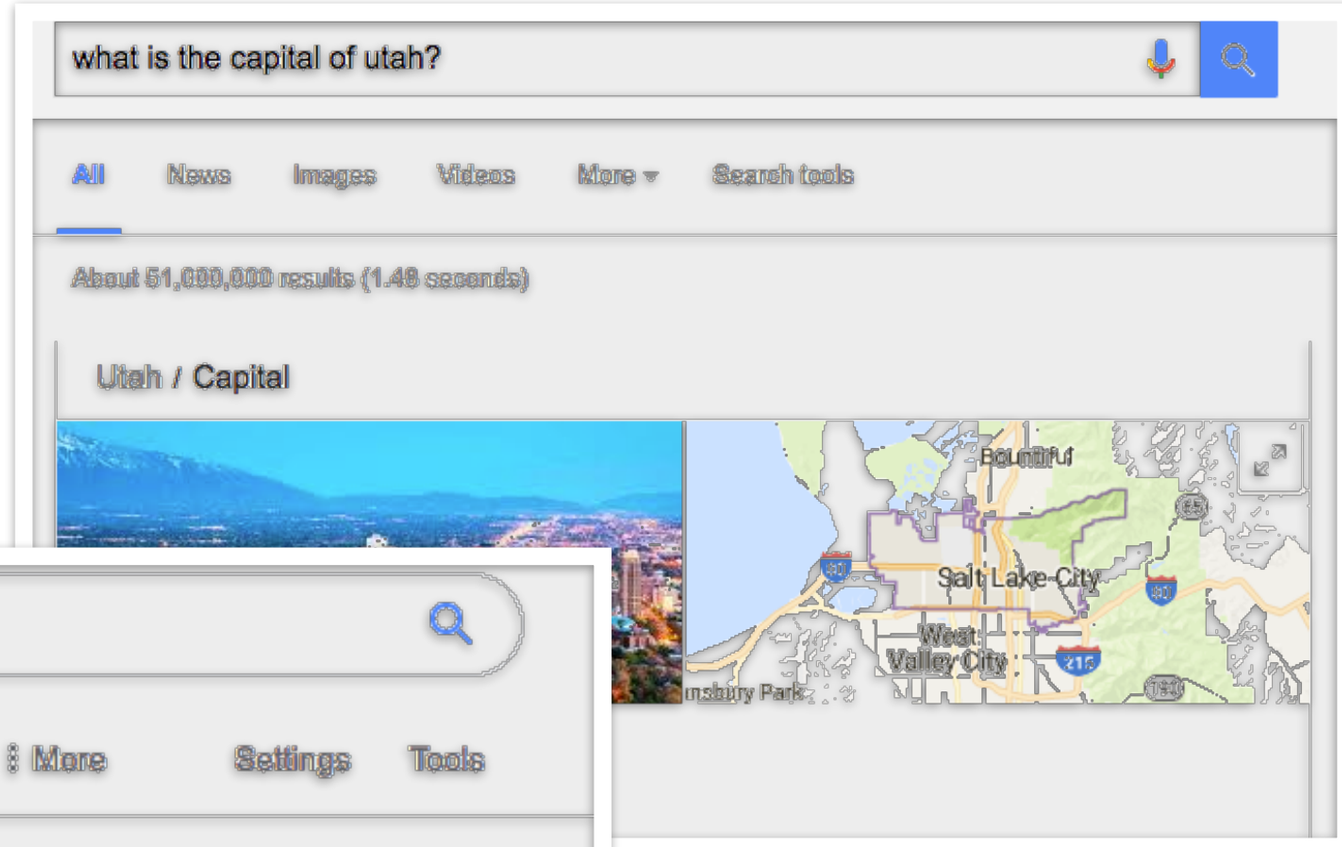
amazon alexa

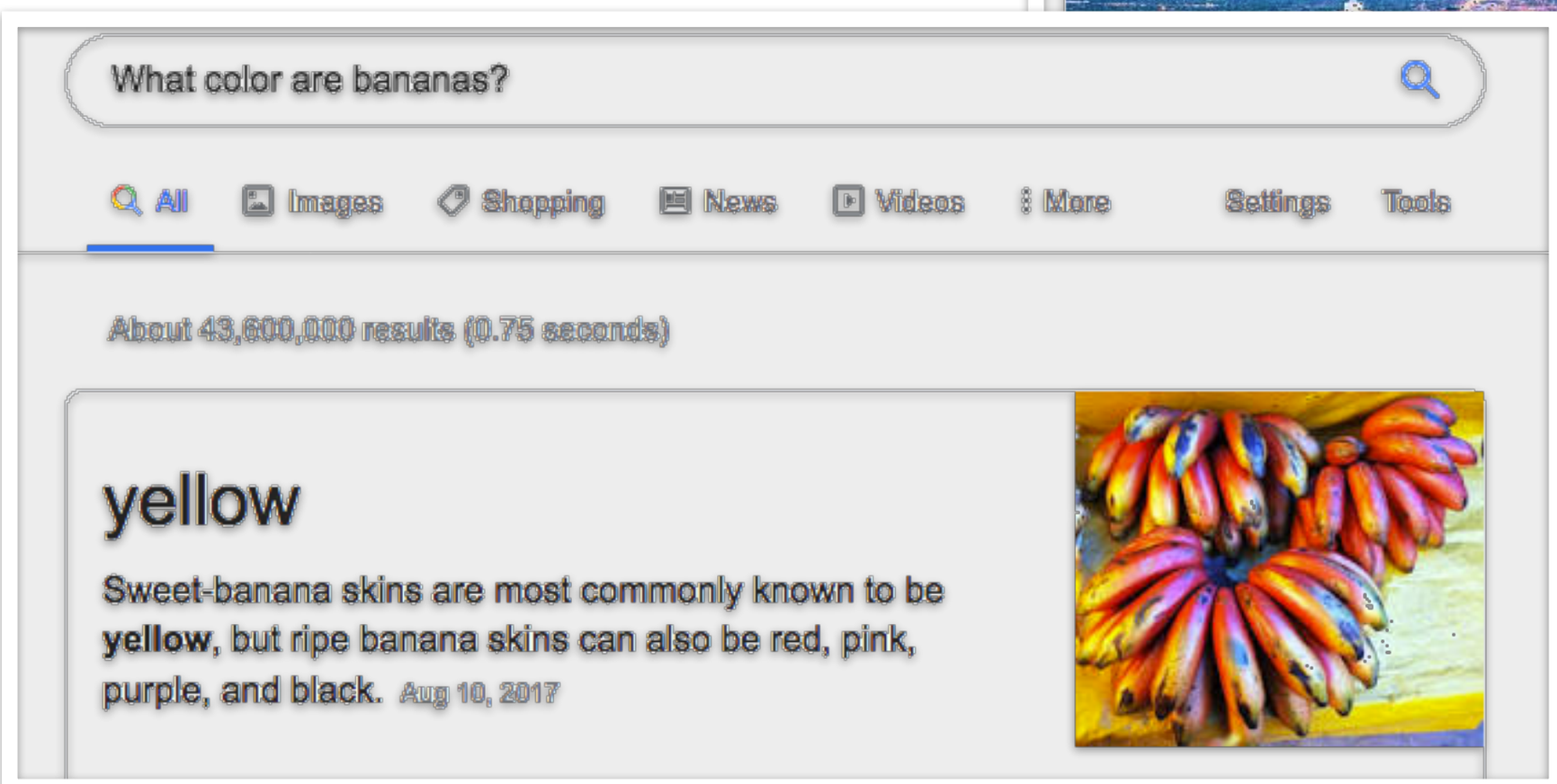






Web Search


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
West Valley 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

Overview

- What is NLP?
- What are the challenges?
- Techniques in NLP
- The Future

Ambiguous

Which seal?

Tuna Recall Blamed on Seal

4 Text

By Katie Moisse
@katiemoisse

Mar 7, 2013 11:47am

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 whic
  - Earless seal, or "true seal"
  - Fur seal
- Seal (emblem), a device to impress an emblem, used as a means of authe
- Seal (mechanical), a device which helps prevent leakage, contain pressure
- 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 th
