CS378: Natural Language Processing
Lecture 1: Introduction

TEXAS
The University of Texas at Austin
Eunsol Choi

Some slides adapted from Greg Durrett, Yejin Choi, Yoav Artzi

Logistics

- Lecture: Tuesdays and Thursdays 3:30pm - 4:45pm
- Course website (including **syllabus**): http://www.cs.utexas.edu/~eunsol/courses/sp22-cs378/
- Piazza, Gradescope: link in the website
- Please complete welcome survey on Piazza.

Course Staff

- Instructor: Eunsol Choi
- Office Hour:
  - Tuesday 10-11AM
  - GDC 3.810

- Graduate TA: Hung-ting Chen
- Office Hour:
  - Friday 2-3PM
  - Online over Zoom: accessible to canvas

- All office hour starts next week!

Format and Accessibility

- Lectures will include times for discussion, in-class exercises, and questions.

- Required equipment: device to make Zoom calls with, some way to do homework
  - Lab machines available via SSH
  - A GPU is **not** required to complete the assignments.
    Having a GPU or GCP credits could be helpful for the final project.
  - We will offer some GPU credits over TACC.

FAQ

- How would course modality change over the course of semester?
  - After first two weeks, lectures will be in person at GDC 1.304.
  - Throughout the semester, the lecture online will be available on canvas *after* the lecture (often within a day).

Coursework

- Homework (40% of grade)
  - HW1-4 : Homework with programming and/or written parts (generally have 2 weeks before it is due).
  - Each homework will count equally to the final grade.
  - HW 1,2,4 is to be done individually.
  - HW 3 will be done in a group.

Coursework

- Final Project (25% of grade)
  - You have an option to do an independent research project or a more structured project.
  - Instruction will be released next week

- Class Participation (10% of grade)
  - Attending guest lectures, helping classmates on Piazza, asking questions during the lecture, insta poll during lecture

Coursework

- Exam (25% of grade)
  - Covering course materials, near the end of the semester.
  - In-class exam

Late Day Policy

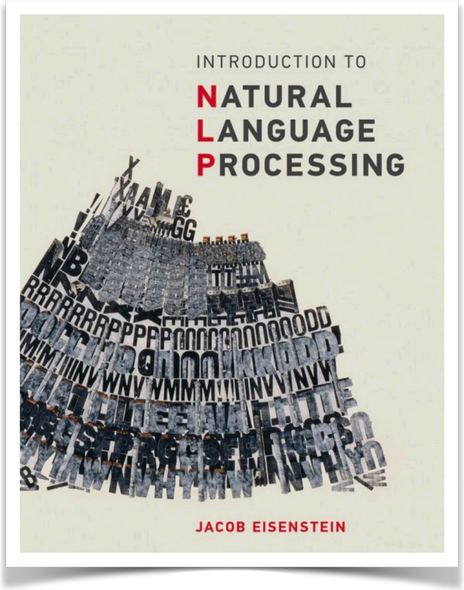
- You have 5 slip days (24 hour extensions) throughout the semester to use for assignments
- If used in team project, will consume both member's slip days.
- Cannot be used for any components of final project

Course Materials

- No need to buy textbooks!
- Tentative readings can be found in the course website, readings will be confirmed 1 week before the lecture.
  - Often chapters from textbooks (freely available as PDFs online), or recent research papers.

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin



Academic Honesty

- You may work in groups, but your final writeup and code **must be your (and your teammate's) own**
- Do **NOT** share code with others!
- If you are unsure, ask!
- For team projects, work TOGETHER with your partner, instead of simply dividing

What will be covered in this course

Machine learning methods to model language

Linguistic concepts

Applications of NLP

ML components covered in this course

- Classification models:
  - Naive Bayes, Logistic Regression
  - Feedforward Neural Network & Backpropagation
- Sequential models:
  - Hidden Markov Models
  - Maximum Entropy Markov Models
  - Expectation Maximization (EM)
- Neural network for Sequential data:
  - Recurrent Neural Network / Convolutional Neural Network
  - Transformers
  - Encoder-Decoder model

