COS 484/584

(Advanced) Natural Language Processing

Spring 2021

Instructors

Danqi Chen

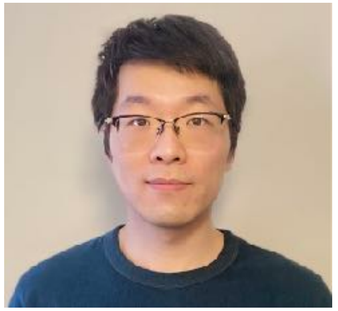
Karthik Narasimhan

Teaching Assistants

Ameet Deshpande

Chris Sciavolino

Kaiyu Yang



Mingzhe Wang

Shunyu Yao

Zexuan Zhong

Logistics (484+584)

- Course webpage: https://nlp.cs.princeton.edu/cos484/
  - Contains all details including office hours, reading lists, assignment policies
- Joint lectures (484+584): **Mondays, Wednesdays 1:30 - 2:50pm**
  (on Zoom, will be recorded)
- TAs will hold precepts for 484 students (please fill out Google Form for timings)
- Canvas for all course related announcements (no Blackboard)
- **Sign up for Ed Discussion**
  - **Forum for all class-related queries.**

Logistics (484+584)

- Assignments:
  - Due on Mondays before class (1:30pm)
  - 96 free late hours (~4 days) in total over all assignments
  - After this, 10% penalty for every day lateness (max 3 days)
  - 5 assignments in total
  - A0 (warm up) released today, due in one week (Feb 8)
- **Sign up for Gradescope**
  - **Assignments and grades will be released here**

Logistics (584 only)

- Precepts: Fridays, 11-11:50am (different Zoom link)
  - Participation counts for 10% of overall grade
  - Readings and pre-lecture questions will be released on Monday
- **Sign up for Perusall**
  - **We will use these for Friday precept readings and provide links to pre-lecture questions here**

Grading

COS 484

- Assignments (40%)
- Mid term (25%)
- Final project (35%): Teams of 3

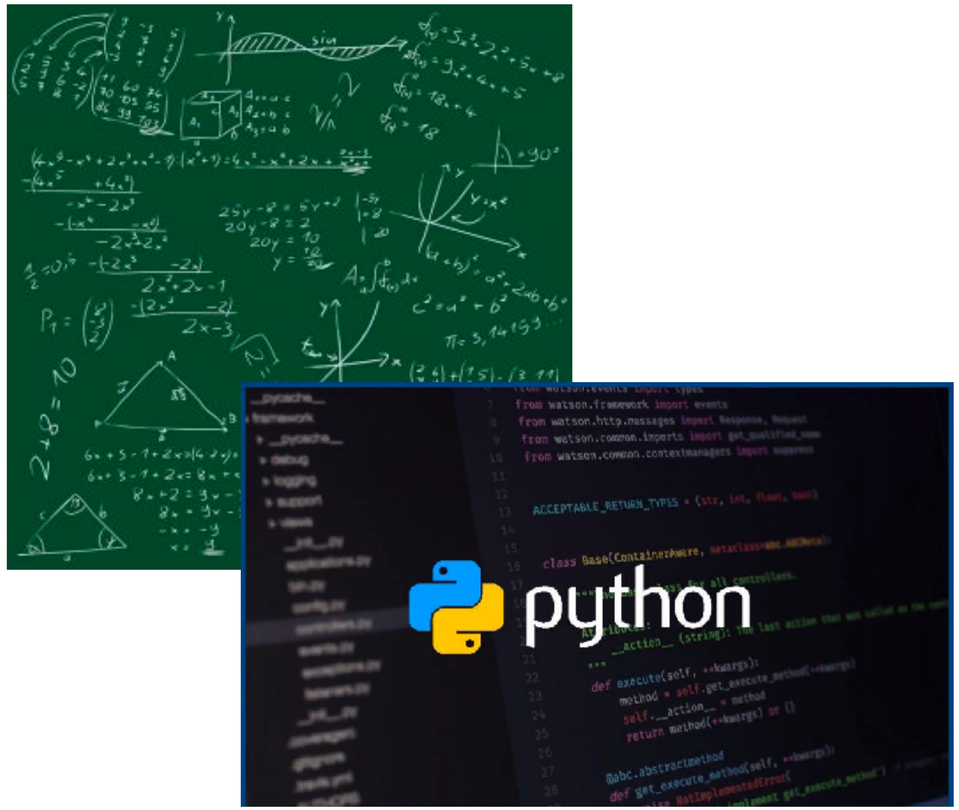
COS 584

- Assignments (30%)
- Mid term (25%)
- Precept participation (10%)
- Final project (35%): Teams of 1-2

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 at the end

Background

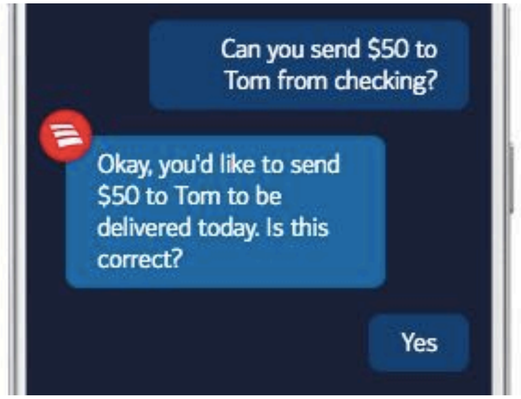


- **Required:** COS 226 (algorithms and data structures), probability, linear algebra, calculus
- COS 324 highly recommended (or be ready to pick up basic ML concepts!)
- Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Let’s do a quick poll!

