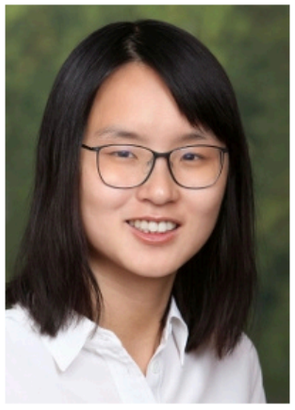
COS 484: Natural Language Processing

Fall 2019

Teaching staff

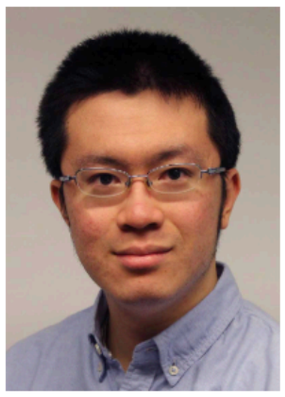
Instructors:



Danqi Chen

Karthik Narasimhan

TAs:



Willie Chang

Pranay Manocha

Runzhe Yang

Logistics

- Course webpage: https://nlp.cs.princeton.edu/cos484/
- Timings: Tuesdays, Thursdays 1:30 - 2:50pm
- Location: CS 104
- Office hours, reading lists, assignment policies on website
- **Sign up for Piazza**
  - **Forum for all class-related queries.**
- **Sign up for Gradescope**
  - **Assignments and grades will be released here**

Course goals

- Gain an understanding of the fundamentals of different sub-fields within NLP
- Understand theoretical concepts and algorithms
- Hands on experience building statistical models for language processing
- Carry out an independent research project

Course structure

- 4 assignments (40%)
  - Released every 2 weeks. Due 11:59pm Monday before lecture.
  - No late submissions. 10% penalty every day of lateness up to 4 days
- 1 in-class mid term (25%): **Thursday, Oct 24, 2019**
- 1 final term project (35%): Teams of 2-3 persons
- Extra credit (5%) for participation (in class/Piazza) and assignment bonus points

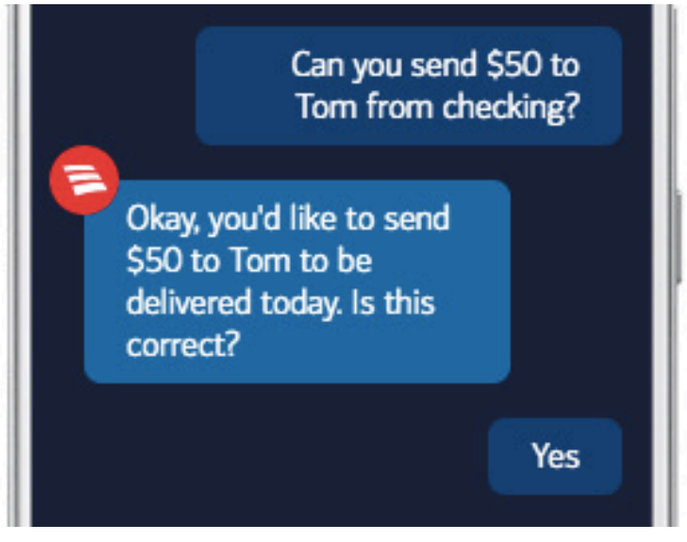
Background

- Required: COS 226 (algorithms and data structures), probability, linear algebra, calculus
- COS 324 recommended

- Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Natural Language Processing

- Making machines understand human language
  - Communication with humans (ex. personal assistants, customer service)



Banking assistant

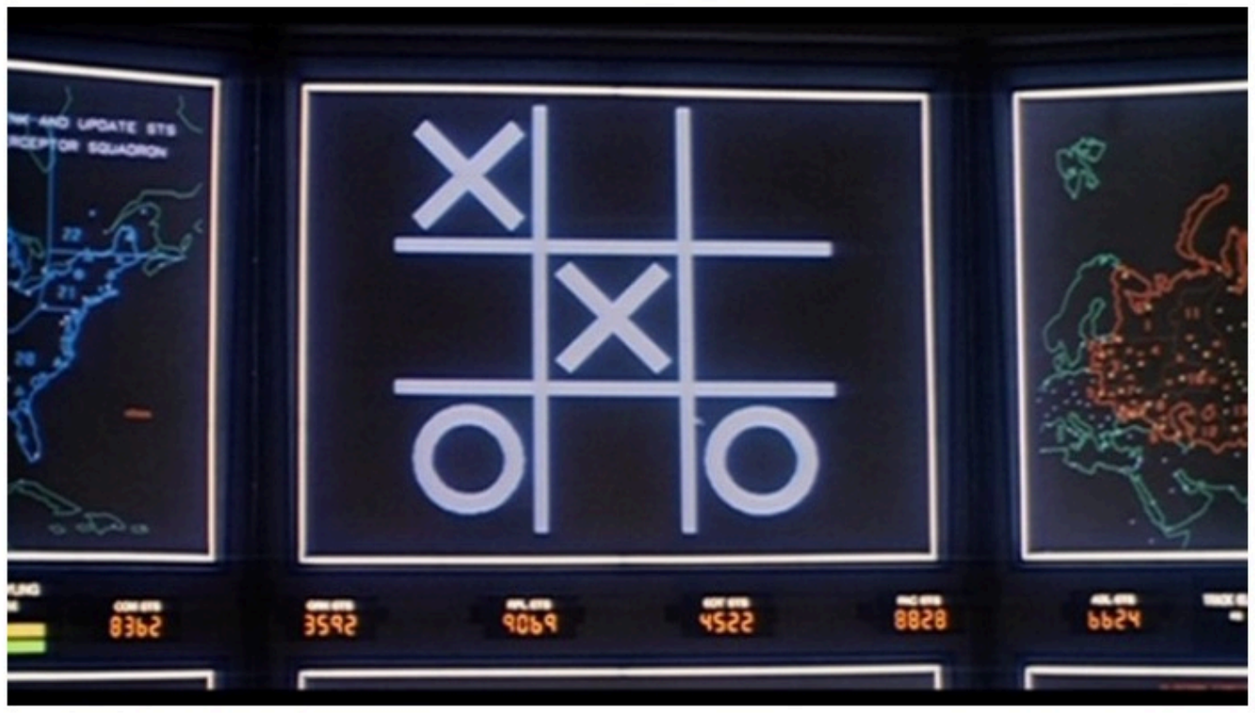
Natural Language Processing

- Making machines understand human language
  - Communication with humans (ex. personal assistants, customer service)
  - Access the wealth of information about the world — crucial for AI systems