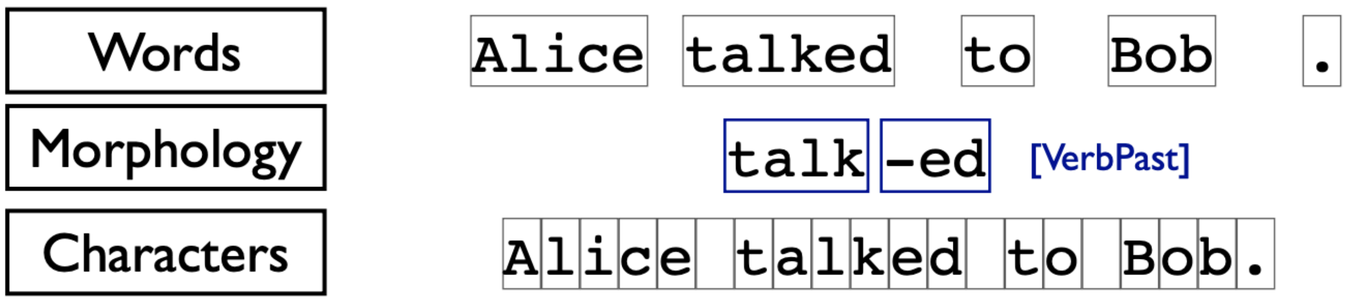
What will be covered in this course

Machine learning methods to model language

Linguistic concepts

Applications of NLP

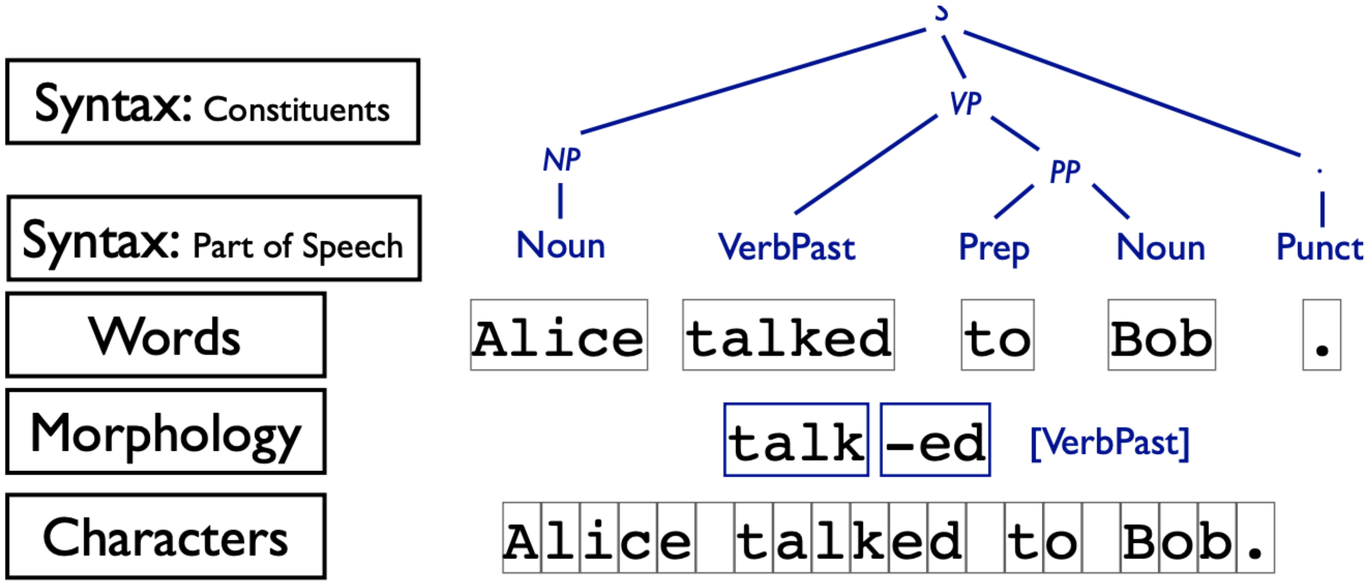
Levels of linguistic structure



- Phonetics / Phonology / Morphology: what words (or subwords) are we studying?

Slide credit: Jason Eisner, Mohit Iyyer

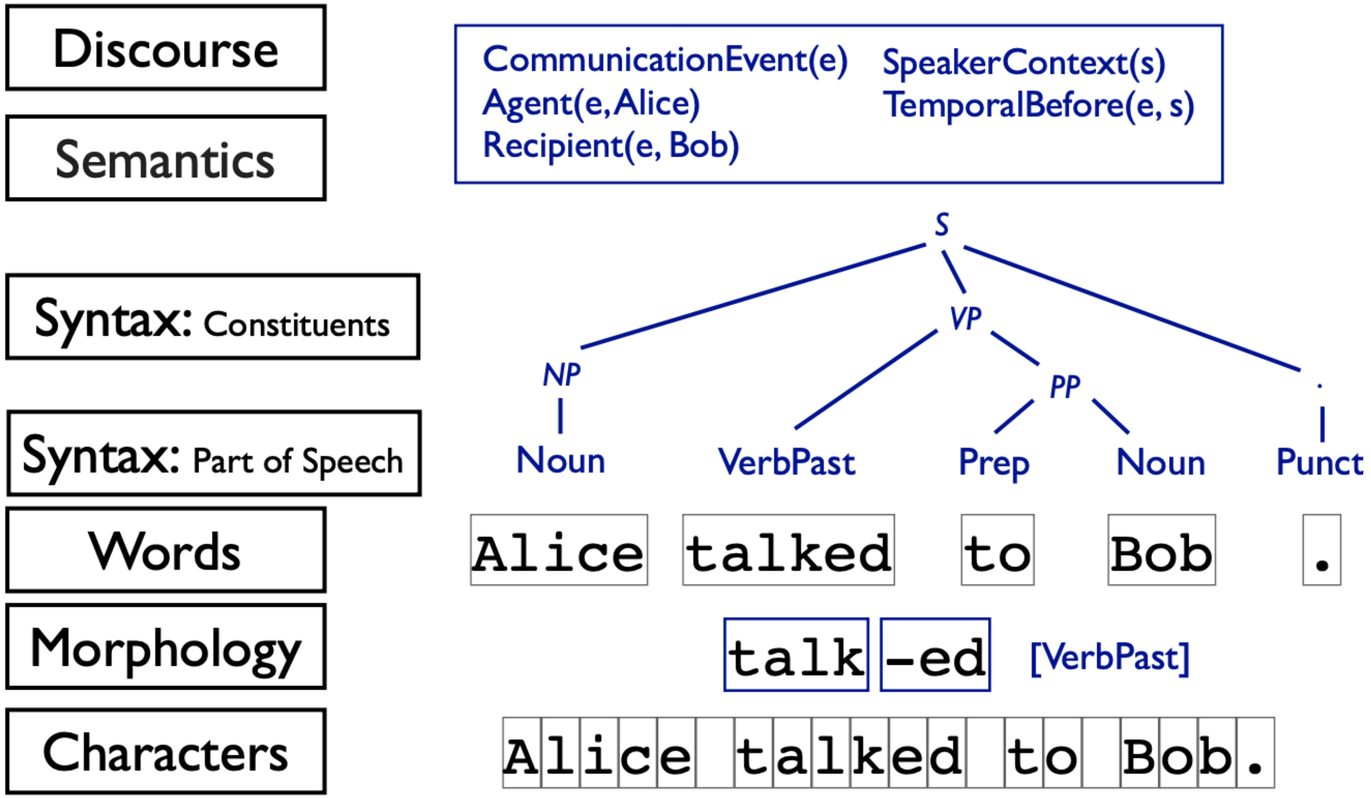
Levels of linguistic structure



- Syntax: What types of phrases are we studying? Which words are modifying one another?
- Phonetics / Phonology / Morphology: what words (or subwords) are we studying?