  - *Seal* (1991 album)
  - *Seal* (1994 album), sometimes referred to as *Seal II*
  - *Seal* (2003 album), sometimes referred to as *Seal IV* (following his thirc
- *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
- Thai Navy Seals, Thai Navy special operations force modelled on the U.S.
- Naval Special Operations Group, also known as the Philippine Navy SEALs
- ROKN UDT/SEALs, a component of the Republic of Korea Navy Special W

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 educa

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

Ambiguous

I ate sushi with tuna.

I ate sushi with chopsticks.

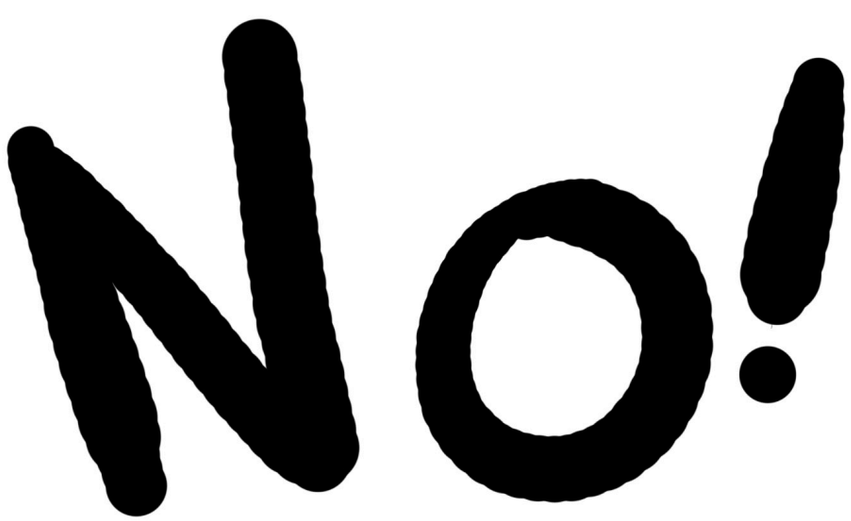
I ate sushi with a friend.

I saw a man with a telescope.