Natural Language Processing

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



Banking assistant

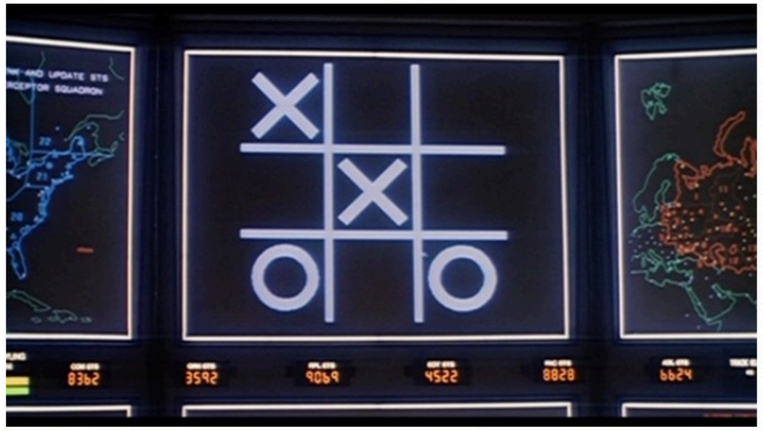
Natural Language Processing

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

Computer learns to play Civilization by reading the instruction manual

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

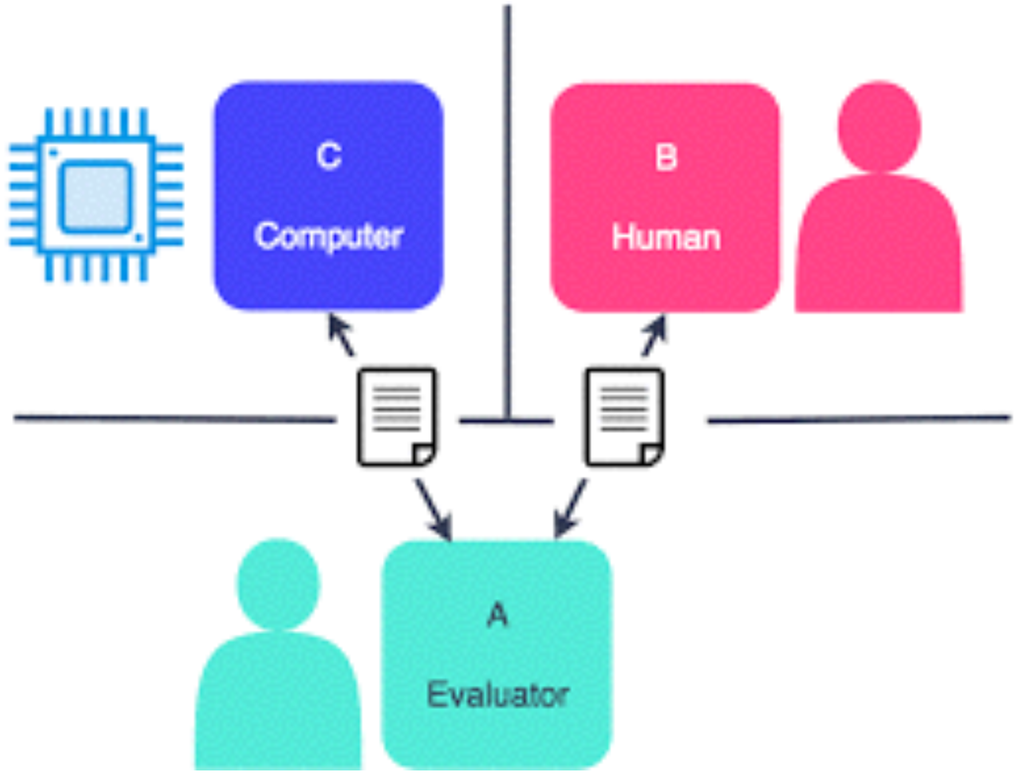
532 SHARES



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

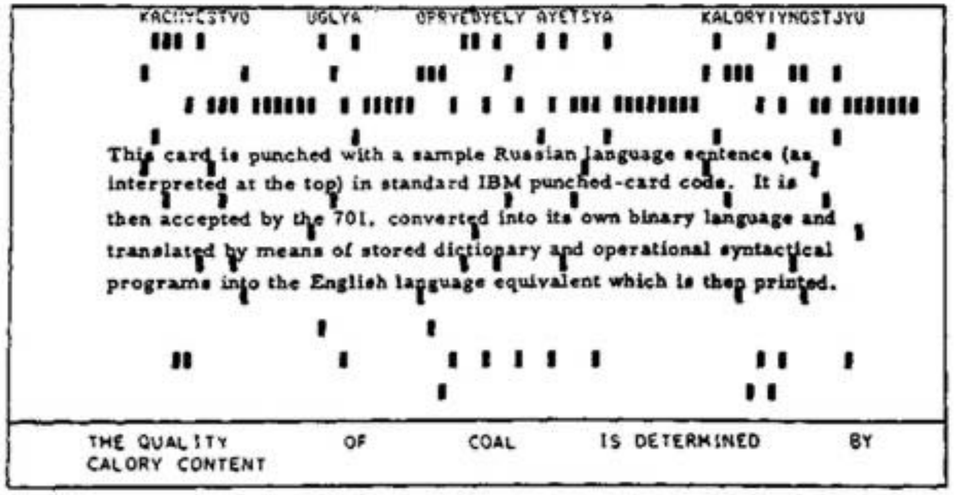
Jordan Zlatev[1,*] and Johan Blomberg[2,3]