Computer learns to play Civilization by reading the instruction manual

By Matthew Rogers on July 14, 2011 at 5:03 pm | 16 Comments

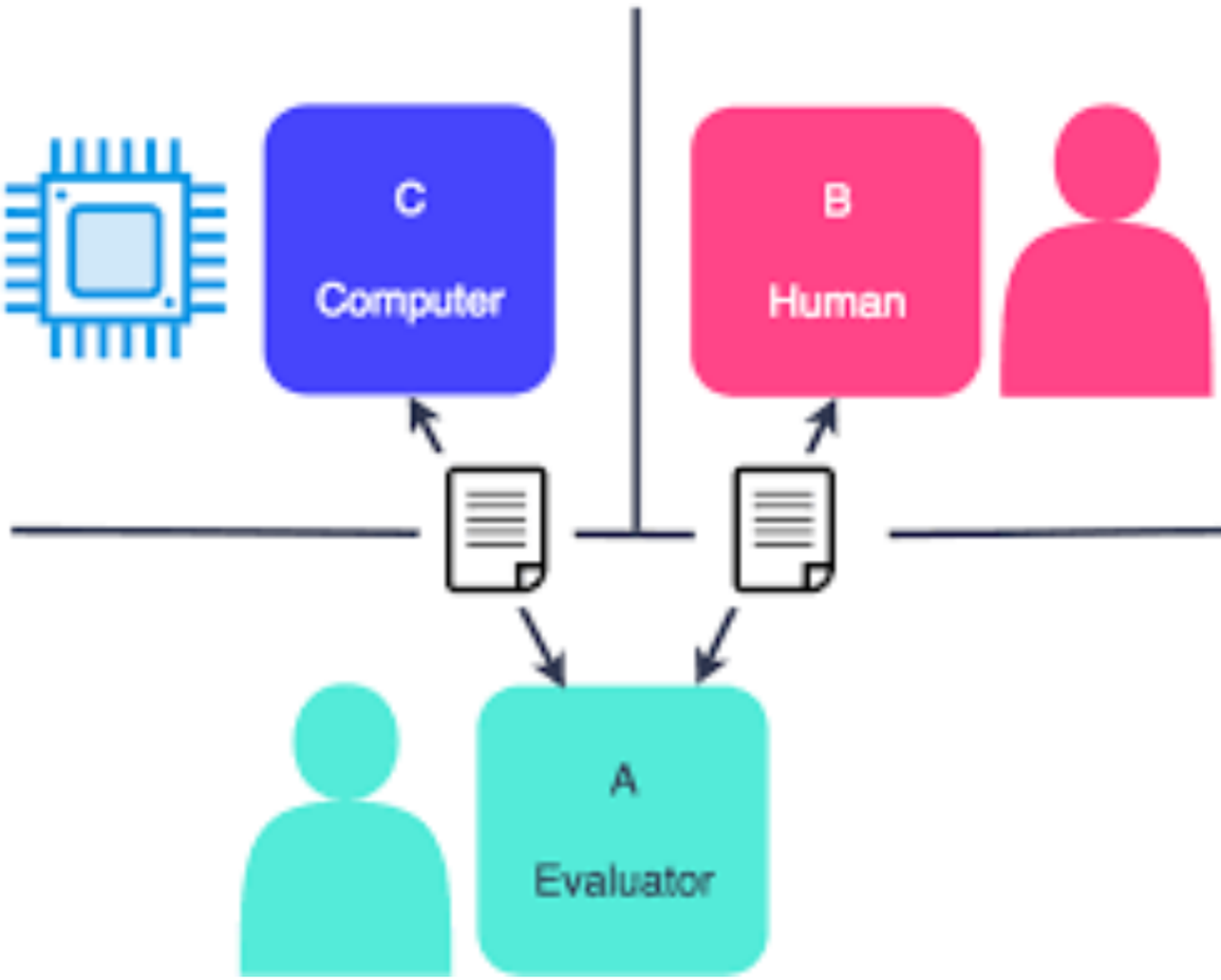
532 SHARES



MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization

Turing Test



Ability to understand and generate language ~ intelligence

Language and thought

Language and Mind

Third Edition

Noam Chomsky

Front Psychol. 2015; 6: 1631.
Published online 2015 Oct 31. doi: 10.3389/fpsyg.2015.01631

Language may indeed influence thought

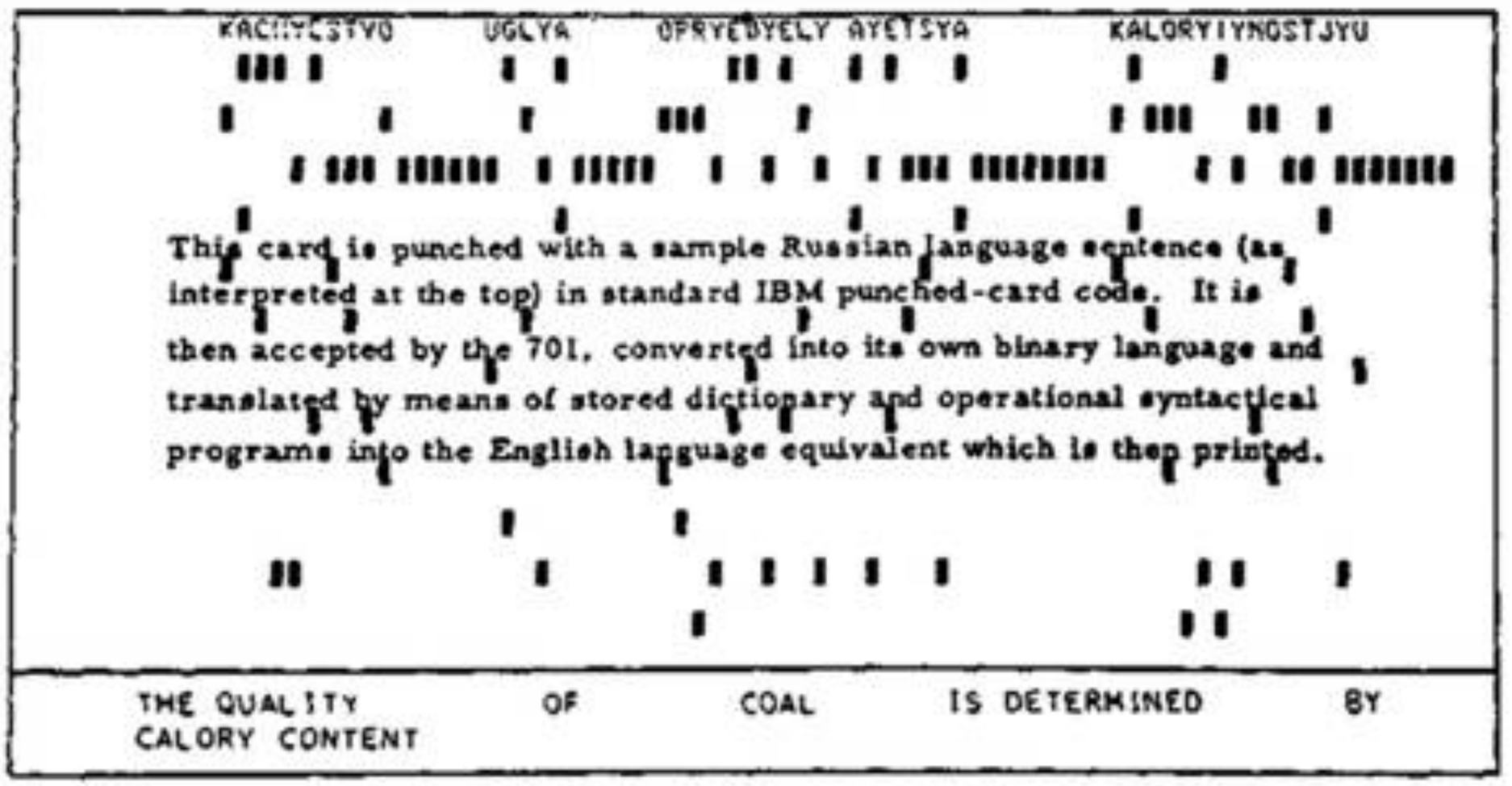
Jordan Zlatev[1,*] and Johan Blomberg[2,3]

Regular Article

Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time ☆

Lera Boroditsky

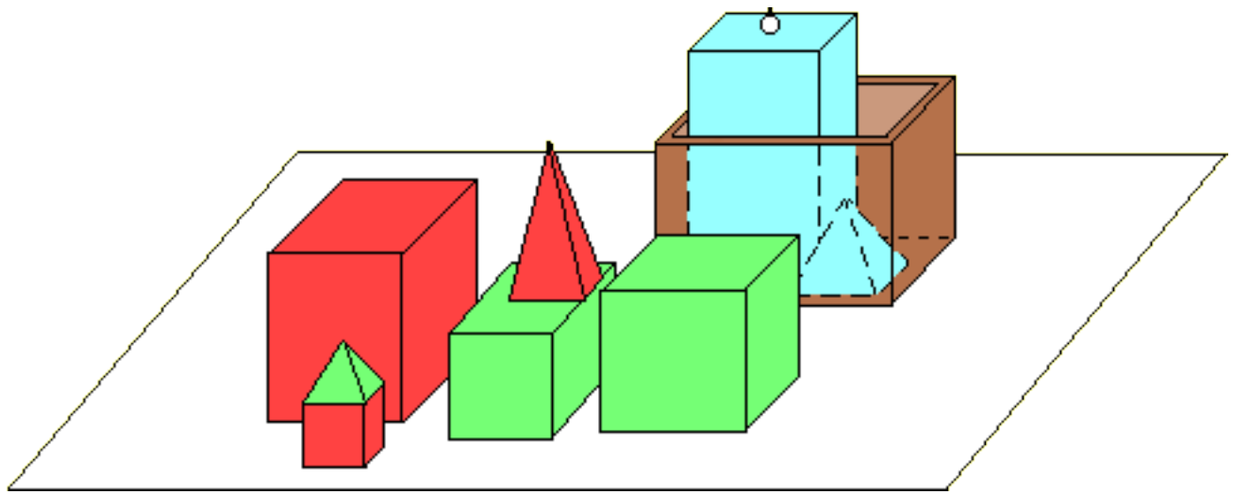
Beginnings



Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown-IBM experiment, 1954

"Within three or five years, machine translation will be a solved problem"