Ambiguity can take many forms: Lexical, syntactic, semantic

Dynamic

- Googlization
- Saaaaad
- No^2

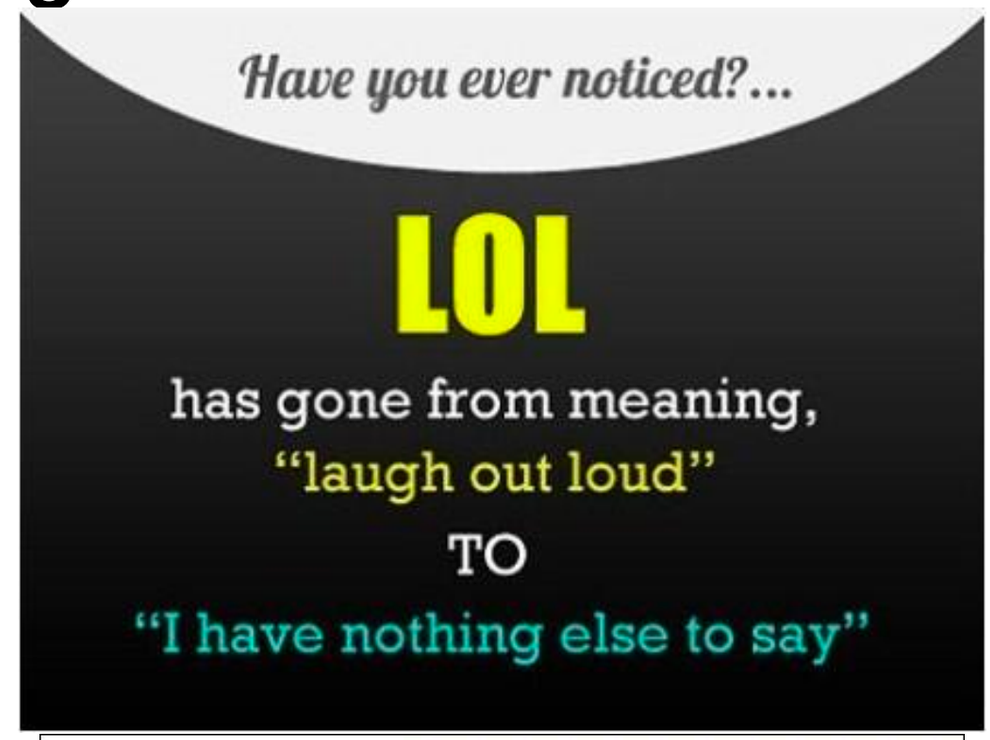


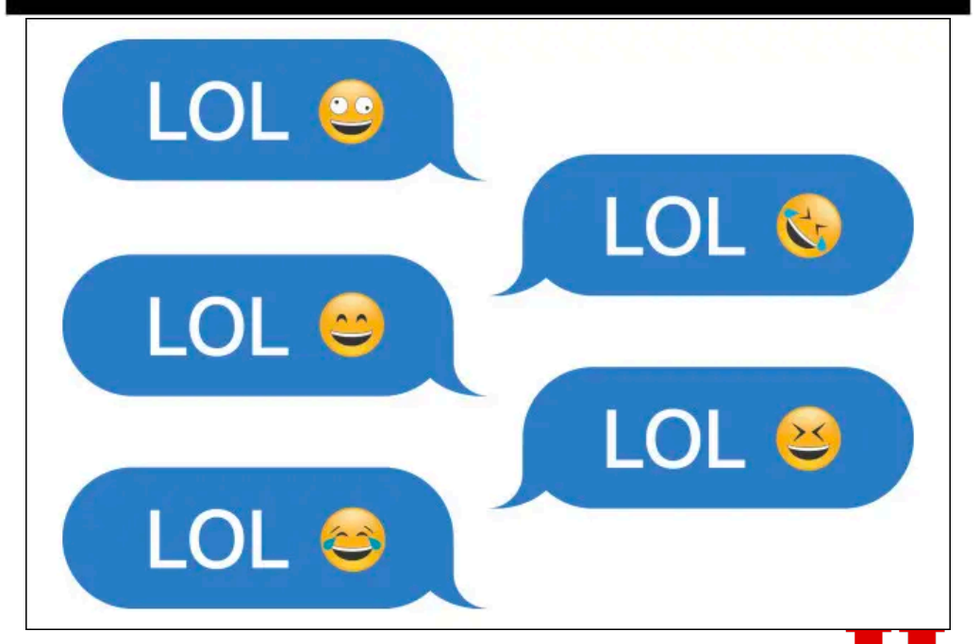
Dynamic

Oneline chatting

SMS/TEXT/CHAT VOCABULARY

- ASAP = as soon as possible
- ATM = at the moment
- B4 = before
- BF = boyfriend
- BRB = be right back
- BTW = by the way
- CU = see you
- FB = facebook
- FML = f**k my life
- FYI = for your information
- GF = girlfriend
- HRU = how are you?
- IMO = in my opinion
- L8R = later
- LOL = laughing out loud
- LMAO = laughing my ass off
- OMG = oh my god
- NP = no problem
- PLS = please
- SBY = somebody
- STH = something
- TBH = to be honest
- THX = thanks
- TMI = too much information
- TTYL = talk to you later
- WKND = weekend
- WTF? = what the f**k
- XOXO = hugs and kisses





Dynamic

cyber lingo

Many, many linguistic phenomena

- Metaphor
  - makes my blood boil, apple of my eye
- Metonymy
  - The White House said today that…
- Idioms
  - Under the weather
- ….