Regular Article

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

Lera Boroditsky

How it started

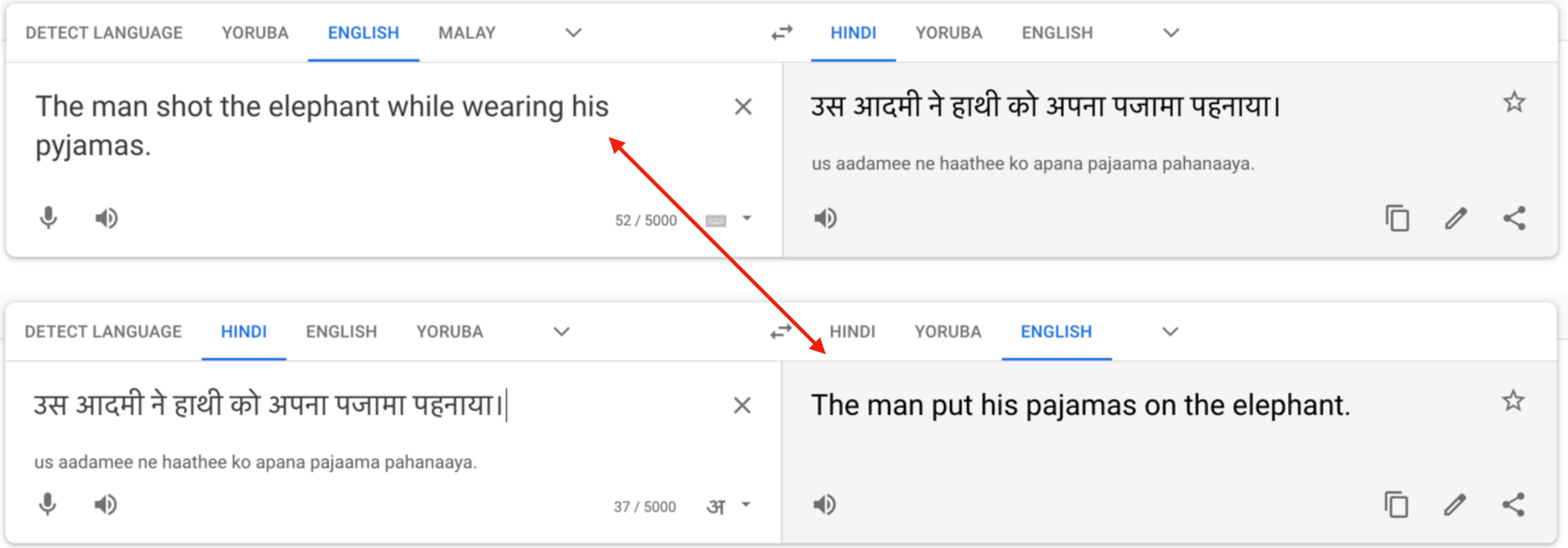


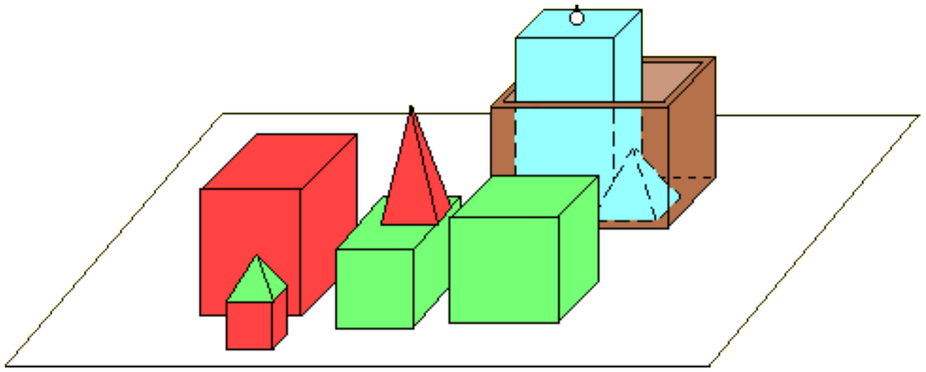
Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown experiment