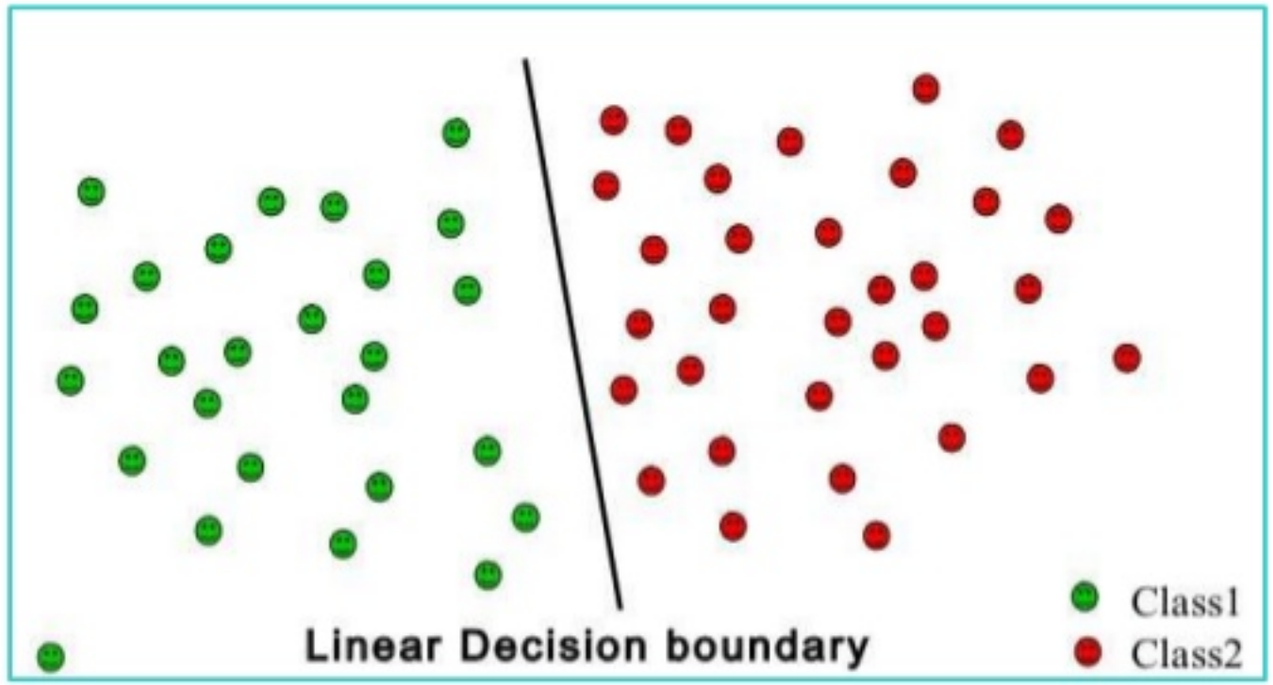
SHRDLU,
1968

> How many red blocks are there?
- THREE OF THEM

> Pick up the red block on top of a green one
OK.

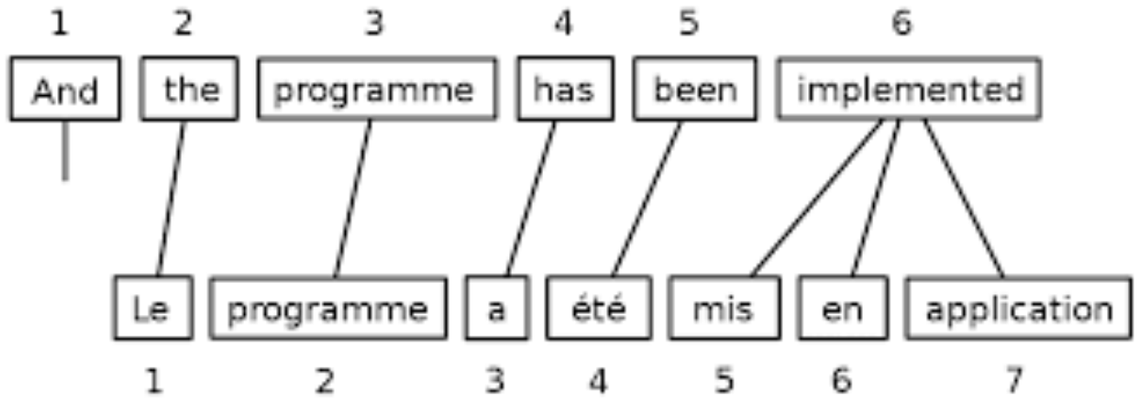
- Rule-based, requiring extensive programming
- Limited domain