Overview

- What is NLP?
- What are the challenges?
- Techniques in NLP
- The Future

Early NLP systems

- 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

The problem with GOFAI for NLP

Good Old Fashinoned Artifical Intelligence

- Worked **only** for a very small set of examples
- Took several years of human effort to create
- And still did not generalize well

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

Power to the data

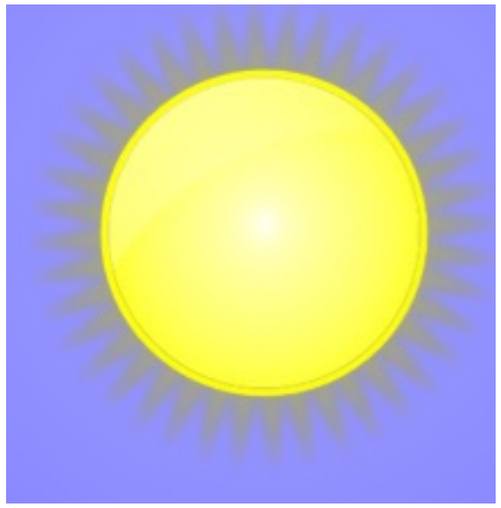
Statistical methods for NLP

Machine learning to the rescue

LITERARIA INSTRUX
ANNO

What do words mean?

- 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

Can we *learn* their meaning from data?

An 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

I would like to eat a _____ of cake

peace or *piece*?

An idea

- Train a binary classifier to make this decision
- use indicators of neighboring words as features

Data + features + learning algorithm = Profit!

Words = vectors

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

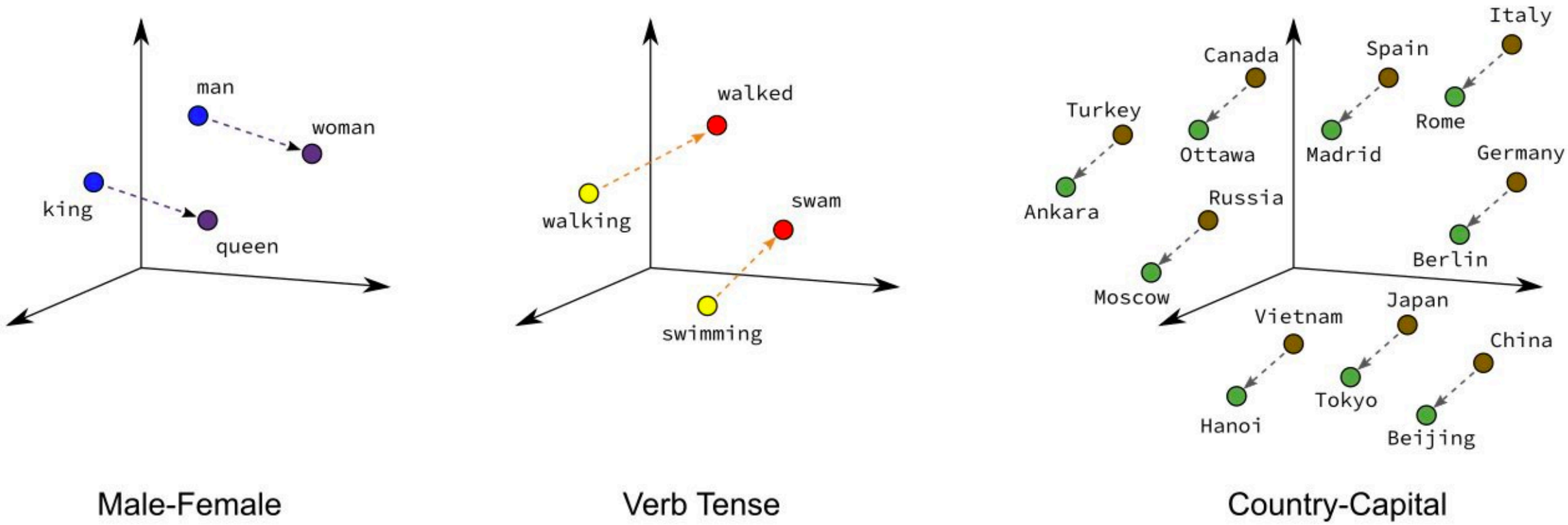
Rome Paris word V

```
Rome   = [1, 0, 0, 0, 0, 0, …, 0]
Paris  = [0, 1, 0, 0, 0, 0, …, 0]
Italy  = [0, 0, 1, 0, 0, 0, …, 0]
France = [0, 0, 0, 1, 0, 0, …, 0]
```