Slide credit: Jason Eisner, Mohit Iyyer

Levels of linguistic structure



- Semantics: What's the literal interpretation of the sentence?
- Syntax: What types of phrases are we studying? Which words are modifying one another?
- Phonetics / Phonology / Morphology: what words (or subwords) are we studying?

Slide credit: Jason Eisner, Mohit Iyyer

Levels of linguistic structure

Pragmatics

Discourse

Semantics

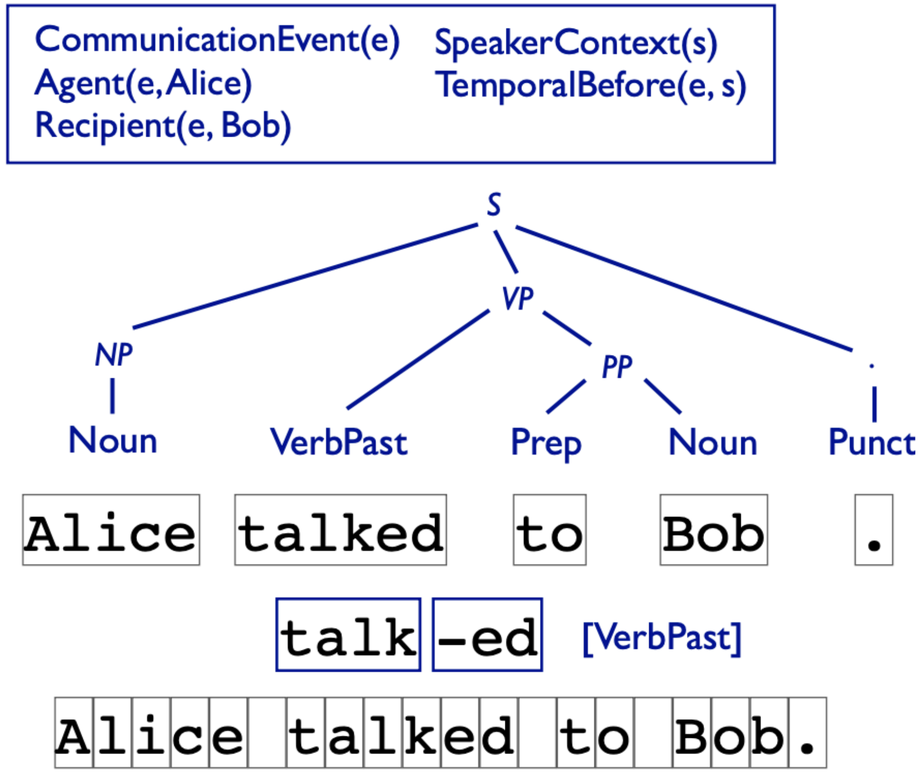
Syntax: Constituents

Syntax: Part of Speech

Words

Morphology

Characters



- What conclusions can we draw from the utterance?
- Semantics: What's the literal interpretation of the sentence?
- Syntax: What types of phrases are we studying? Which words are modifying one another?
- Phonetics / Phonology / Morphology: what words (or subwords) are we studying?

Slide credit: Jason Eisner, Mohit Iyyer

What will be covered in this course

Machine learning methods to model language

Linguistic concepts

Applications of NLP

What’s the goal of NLP?

?

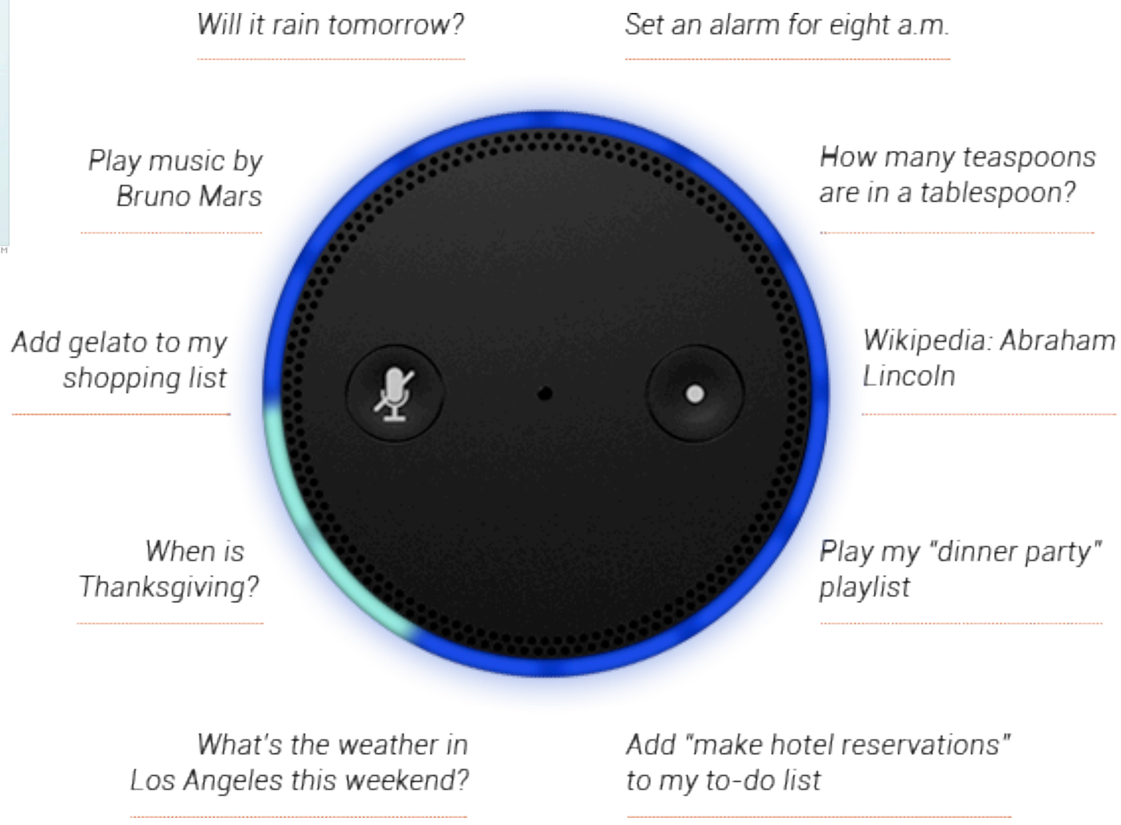
?