Rise of statistical learning

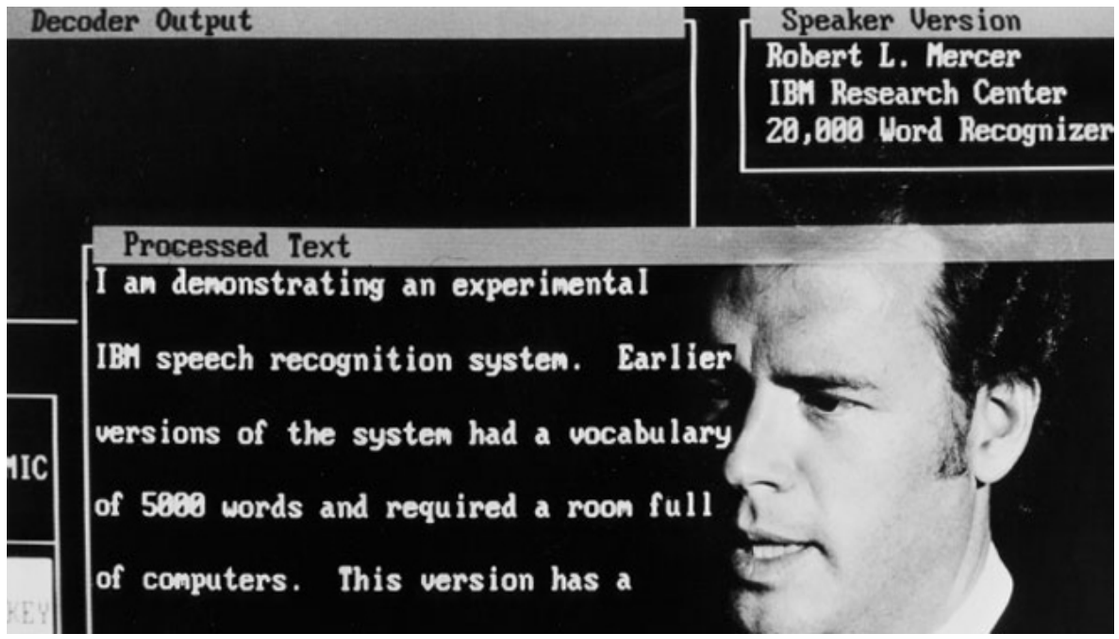


- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

Rise of statistical learning



IBM Models
for translation

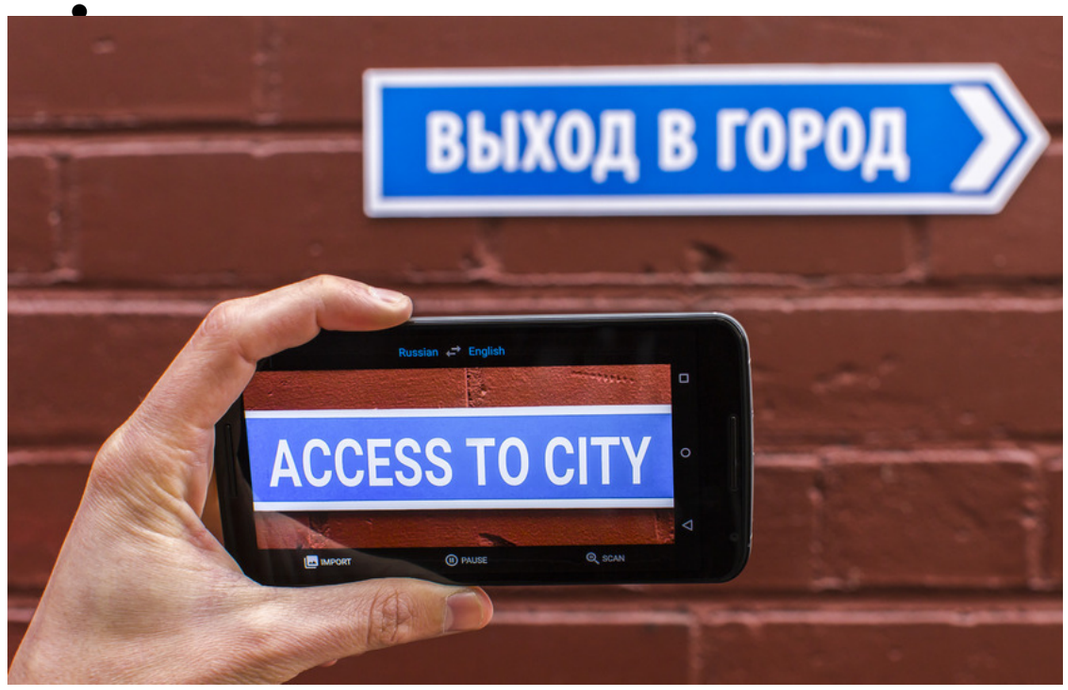


Speech
recognition

Anytime a linguist leaves the group the (speech) recognition rate goes up
- Fred Jelinek

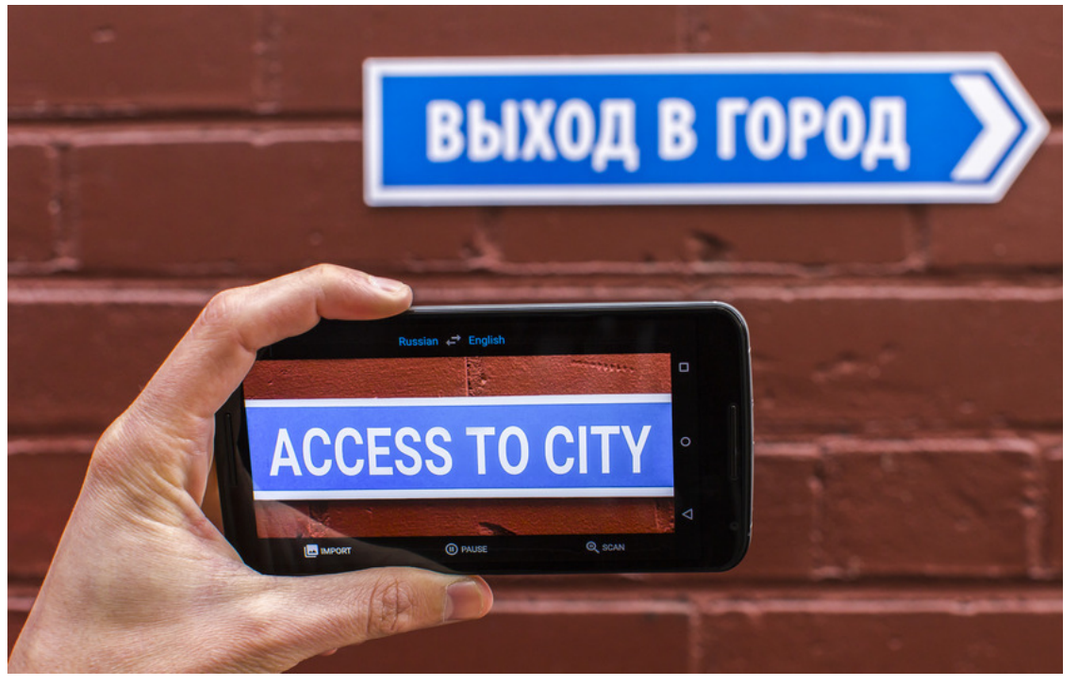
Deep Learning era

- Significant advances in core NLP technologies



Deep Learning era

- Significant advances in core NLP technologies
- Essential ingredient: large-scale supervision, lots of compute
- Reduced manual effort - less/zero feature engineering



36M sentence pairs

Russian: Машинный перевод - это круто!

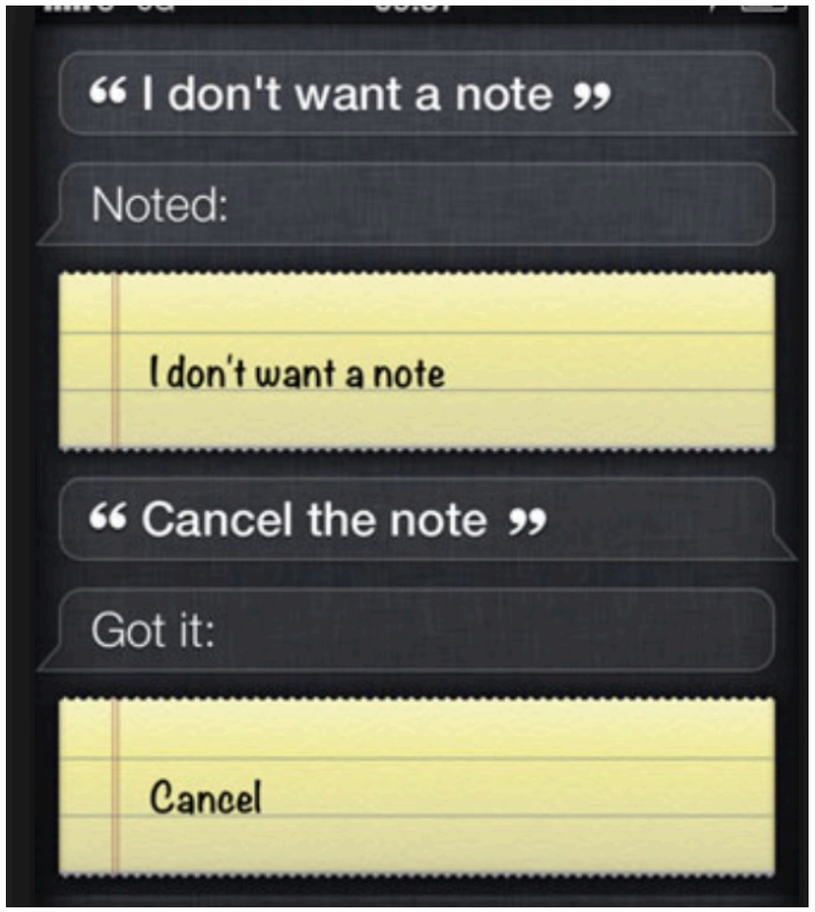
English: Machine translation is cool!

Turing test solved?