A newer idea: Represent words by much lower dense vectors

Word embeddings

Word embeddings



$$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$$

Part of speech tagging

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

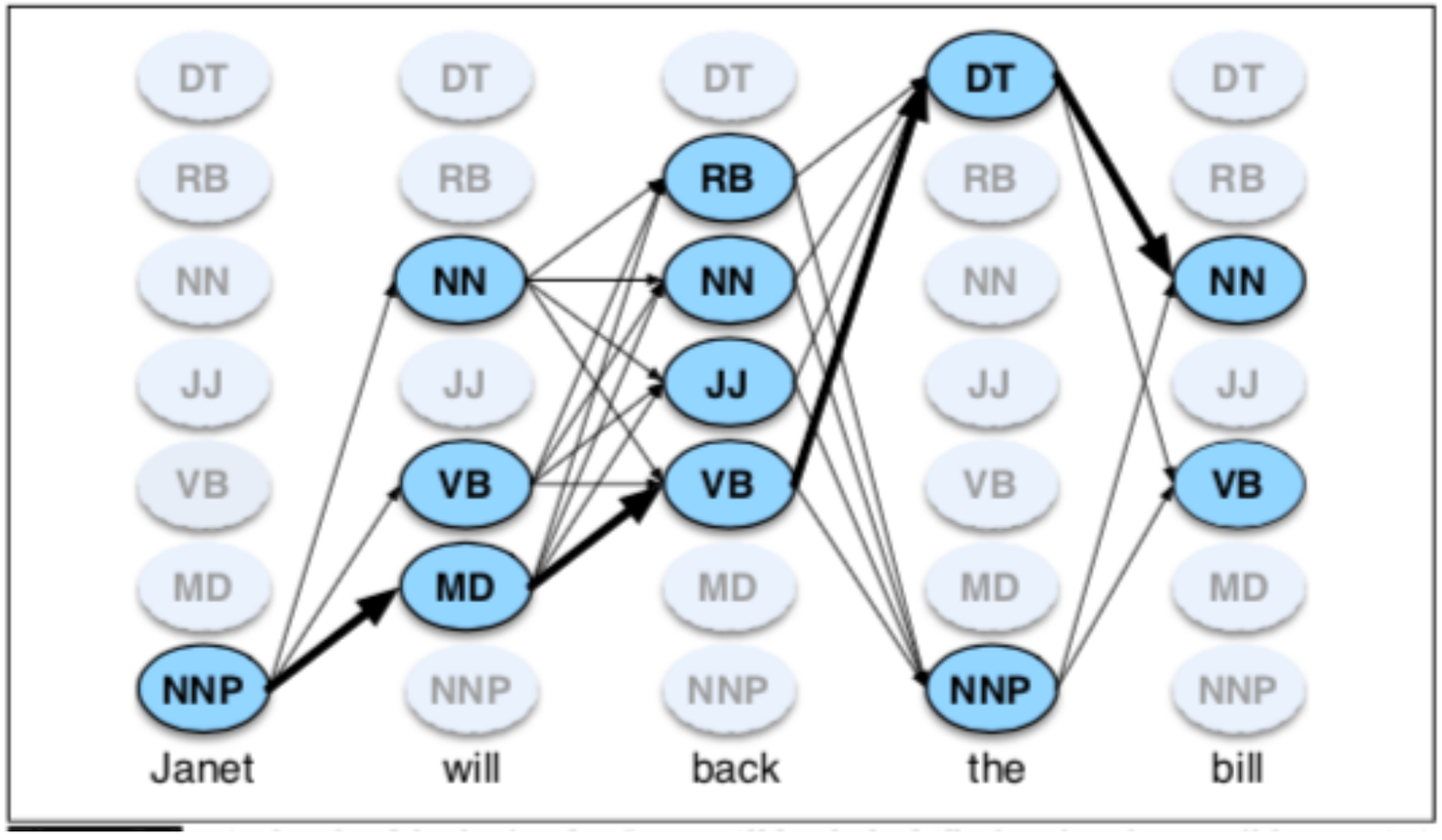
The koala drank the green ink

Determiner, Noun, Verb, Determiner, Adjective, Noun

Part of Speech

The standard solution: Train a sequence model

- Assign a probability to the entire sequence
- Many variants exist



Driven by liguistics

Powered by probabilistic models

Information Extraction

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

One task: Named entity recognition

- Identify all named entities in text and their types

Martin lives in Bananaville and works at Bananas Inc.