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

How it's going





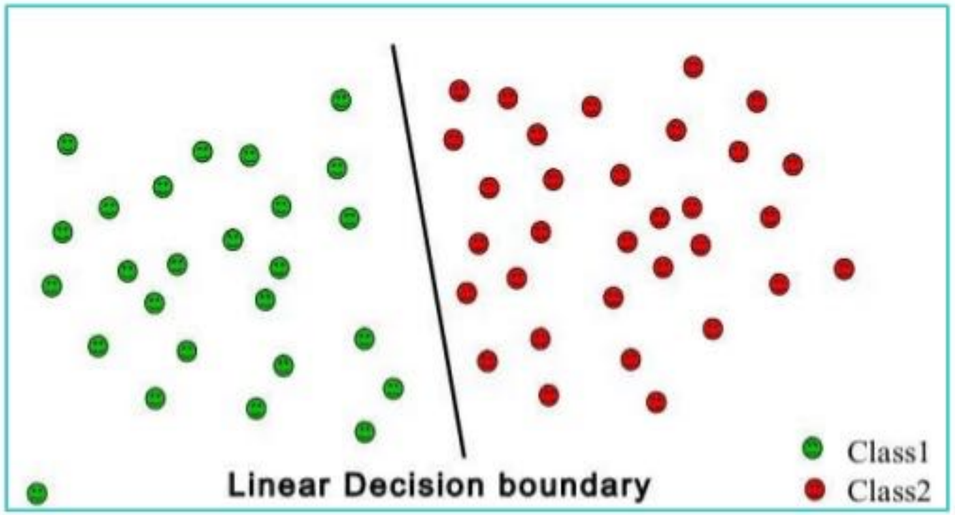
SHRDLU,
1968

> 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

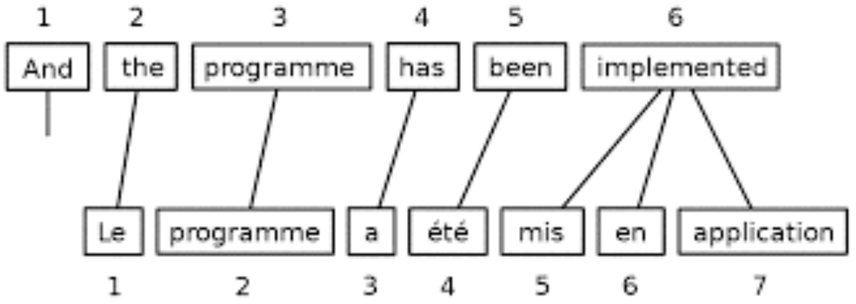
Statistical learning



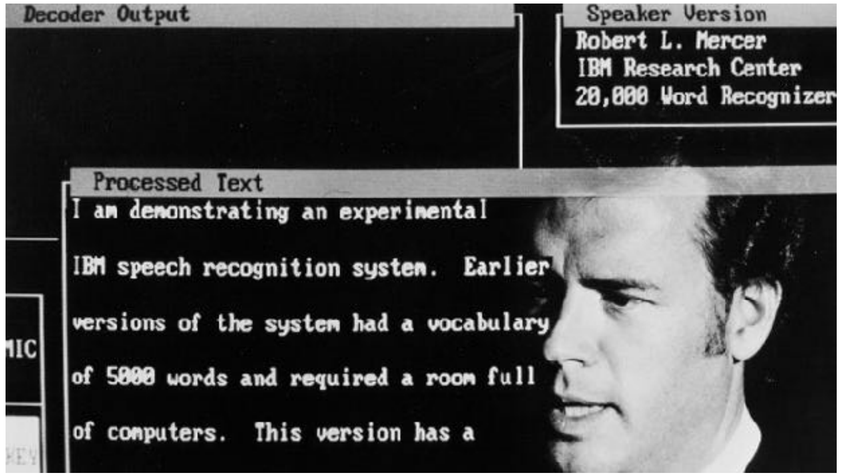
- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

Statistical learning

IBM translation models



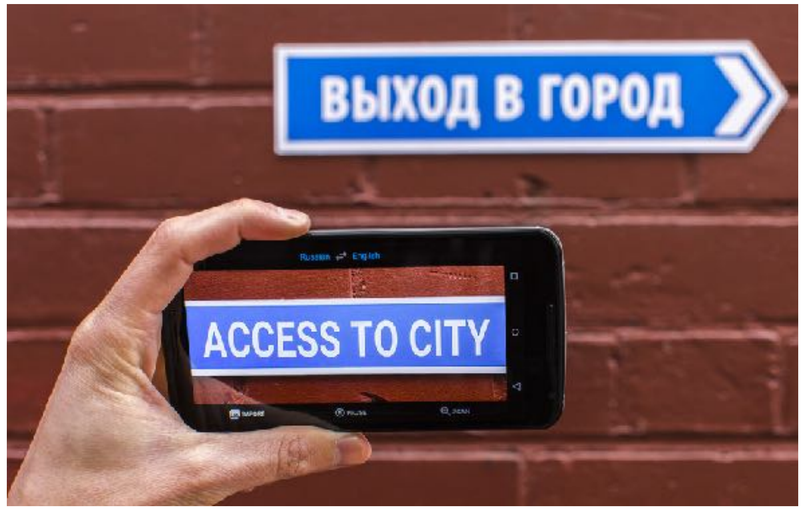
Speech recognition



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

Deep Learning era

- Significant advances in core NLP technologies
- 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!

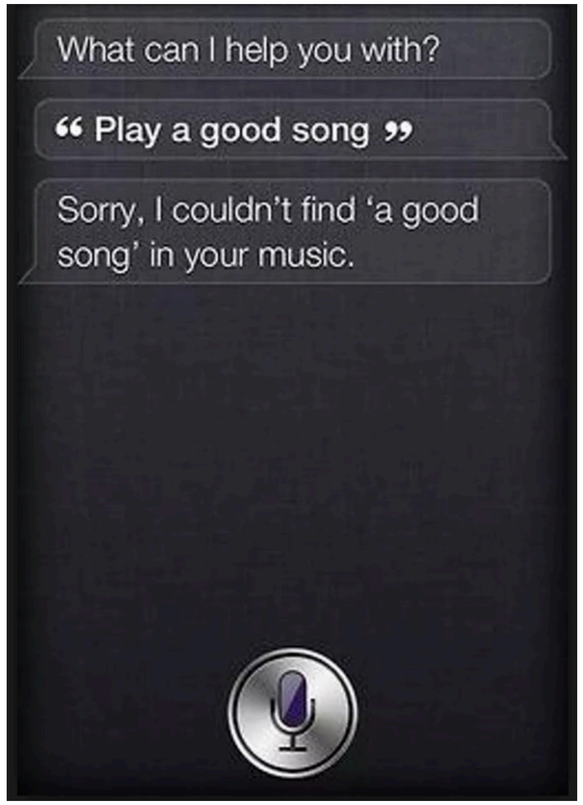
Turing test solved?