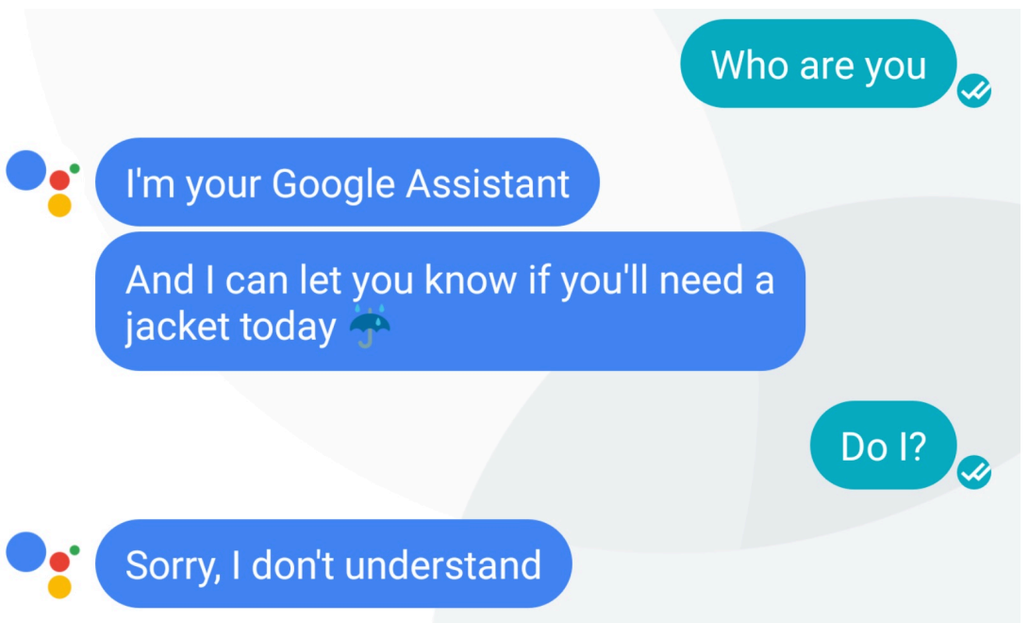
Talking to Google Duplex: Google's human-like phone AI feels revolutionary

Believe the hype—Google's phone-call bot is every bit as impressive as promised.


What can I help you with?
“ Play a good song ”
Sorry, I couldn’t find ‘a good song’ in your music.


“ I don't want a note ”
Noted:
I don't want a note
“ Cancel the note ”
Got it:
Cancel

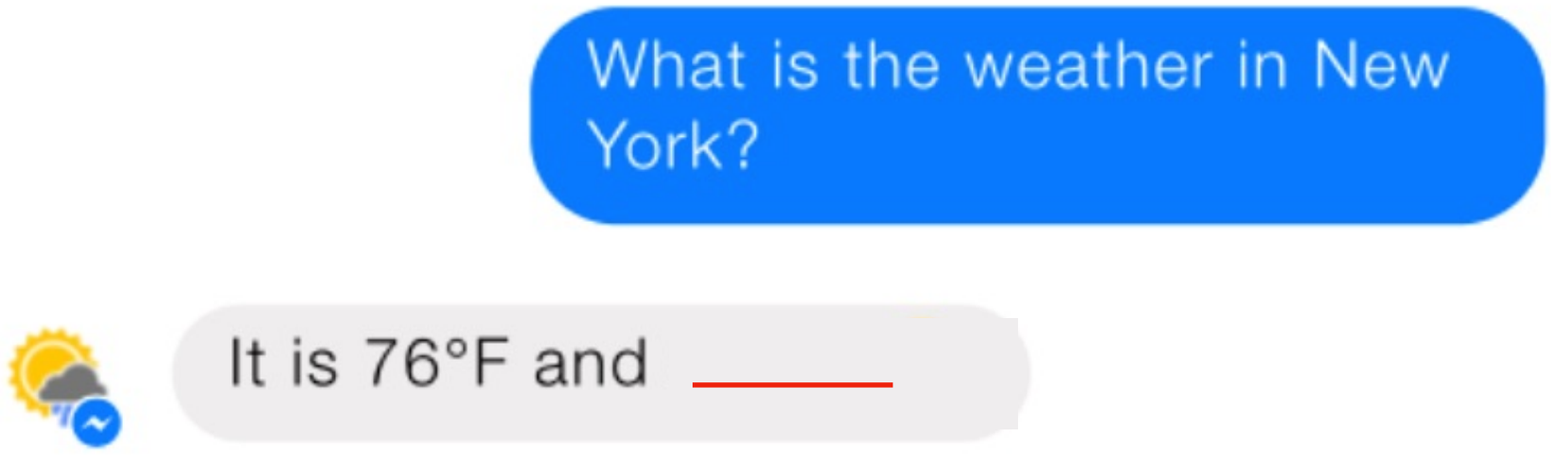

“ Tell my wife I love her ”
OK, I’ll send your message.
To: Sarah Fox
I love her


Who are you
I'm your Google Assistant
And I can let you know if you'll need a jacket today ☔
Do I?
Sorry, I don't understand

… maybe not.

Why is language difficult to understand?

Generating responses



- red ?
- 24.44 C ?
- sunny ?

Language models

- Probabilistic models over word sequences

Sentence "Princeton is in New Jersey"

Chain rule

$$p(w_1, w_2, w_3, \ldots, w_N) =$$
$$p(w_1)\, p(w_2|w_1)\, p(w_3|w_1, w_2) \times \ldots \times p(w_N|w_1, w_2, \ldots w_{N-1})$$

Completion Princeton is in New ?

$$\arg\max_x p(\text{Princeton, is, ...New, } x)$$

Some language humor

Kids make nutritious snacks

Stolen painting found by tree

Miners refuse to work after death

Squad helps dog bite victim

Killer sentenced to die for second time in 10 years

Lack of brains hinders research

Real newspaper headlines!

Lexical ambiguity

The fisherman went to the *bank*.

bank[1]

/baNGk/

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

One word can mean several different things

Lexical ambiguity

The fisherman went to the *bank*. He deposited some money.

bank[1]

/baNGk/

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

Word sense disambiguation

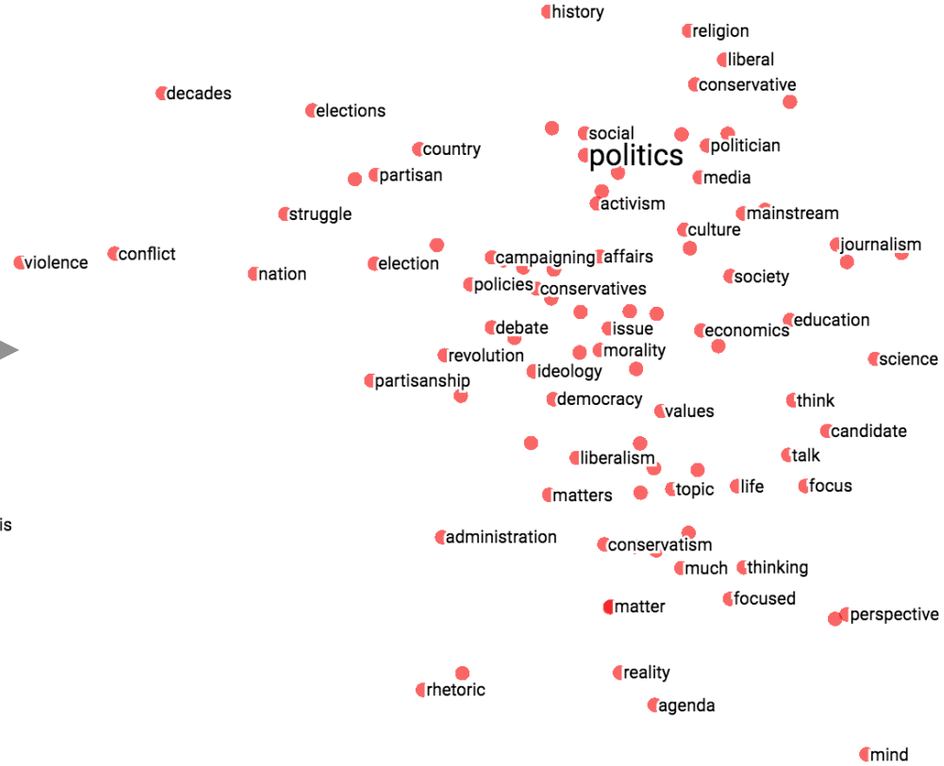
Lexical variations

ACCORDING TO THE THESAURUS, "THEY'RE HUMID, PREPOSSESSING HOMOSAPIENS WITH FULL SIZED AORTIC PUMPS" MEANS "THEY'RE WARM, NICE PEOPLE WITH BIG HEARTS."