- Be able to solve **problems** that require **deep** understanding of **text**

?

Already in our daily lives

FROM: AMAZON.COM



Machine Translation

"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959

Vidéo Anniversaire de la rébellion

"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

Video Anniversary of the Tibetan rebellion: China on guard

- Translate text from one language to another
- Challenges:
  - How to make efficient? [fast translation search]
  - Fluency vs. Fidelity

Question Answering

When was Abraham Lincoln born?

map to **Birthday** field

Name	**Birthday**
Lincoln, Abraham	2/12/1809
Washington, George	2/22/1732
Adams, John	10/30/1735

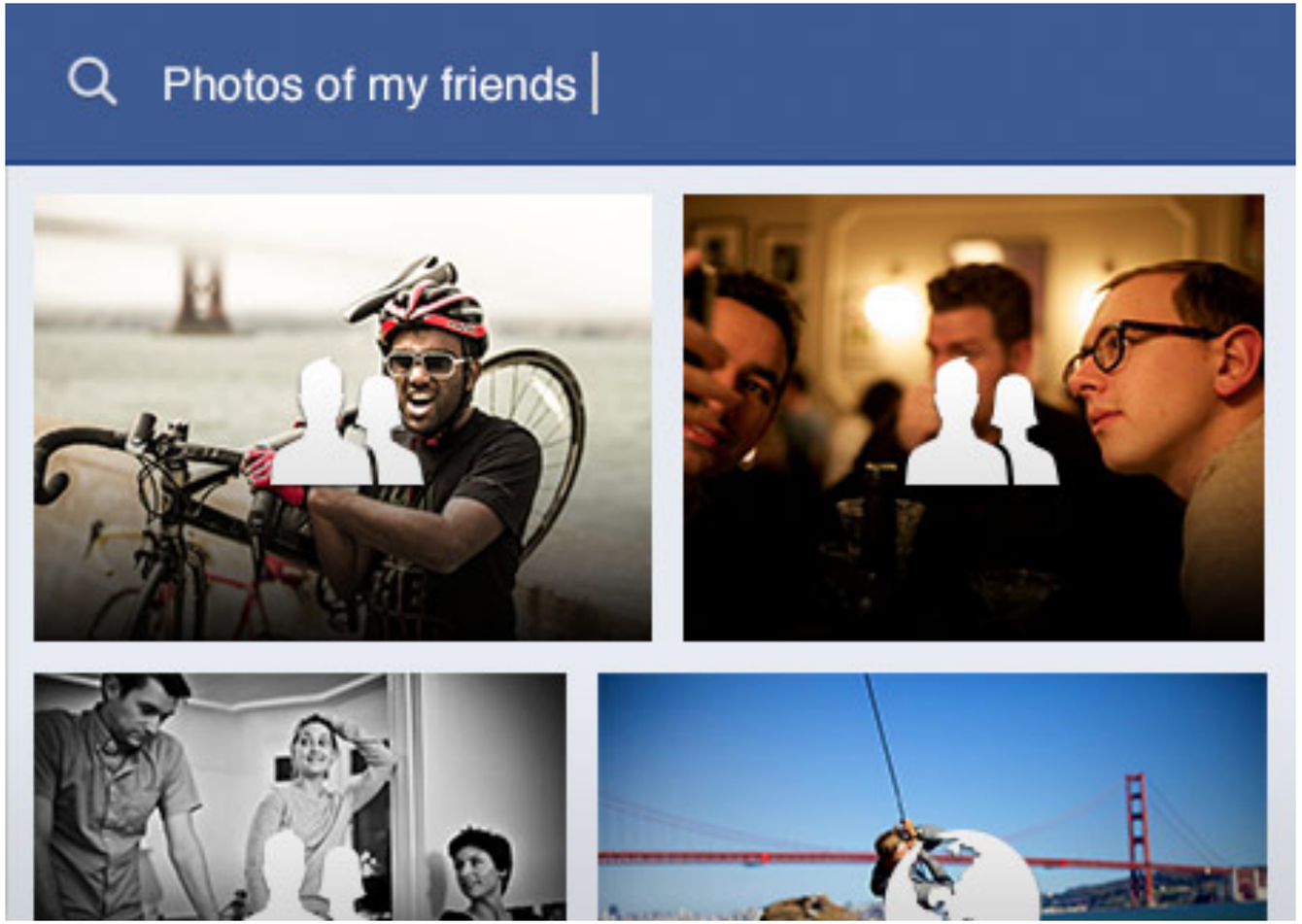
February 12, 1809

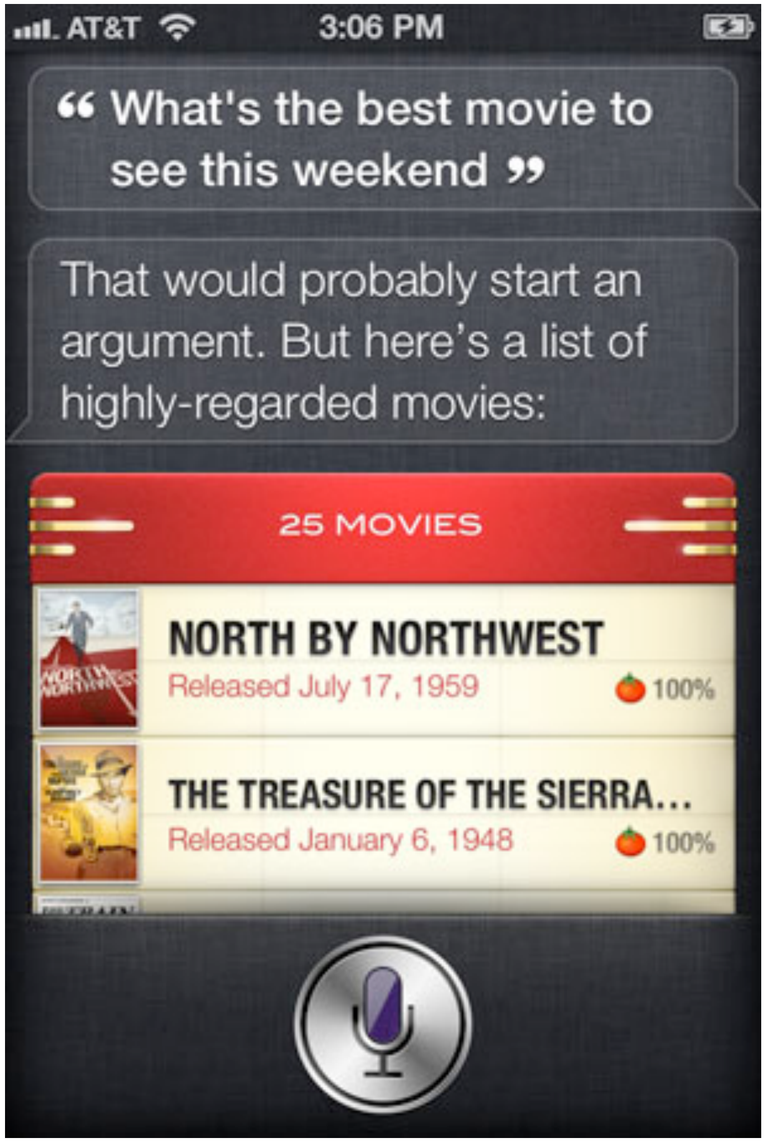
How many visitors centers are there in Rocky Mountain National Park?

The park has a total of five visitor centers