Person

Location

Organization

Sentiment Analysis

Predict whether the given text has positive or negative polartity

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

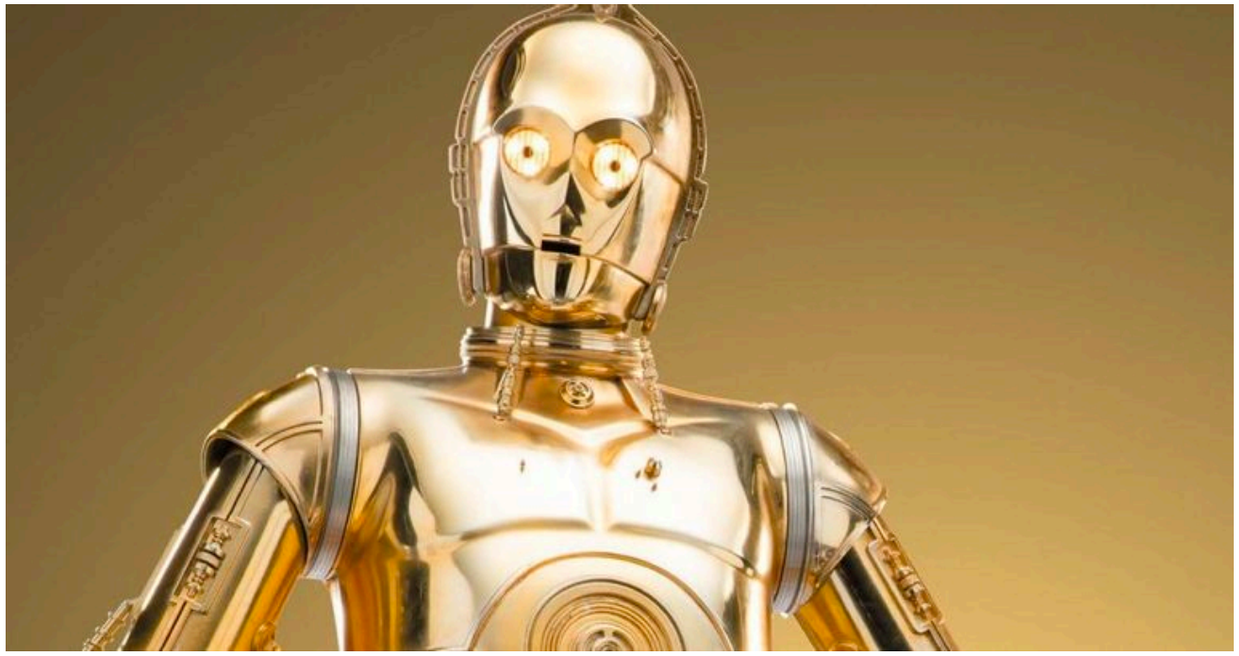
The movie was terrible

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

This movie was long and convoluted, but I liked it anyway

Overview

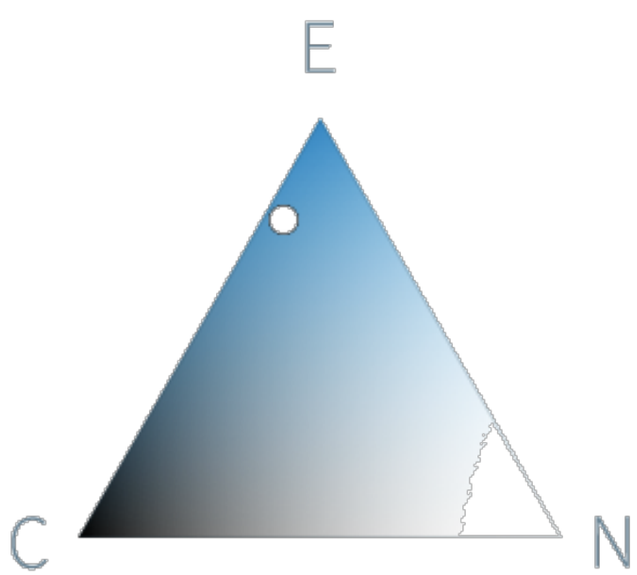
- What is NLP?
- What are the challenges?
- Techniques in NLP
- The Future



Natural language inference

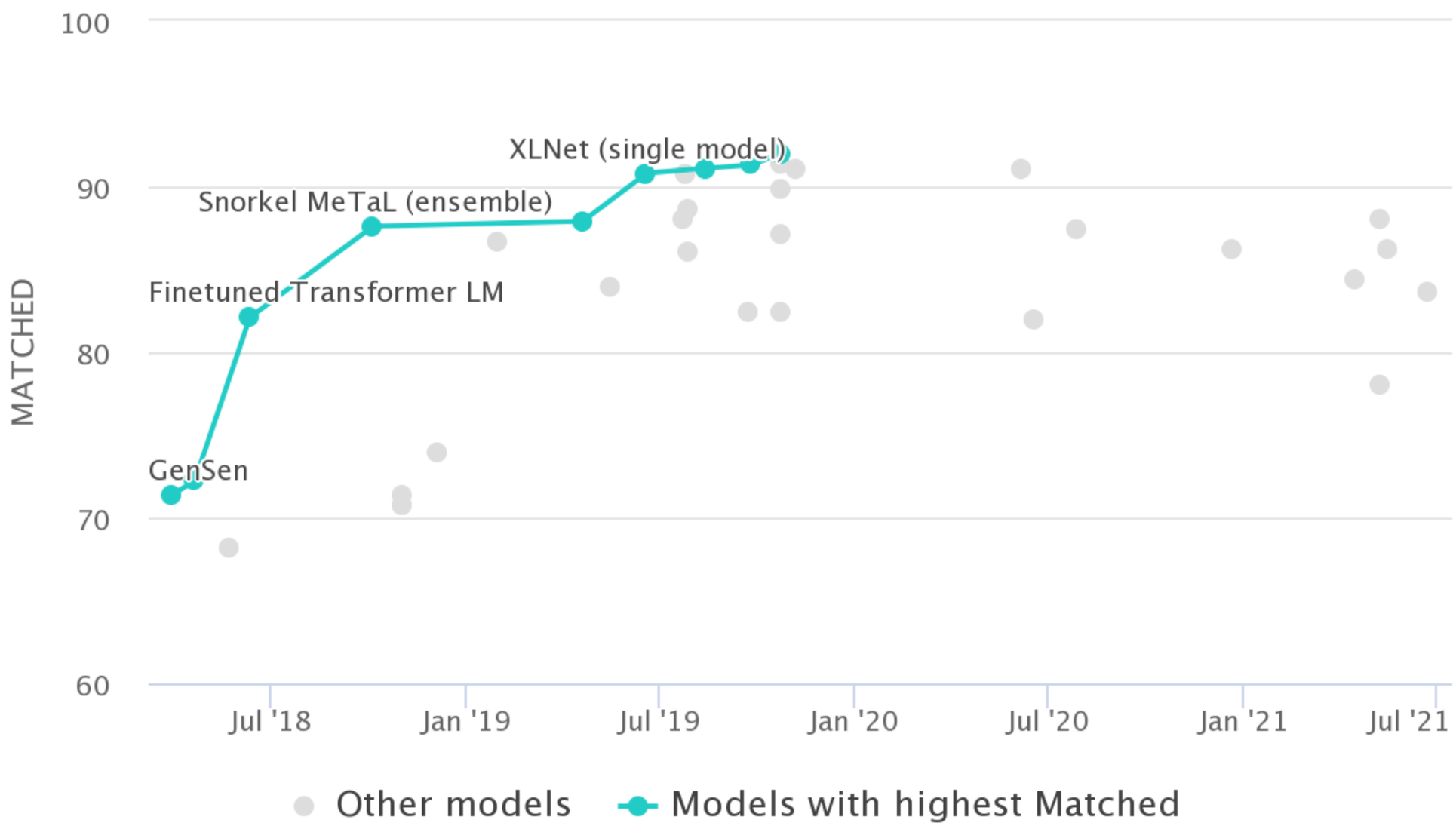
Premise Before it moved to Chicago, aerospace manufacturer Boeing was the largest company in Seattle