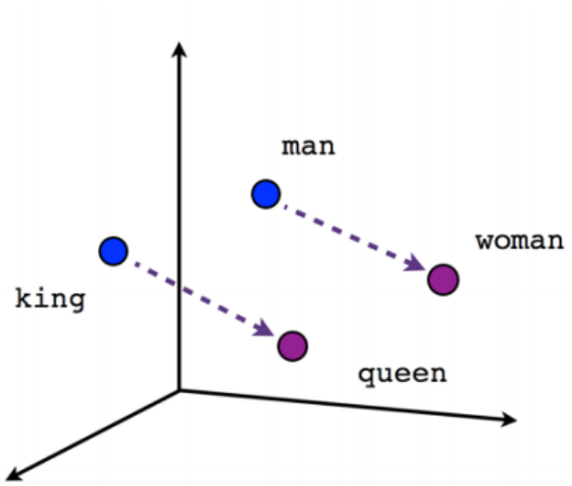
Several words can mean the same thing!

Distributed representations

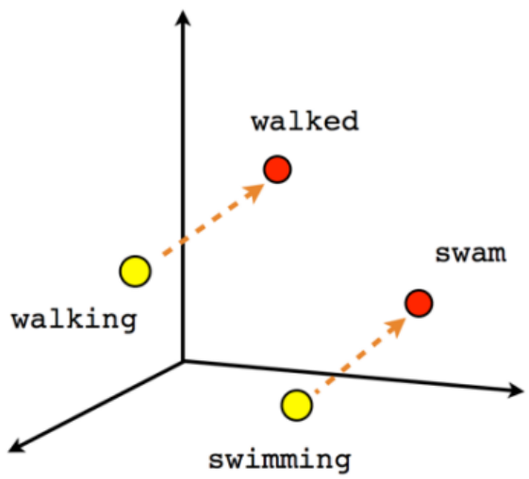
Project words onto a continuous vector space



Similar words closer to each other



Male-Female



Verb tense

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) = v(\text{queen})$$

Comprehending word sequences

- My brother went to the park near my sister’s house
- Park my went house near to sister’s my brother the
- “My brother went park near sister’s house”?
- The old man the boat

Comprehending word sequences

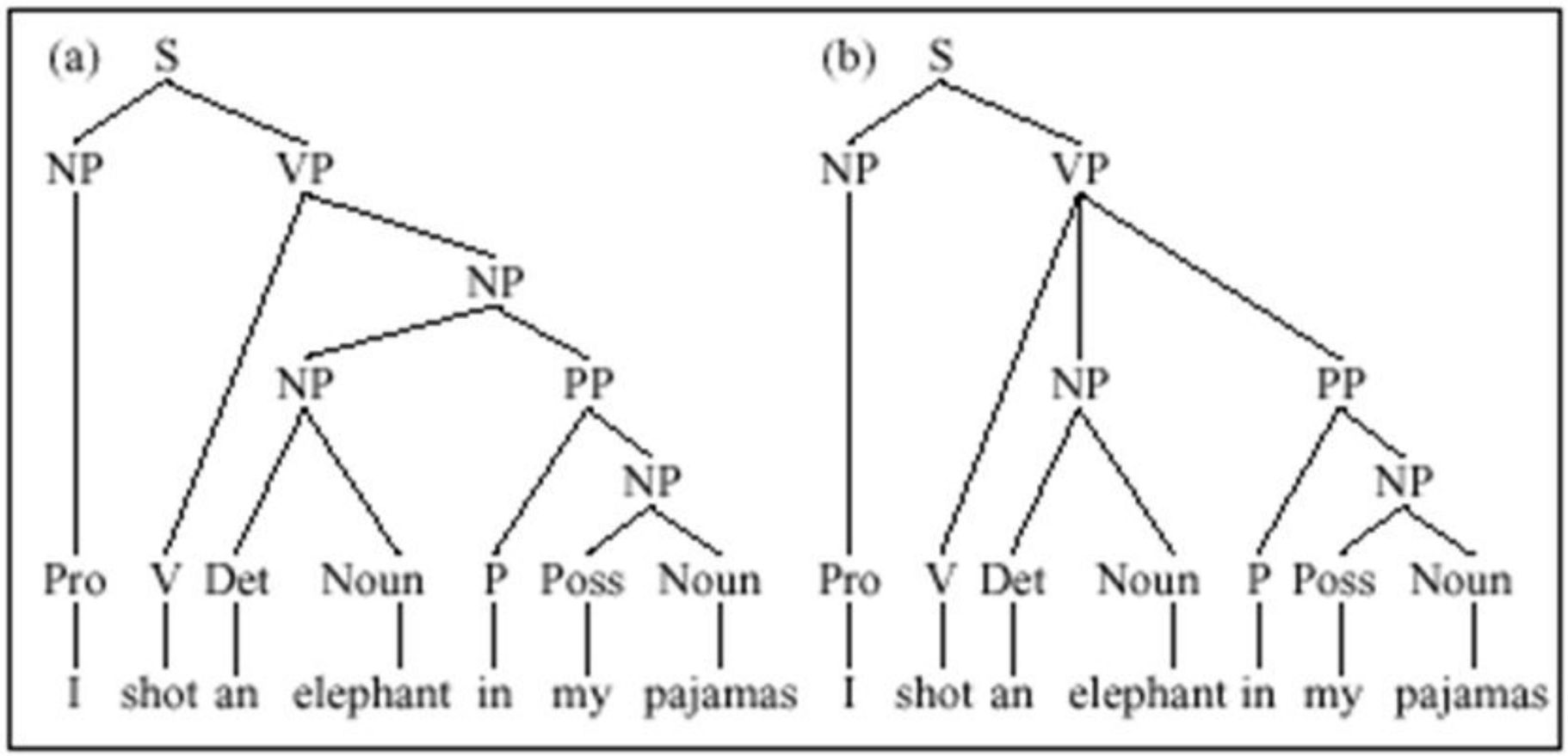
- My brother went to the park near my sister's house
- Park my went house near to sister's my brother the
- "My brother went park near sister's house"?
- The old **man** the boat Garden Path sentence

Structure in language

- Implicit structure in all languages
  - Crucial to understanding
- Coarse-to-fine levels (recursive)
- What are some good data structures to represent this?

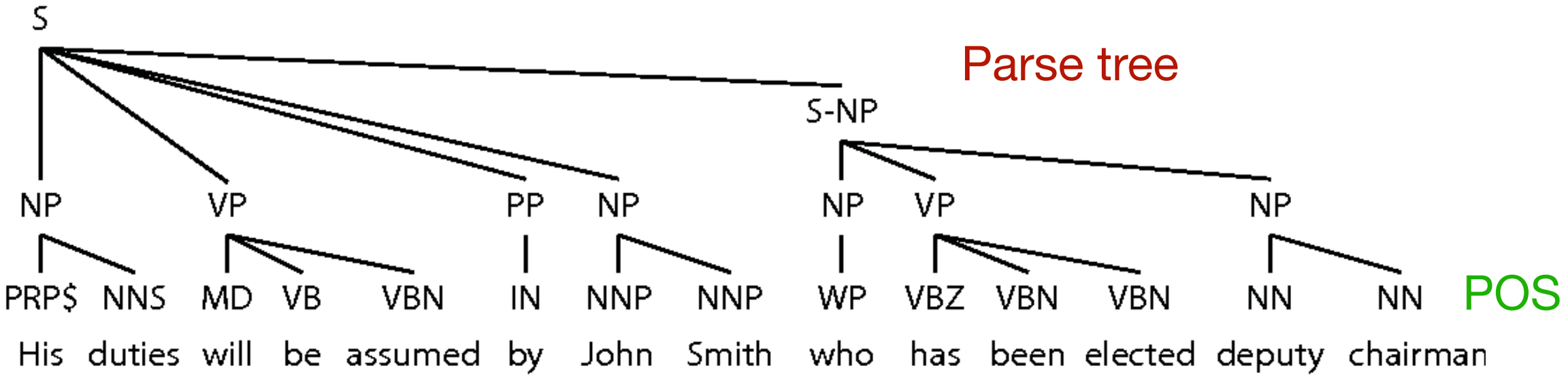
Syntactic ambiguity

I shot an elephant in my pajamas



Human language is full of such examples!