five

Question Answering in the Wild





Automatic Summarization

POLITICS

Google Critic Ousted From Think Tank Funded by the Tech Giant

WASHINGTON — In the hours after European antitrust regulators levied a record $2.7 billion fine against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

• • •

But not long after one of New America's scholars posted a statement on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president, Anne-Marie Slaughter, according to the scholar.

• • •

Ms. Slaughter told Mr. Lynn that "the time has come for Open Markets and New America to part ways," according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

compress text

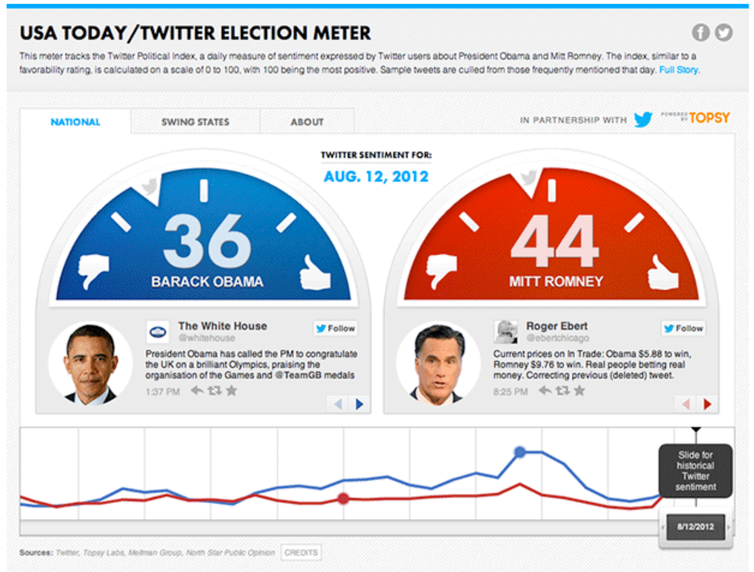
provide missing context

One of New America's writers posted a statement critical of Google. Eric Schmidt, Google's CEO, was displeased.

The writer and his team were dismissed.