Hypothesis Boeing is a Chicago-based aerospace manufacturer



Judgment	Probability
Entailment	75.6%
Contradiction	19.9%
Neutral	4.5%

Natural language inference



MultiNLI

[Lan et al., 2019]

Question Answering

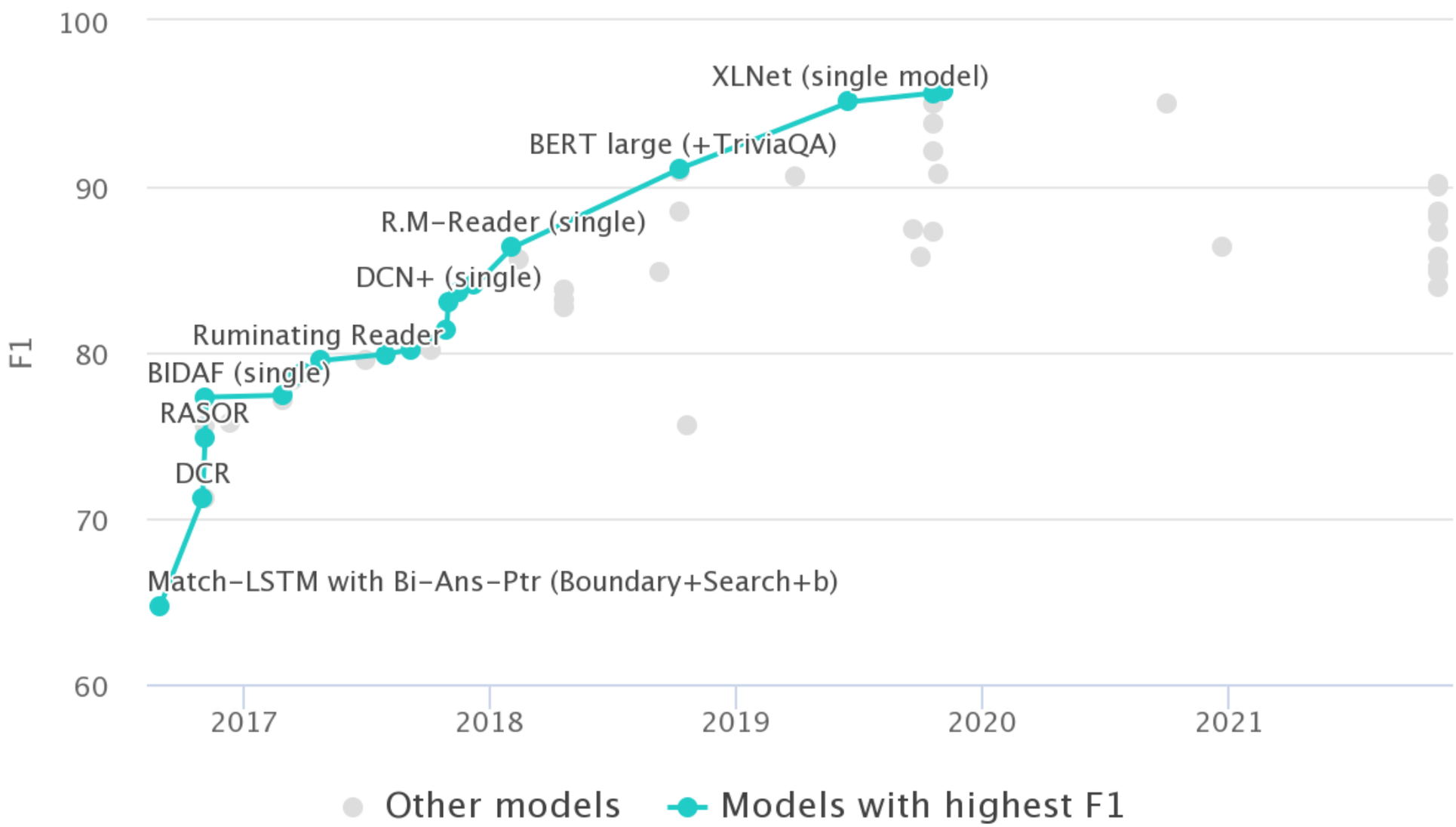
Passage Segment

...The Rankine cycle is sometimes referred to as a practical Carnot cycle...

Question

What is the Rankine cycle sometimes called?