Syntactic parsing



Parse tree

POS

Sentence

Penn Treebank (PTB) : ~40k sentences, 950k words

Online tools: http://nlp.stanford.edu:8080/corenlp/

Discourse ambiguity

Alice invited Maya for dinner but she cooked her own food

she = Alice or Maya?

… and brought it with her. Maya

… and ordered a pizza for her guest. Alice

Anaphora resolution

Semantics

Tell my wife I love her

Action: Text

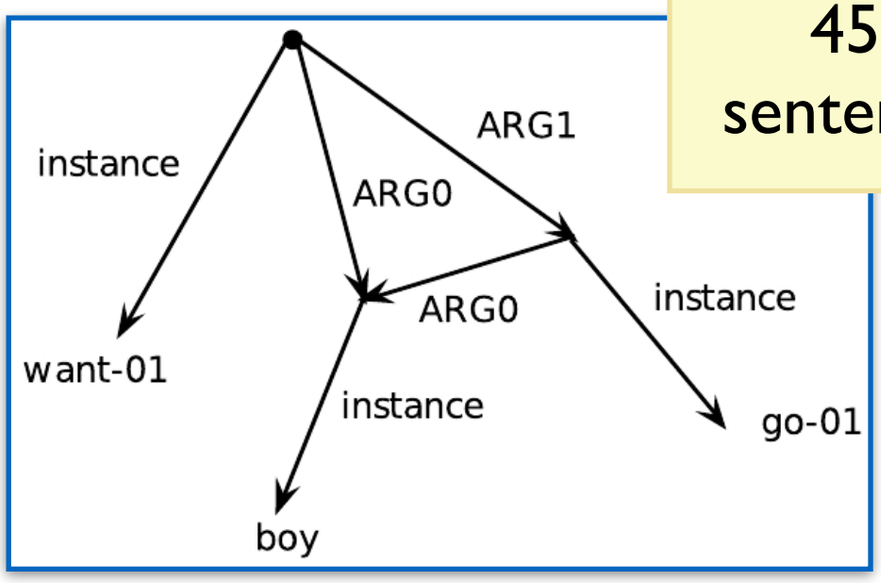
Arg1: Sarah Fox

Arg2: I love you

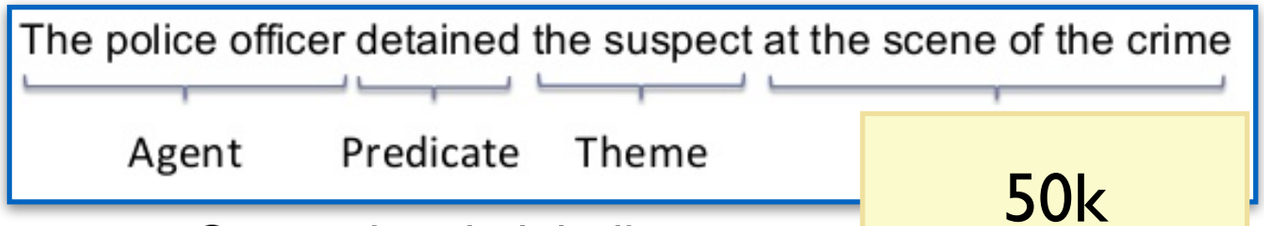
Semantic Parsing

Trained on manually annotated representations

45k sentences



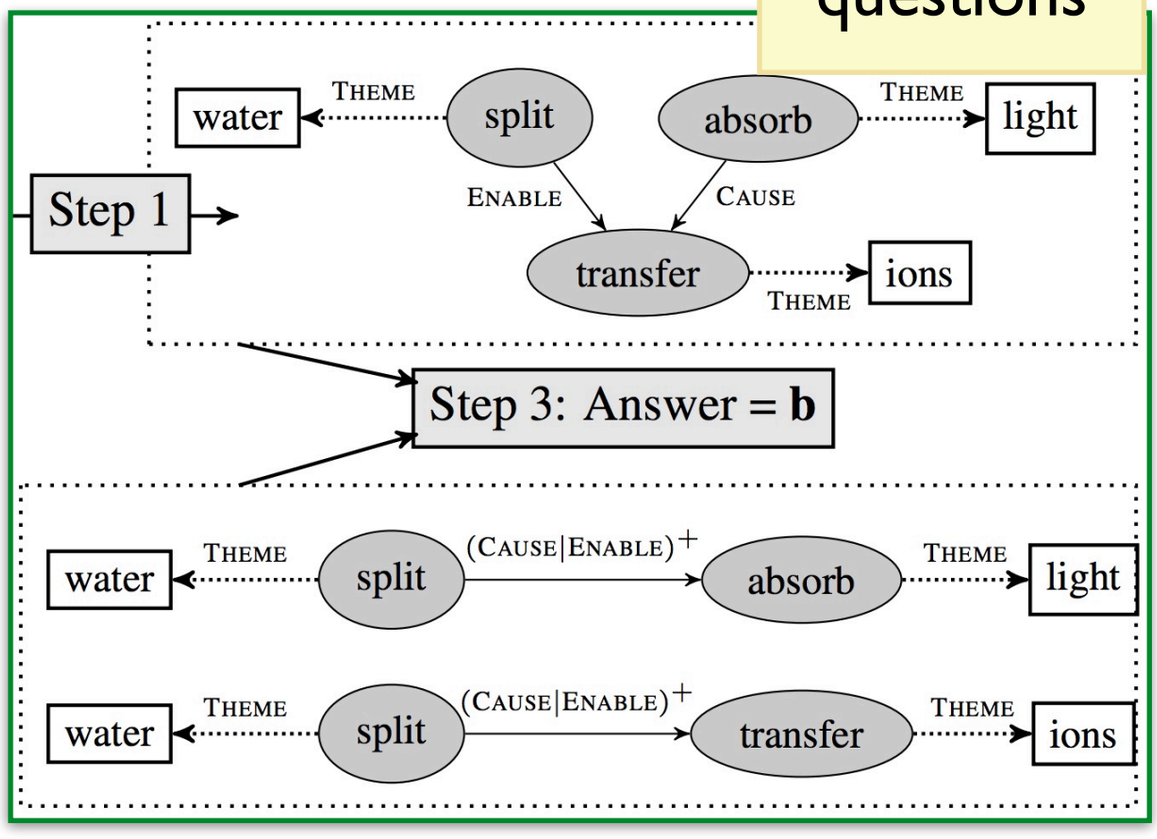
Abstract meaning representation



Semantic role labeling

50k arguments

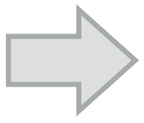
585 questions



(Berant et al., 2014)

Information Extraction

The Massachusetts Institute of Technology (MIT) is a private research university in Cambridge, Massachusetts, often cited as one of the world's most prestigious universities.
Founded in 1861 in response to the increasing industrialization of the United States, …



***City**: Cambridge, MA*

***Founded**: 1861*

***Mascot:** Tim the Beaver*

…

Article

Database

Information Extraction: State of the Art

Dependence on large training sets

ACE: 300K words

Freebase: 24M relations

Not available for many domains (ex. medicine, crime)

Challenging task: even large corpora do not guarantee high performance

~ 75% F1 on relation extraction (ACE)

~ 58% F1 on event extraction (ACE)