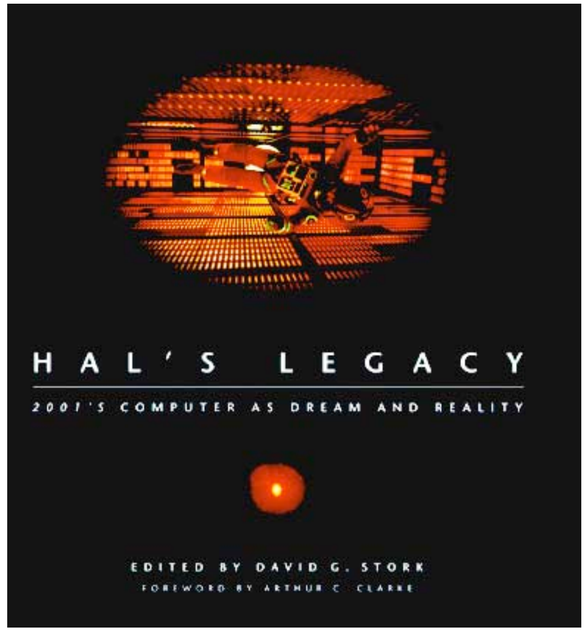
paraphrase to provide clarity

Analyzing media, making predictions



- Today: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- Past: “Sentiment Analysis” research started in 2002
- Future: computational social science and NLP for digital humanities (psychology, communication, literature and more)

Language And Vision



"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"

-- David Stork (HAL's Legacy, 2001) on A. Rosenfeld's vision

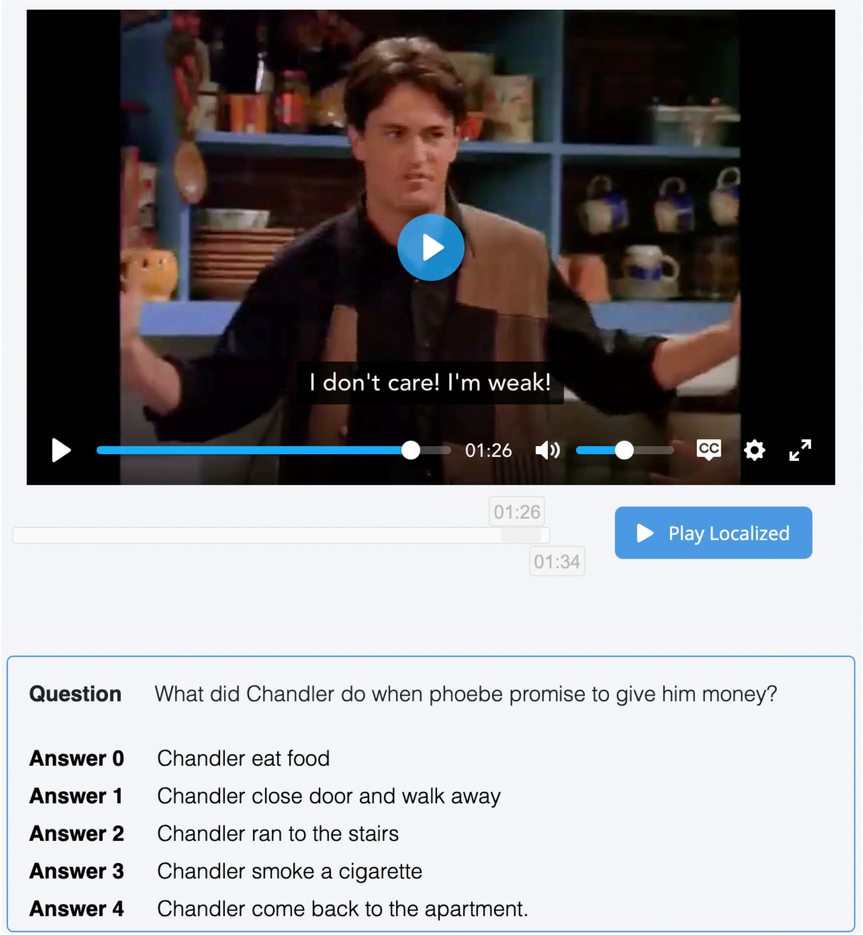
Language And Vision

https://openai.com/blog/dall-e/

https://visualqa.org/

Language And Vision & Speech

- Video understanding



https://tvqa.cs.unc.edu/explore.html

Applications of NLP

- Machine Translation
- Question Answering
- Spelling correction, grammar checking
- Psychotherapy & Analysis
- Providing new interfaces to assess information:
  - Dialogue systems
  - Speech recognition
  - Image Retrieval

Outline of the Course

- Classification: linear and neural (2 weeks)
- Sequence Modeling (1.5 weeks)
- Meta - NLP / Ethics in NLP (1 week)
- Word Embeddings / Language Model (1.5 week)
- Sequence Modeling, revisited with Neural Network (1.5 weeks)
- Contextualized Word Embeddings (1 week)
- Tree Modeling (1 weeks)
- Machine Translation (1 week)
- Grounding (1 week)
- QA / Dialogue (1 week)

Guest Lectures

Chenhao Tan (3/24)

Assistant Professor at University of Chicago

Human-centered machine learning, Language and social dynamics

Peter Anderson (4/26)