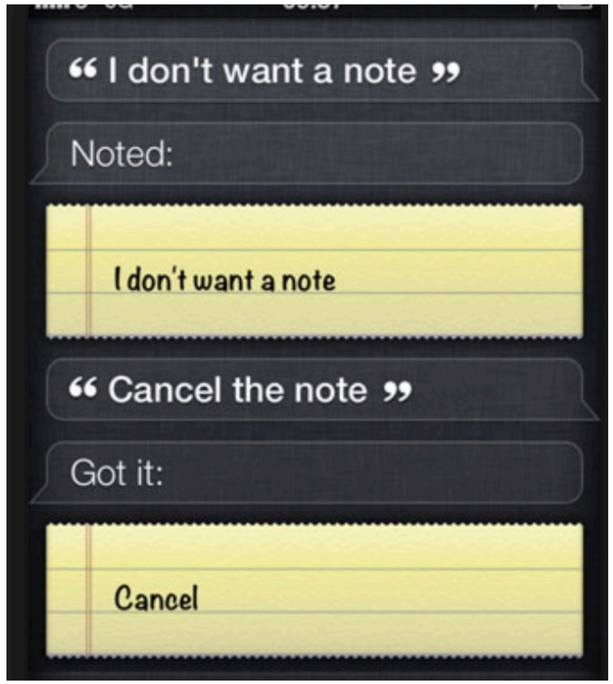
Talking to Google Duplex: Google's human-like phone AI feels revolutionary

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

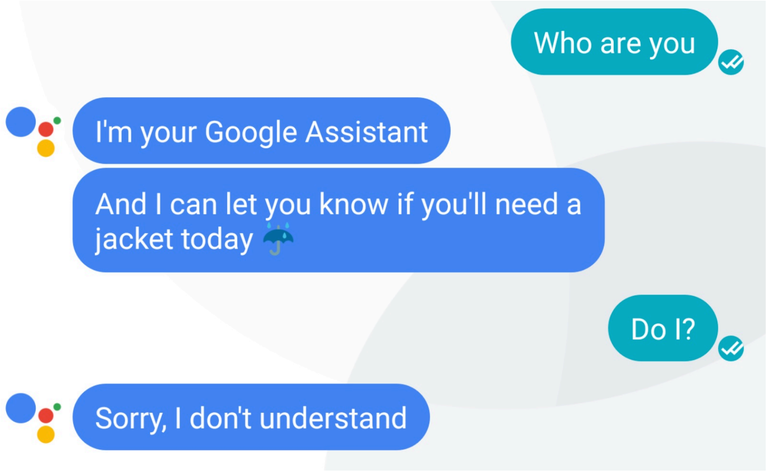
Who was the human in the call?

a) the receptionist
b) the person making the appointment
c) both of them
d) neither of them


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?

Type your thoughts in the chat!

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

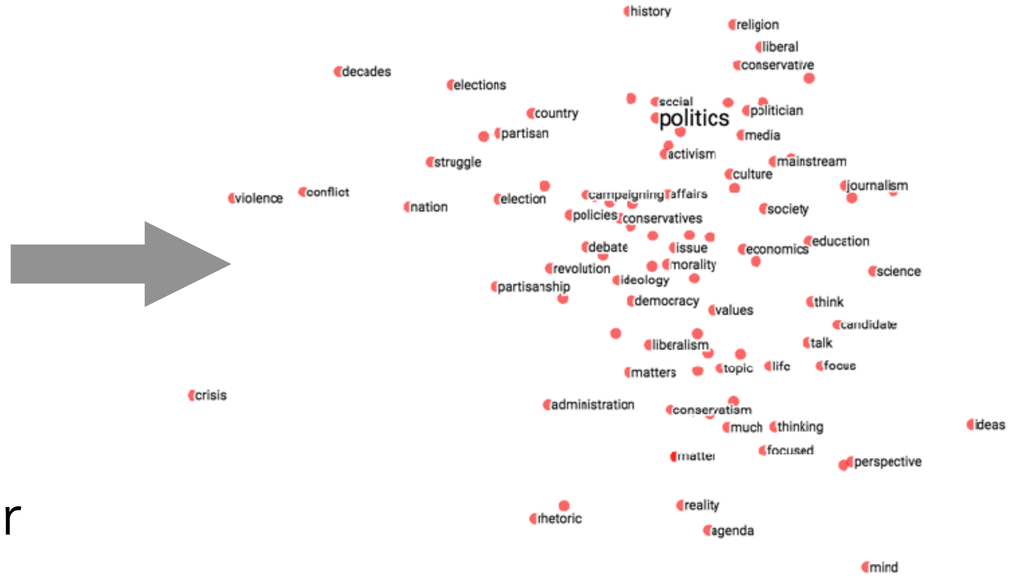
Lexical variations

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

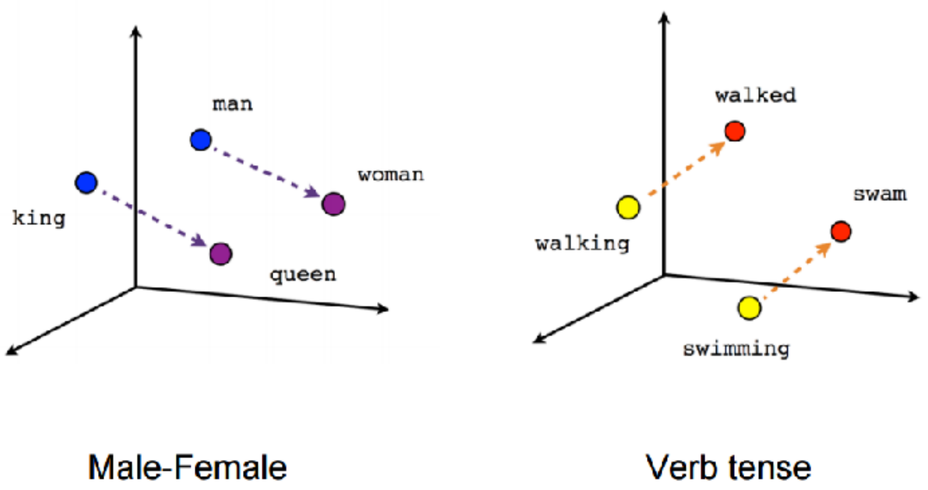
Several words can mean the same thing!