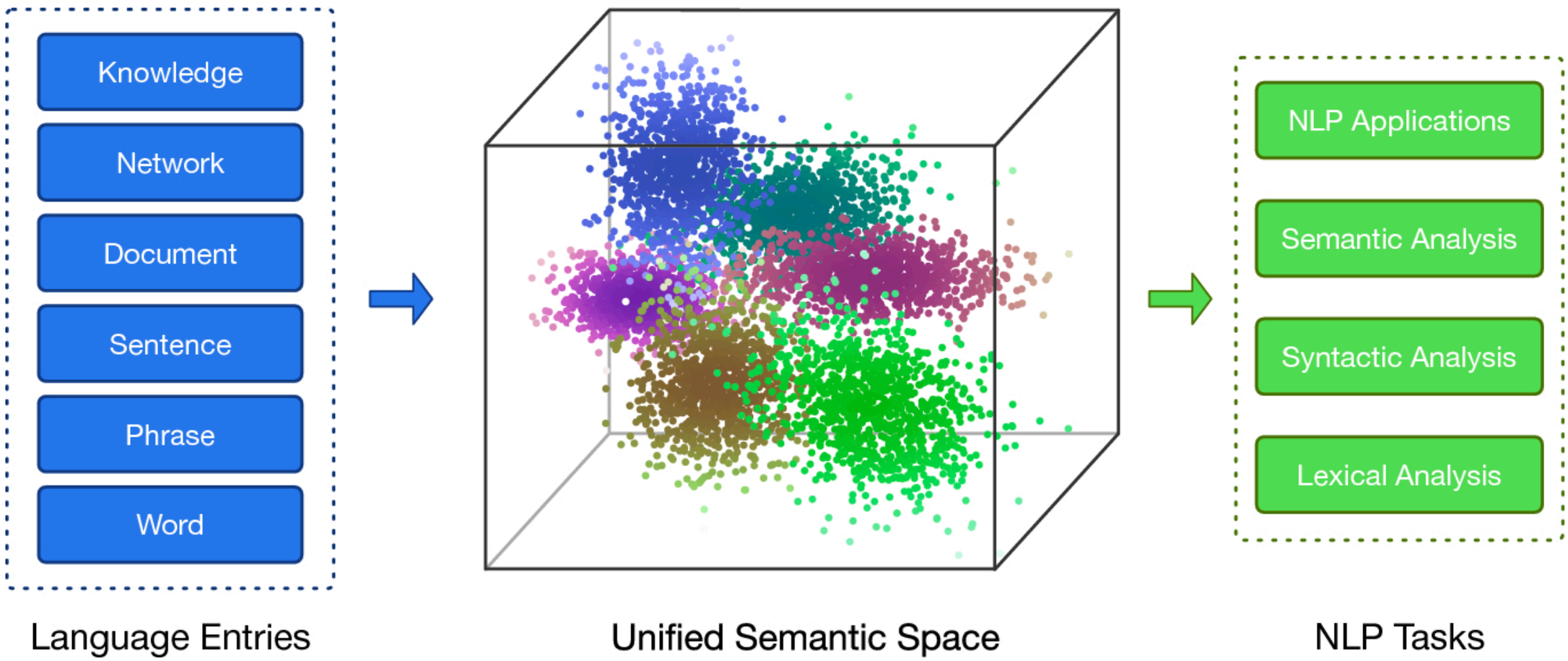
Question Answering



SQuAD1.1 dev

[Yang et al., 2019]

Representations



[Devlin et al., 2019, Peters et al., 2018]

More future directions

- Mental health therapy
- Machine Translation
- Medical diagnose
- …

NLP can transform communication.

But, still a long way to go…

Challenges

Representations

- What is a good representation of language? How can we get it?

Data

- How do we get around the need for annotated data?

Reasoning

- How do we build a system that can reason other than memorizing?

Common knoledge

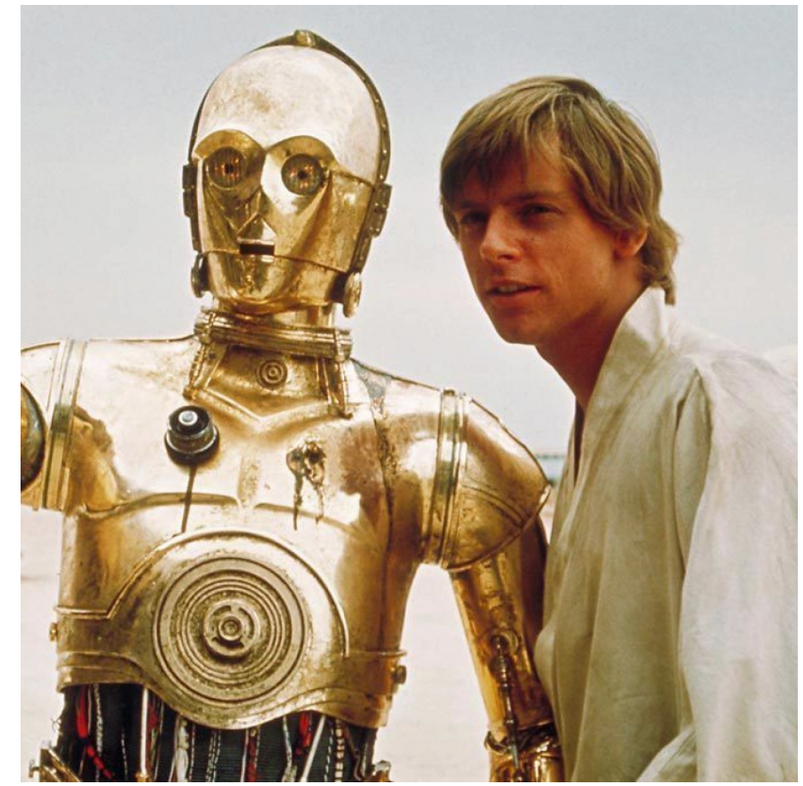
- How do we inform the model the common knowledge?

Bigger Picture

We **want**
programs that read, understand and reason about text

Natural Language Processing

Working in Progress



Lessons from the field

Representation of text matters for building application

Statistical machine learning is here to stay

Domain konwledge helps

Good data helps train a better model