Research Scientist at Google AI Austin

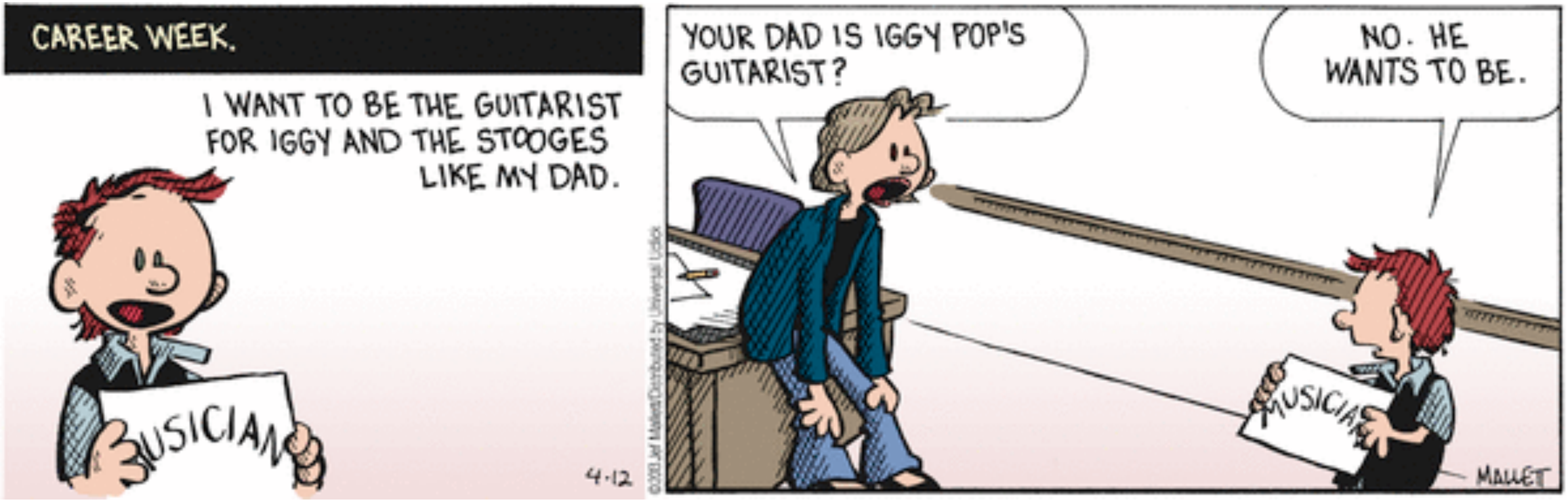
Vision + Language / Navigation

Today

- Why study NLP?
- **Why is NLP hard?**
- Little bit of history
- Current state of the field

Why is language hard?

Language is Ambiguous!



Semantic Ambiguity

At last, a computer that understands you like your mother.

- Direct meanings:
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
- "*mother*" could mean:
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
- Context matters, e.g. what if previous sentence was:
  - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients.

[Example from Lillian Lee]