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

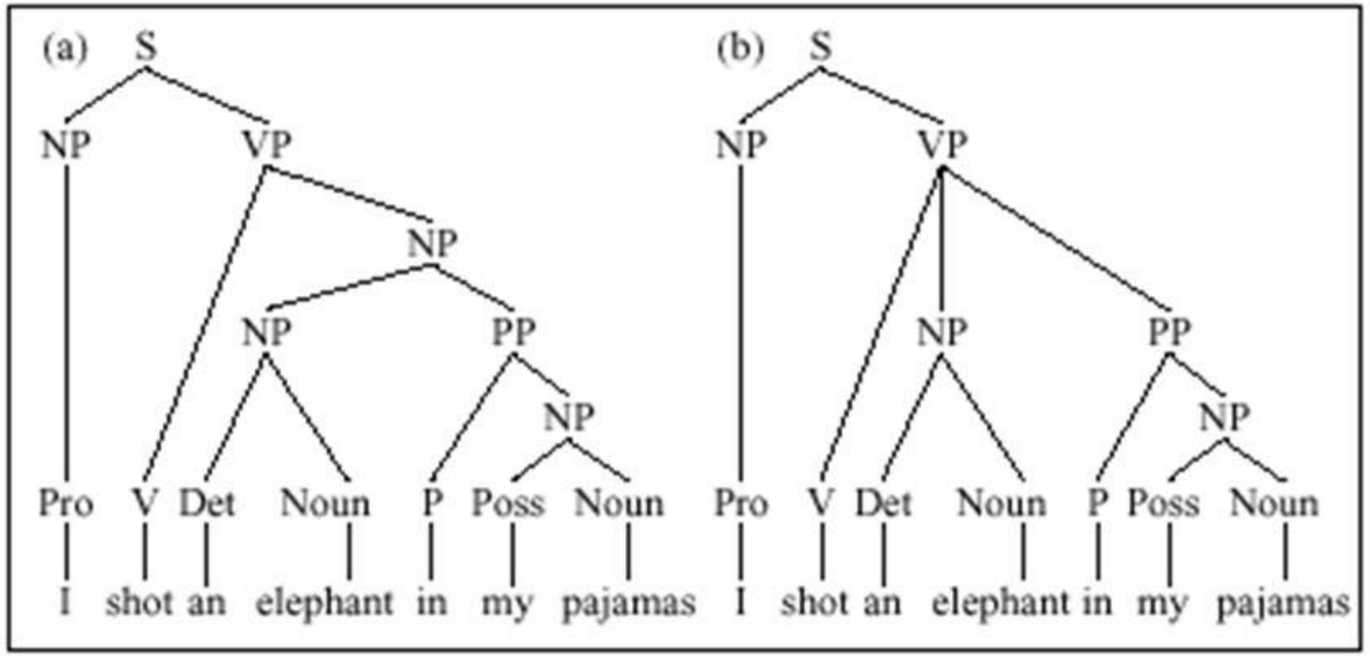
- My brother went to the park near my sister's house
- Park my went house near to sister's my brother the
- "My brother went park near sister's house"?
- The old man the boat

Garden Path sentence

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

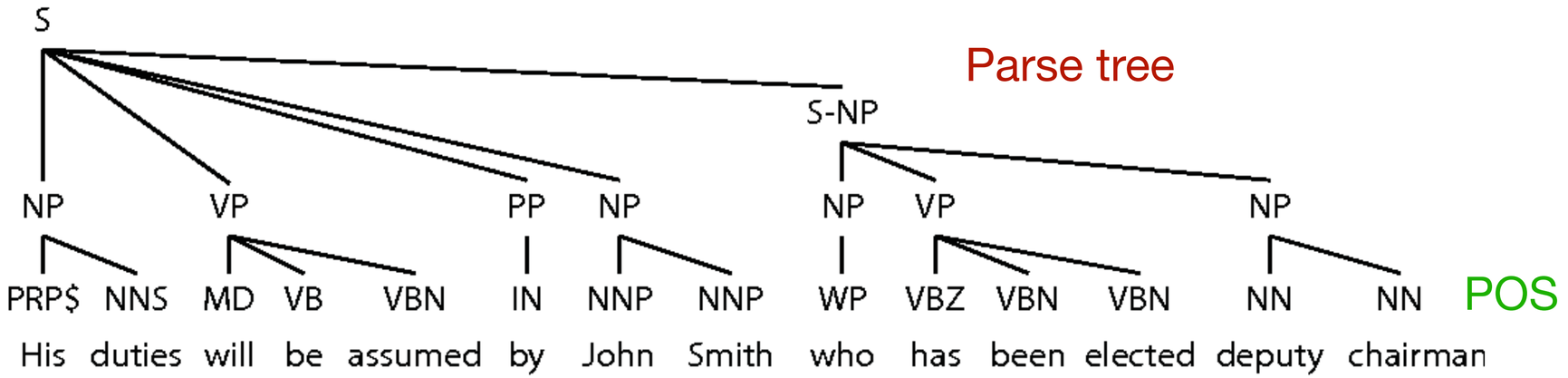
Syntactic ambiguity

I shot an elephant in my pajamas



Human language is full of such examples!