Machine Translation

巴拿马首任驻华大使专访：与台湾断交之后

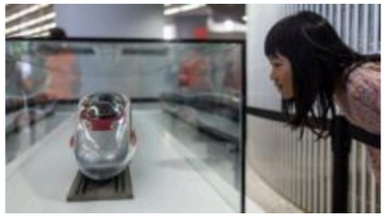
他说，不认同中国“买走”台湾邦交国的说法，与中国建立关系对巴拿马有利，不担心影响与美国关系。

1小时前

巴拿马外交转向周年 中美大国博弈内幕解密

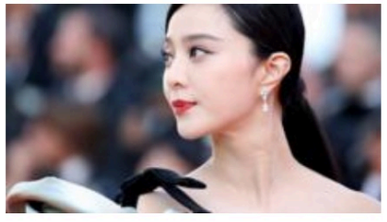
尼加拉瓜运河成谜：人走楼空 “不再提及”

触发萨尔瓦多与台湾断交的港口令美国担忧



观点：高铁“一地两检”——中国的强势 港人的无力

江沂：两地价值观和制度差异巨大，冲突无可避免。香港没有反对的权利，也没有反对的能力。



范冰冰消失百日后 中国娱乐业的寒蝉效应

日前发布的《中国影视明星社会责任研究报告》中，最高分徐峥为78分，最低分范冰冰为0分。

2小时前

中蒙参加俄罗斯军演 “中俄靠拢论”再吸眼球

2018年 9月 11日

日本人脚踹慰安妇铜像引爆台湾人抗议

2018年 9月 11日

特别推荐

人民币贬值、楼价高涨：消费降级中产失去的优质生活

台湾民间发起东京奥运“正名”公投的意义

RESEARCH › PUBLICATIONS ›

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

(Wu et al., 2016)

Machine Translation

BBC Sign in Option (English) 检索

NEWS | 中文 Traditional

Homepage | International | Cross-strait | United Kingdom | comment | Technology | Finance | Picture series | Audio material

Interview with Panama's first ambassador to China: After breaking diplomatic relations with Taiwan

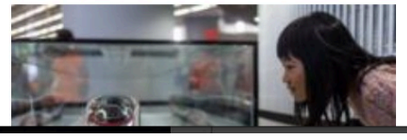
He said that he does not agree with China's saying that "buy" Taiwan's diplomatic relations with China. Establishing relations with China is beneficial to Panama and does not worry about affecting relations with the United States.

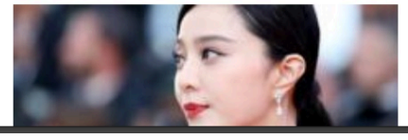
1 hour ago

Panamanian Diplomacy Turns to Anniversary

The Nicaragua Canal is a mystery: people go to the floor and "no longer mention"

The port that triggered the break of El Salvador and Taiwan has worried the United States





China and Mongolia participate in the Russian military

Special recommendation

Renminbi depreciation, property prices are rising: consumption downgrades the loss of quality life in the middle class

ADVERTISING

The significance of the Taiwanese people's "referred to" referendum

Machine comprehension

Amazon_rainforest

The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

Which name is also used to describe the Amazon rainforest in English?
Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, | Amazonia or the Amazon Jungle | Amazonia
Prediction: Amazonia

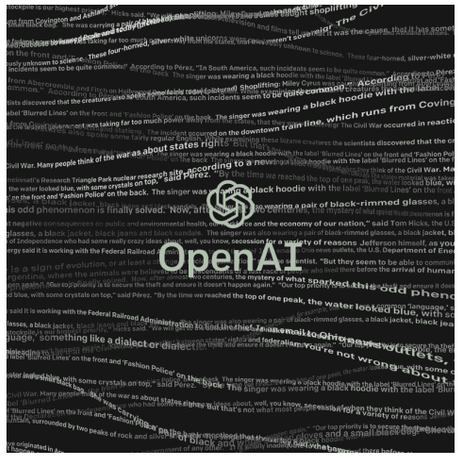
How many square kilometers of rainforest is covered in the basin?
Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. | 5,500,000 | 5,500,000
Prediction: 5,500,000

How many nations control this region in total?
Ground Truth Answers: This region includes territory belonging to nine nations. | nine | nine
Prediction: nine

How many nations contain "Amazonas" in their names?
Ground Truth Answers: States or departments in four nations contain "Amazonas" in their names. | four | four
Prediction: four

What percentage does the Amazon represents in rainforests on the planet?

Language generation



Better Language Models and Their Implications

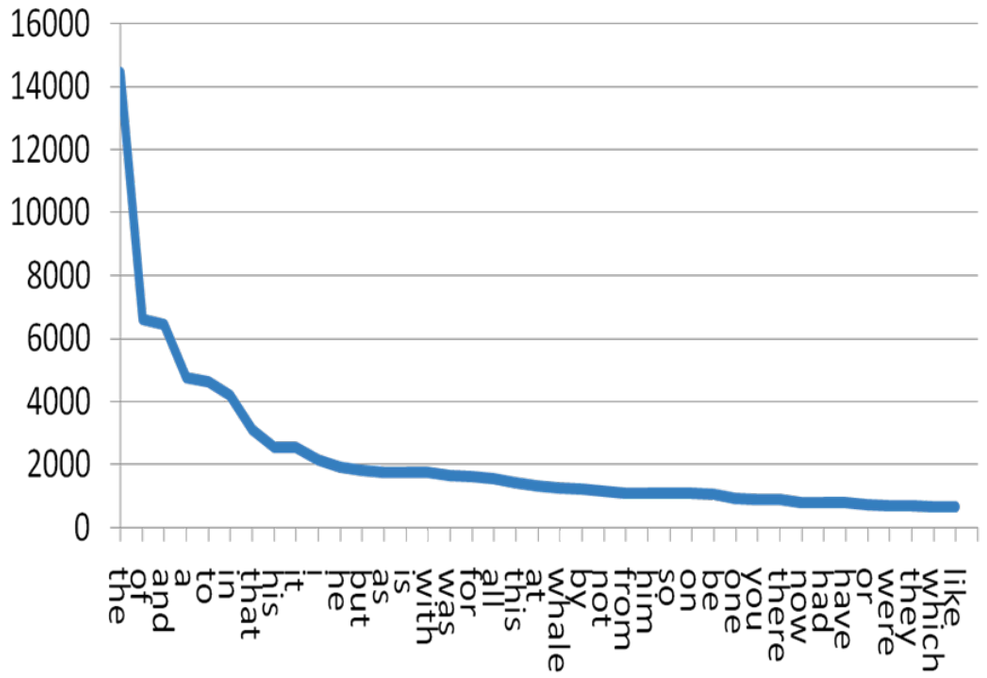
We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

With the start of the new academic year, Princeton has an opportunity to help provide a new generation of women with a diverse set of academic resources for higher education.
We are offering the resources of the Princeton-McGill program specifically to women with undergraduate degrees who would like to enhance their academic experience. Princeton-McGill offers a comprehensive suite of services for women and their families including a variety of graduate programs, support programs, and the opportunity to serve as leaders in their communities with a wide variety of programs, activities and services. For the upcoming fall, Princeton-McGill will also offer its Women's Center , which is located in a renovated women's dorm.
At Princeton, we are working with the Princeton-McGill community to develop a suite of programs that are designed to give new and returning students a strong foundation for a successful, rewarding graduate career. The Women's Center , the Princeton-McGill Women's Center provides a range of supports to address the specific needs of female doctoral degree graduates. Programs are tailored to meet the unique needs of women under the age of 28, women and families

https://talktotransformer.com/

Challenges in modern NLP

- *Scale*: Large number of phenomena
- *Sparsity*: Text data is often heavy-tailed



Challenges in modern NLP

- *Bias*: Models learn biases in available data

man

woman

- *Context:* Knowledge bases, perception, interaction

Outline

Words

- Language models
- Text classification
- Word embeddings

Sequences and structures

- HMMs, recurrent neural networks
- Syntactic Parsing
- Machine Translation

Applications

- Coreference resolution
- Question Answering
- Multimodal NLP