Syntactic parsing



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

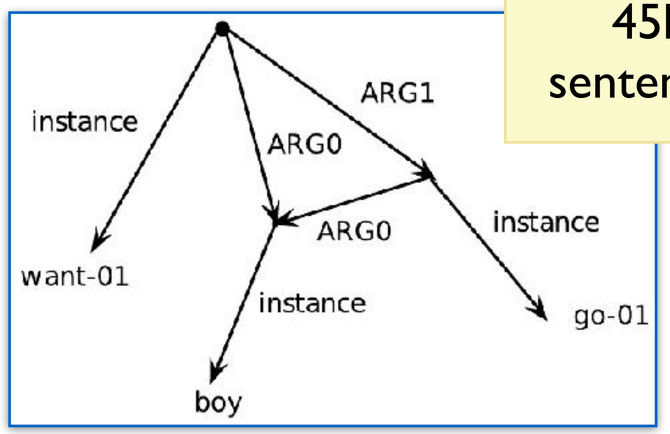
Tell my wife I love her

Action: Text

Arg1: Sarah Fox

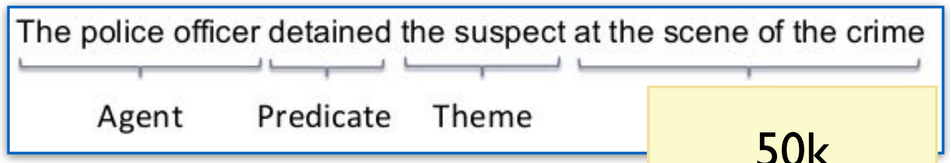
Arg2: I love you

Semantic Parsing



45k sentences

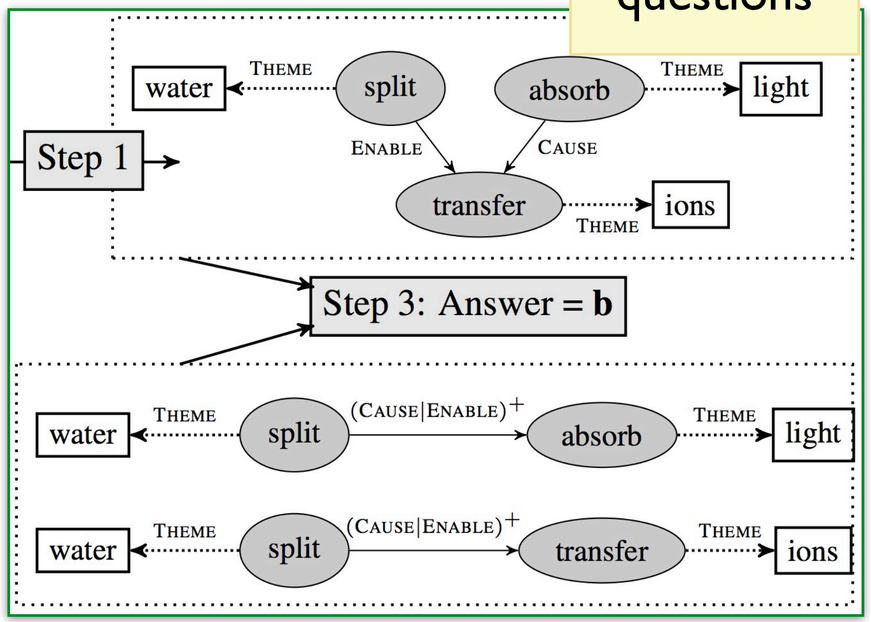
Abstract meaning representation



Semantic role labeling

50k arguments

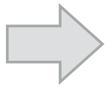
585 questions



(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

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

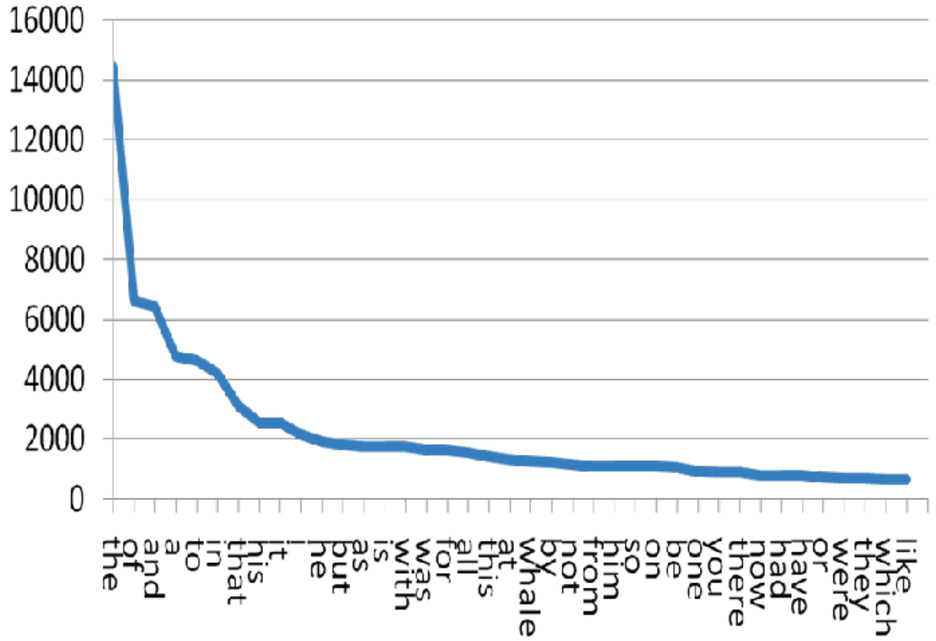
We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific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

- *Context:* Knowledge bases, perception, interaction

Outline

Words

- Language models
- Text classification
- Word embeddings

Sequences and structures

- HMMs, recurrent neural networks
- Syntactic Parsing
- Contextualized language models

Applications

- Machine Translation
- Question Answering
- Natural language inference
- Multimodal NLP

Language Models

Generating responses

- red ?
- 24.44 C ?
- sunny ?

Today, in New York, it is 76 F and red

vs

Today, in New York, it is 76 F and sunny

- Both are grammatical
- But which is more likely?

Language models

Probabilistic models over word sequences

Sentence "Princeton is in New Jersey"

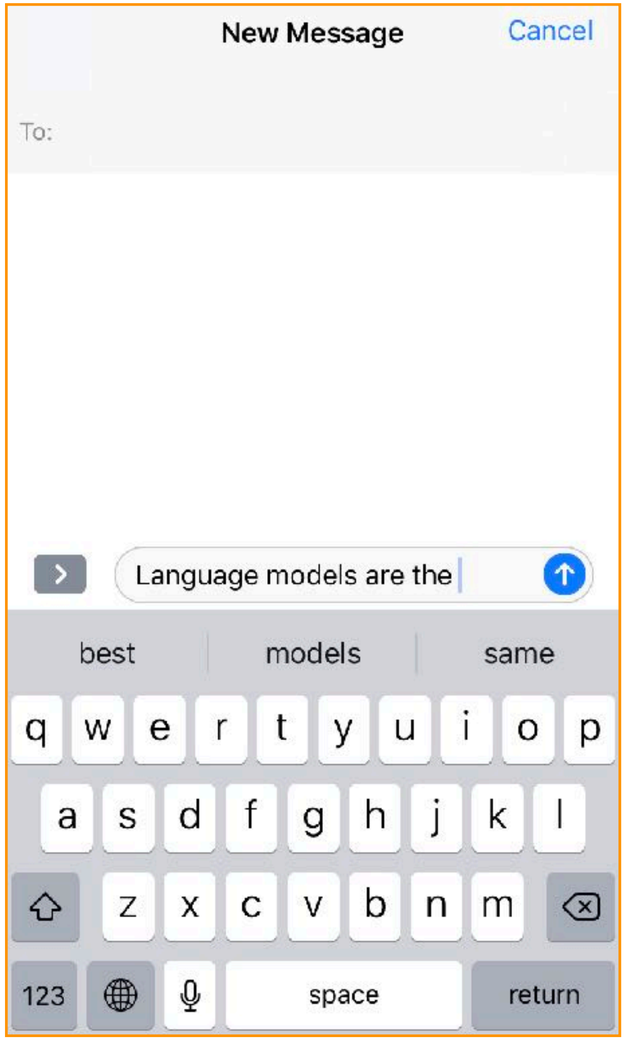
Chain rule

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

Completion Princeton is in New ?

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

Language models are everywhere



Applications of LMs

- Predicting words is important in many situations
  - Machine translation

$$P(\text{a } \mathbf{smooth} \text{ finish}) > P(\text{a } \mathbf{flat} \text{ finish})$$

  - Speech recognition/Spell checking

$$P(\text{high school } \mathbf{principal}) > P(\text{high school } \mathbf{principle})$$

  - Information extraction, Question answering

Impact on downstream applications

Language Resources	Adaptation	Word	
		Cor.	Acc.
1. Doc-A		54.5%	45.1%
2. Trans-C(L)		63.3%	50.6%
3. Trans-B(L)		70.2%	60.3%
4. Trans-A(S)		70.4%	59.3%
5. Trans-B(L)+Trans-A(S)	CM	72.6%	63.9%
6. Trans-B(L)+Doc-A	KW	72.1%	64.2%
7. Trans-B(L)+Doc-A	KP	73.1%	65.6%
8. Trans-A(L)		75.2%	67.3%

PP
49972
1856.5
318.4
442.3
225.1
247.5
259.7
148.6

(Miki et al., 2006)

New Approach to Language Modeling Reduces Speech Recognition Errors by Up to 15%

December 13, 2018
Ankur Gandhe

Alexa | Alexa research | Alexa science

What is a language model?

- Probabilistic model of a sequence of words
  - How likely is a given phrase/sentence/paragraph/document?
- Joint distribution of words $w_1, w_2, w_3, \ldots, w_n$:

$$P(w_1, w_2, w_3, ..., w_n)$$

Chain rule

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

Sentence: “the cat sat on the mat”

$$P(\text{the cat sat on the mat}) = P(\text{the}) * P(\text{cat}|\text{the}) * P(\text{sat}|\text{the cat})$$
$$*P(\text{on}|\text{the cat sat}) * P(\text{the}|\text{the cat sat on})$$
$$*P(\text{mat}|\text{the cat sat on the})$$

Estimating probabilities

$$P(\text{sat}|\text{the cat}) = \frac{\text{count}(\text{the cat sat})}{\text{count}(\text{the cat})}$$

$$P(\text{on}|\text{the cat sat}) = \frac{\text{count}(\text{the cat sat on})}{\text{count}(\text{the cat sat})}$$

$$\vdots$$

Maximum likelihood estimate (MLE)

- With a vocabulary of size v,
  - number of sequences of length n = v^n
- Typical vocabulary ≈ 40000 words
  - even sentences of length <= 11 results in more than $4 * 10^{50}$ sequences! *(more than the number of atoms in the earth)*

We'll tackle this in the next class!