Ambiguity in the wild

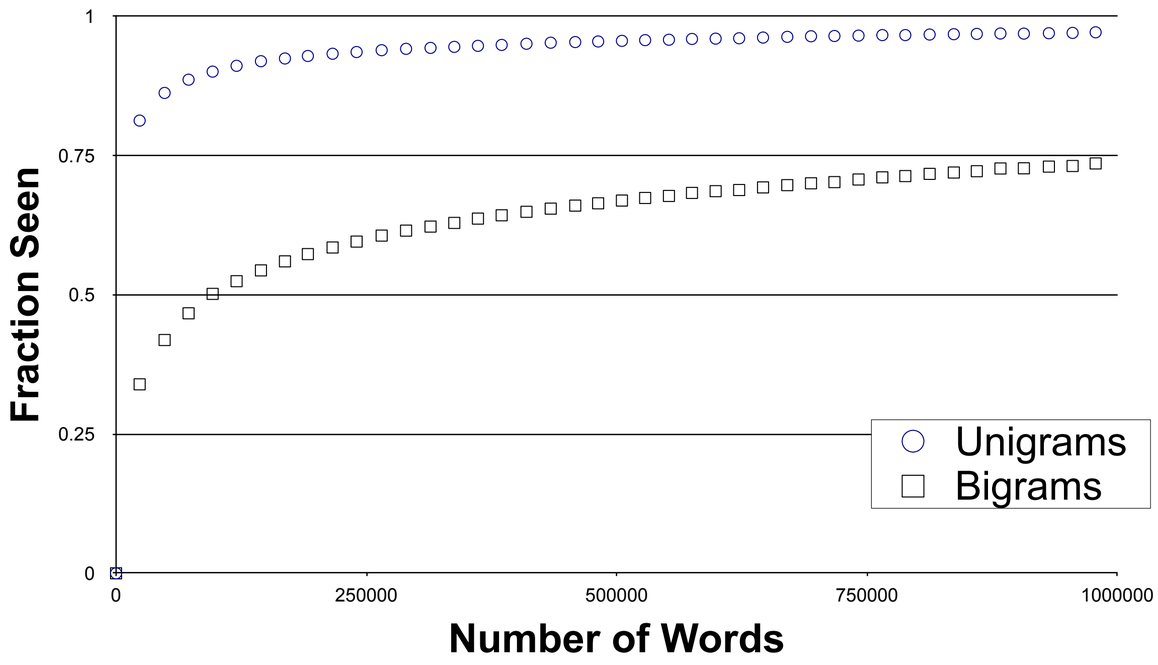
Situated understanding of language

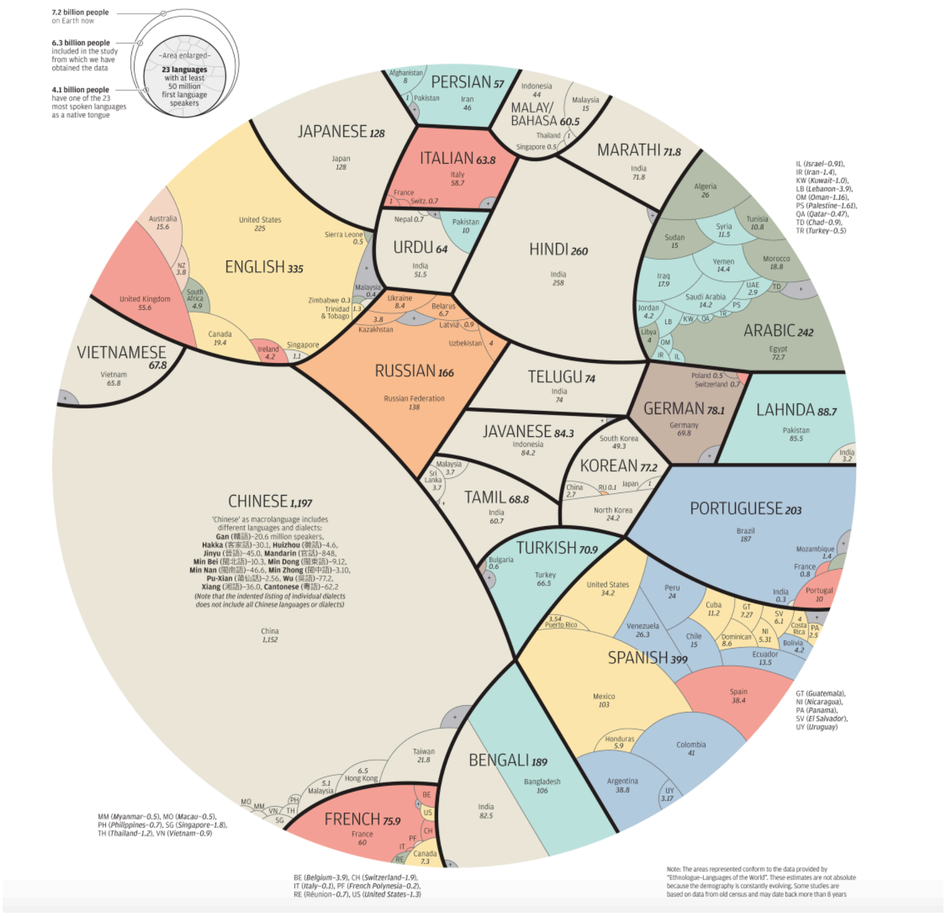
What the British say	What others understand	What the British mean
I hear what you say	He accepts my point of view	I disagree and do not want to discuss it further
With the greatest respect...	He is listening to me	I think you are an idiot
That's not bad	That's poor	That's good
That is a very brave proposal	He thinks I have courage	You are insane
Quite good	Quite good	A bit disappointing
I would suggest...	Think about the idea, but do what you like	Do it or be prepared to justify yourself
Oh, incidentally / by the way...	That is not very important	The primary purpose of our discussion is...
I was a bit dissapointed that	It doesn't really matter	I am annoyed that
Very interesting	They are impressed	That is clearly nonsense.
I'll bear it in mind	They will probably do it	I've forgotten it already
I'm sure it's my fault	Why do you they think it was their fault?	It's your fault
You must come for dinner	I will get an invitation soon	It's not an invitation, I'm just being polite

https://greatbritishmag.co.uk/uk-culture/what-the-british-say-and-what-they-actually-mean/

Sparsity of the data

- New word constantly comes up!
- Even worse for low resource languages





https://www.lucasinfografia.com/Mother-tongues

Today

- Why study NLP?
- Why is NLP hard?
- **Little bit of history**
- Current state of the field

Brief History of NLP

- 1940-50s: introducing probability
- 1950-80s: expert hand-written rules
- 1990s: statistical model coming back

Analyzing the dependent probabilities of letters and words appearing in combination with each other: statistical modeling of English

First attempt: Statistical Modeling of Language

- The Shannon Game:
  - How well can we predict the next word?

    When I eat pizza, I wipe off the ____

    Many children are allergic to ____

    I saw a ____

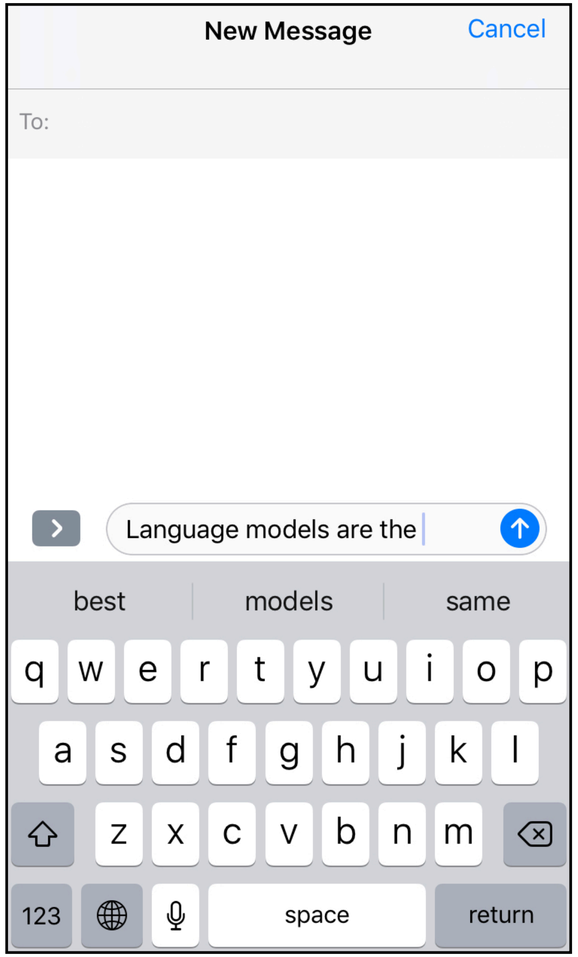
grease 0.5
sauce 0.4
dust 0.05
….
mice 0.0001
….
the 1e-100

Claude Shannon

- How good are we doing?
- Compute per word log likelihood (total n words):

$$l = \frac{1}{n}\sum_{i=1}^{n} \log P(x_i \,|\, x_1, x_2 \ldots x_{i-1})$$

Predicting next word is useful!



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

Brief History of NLP

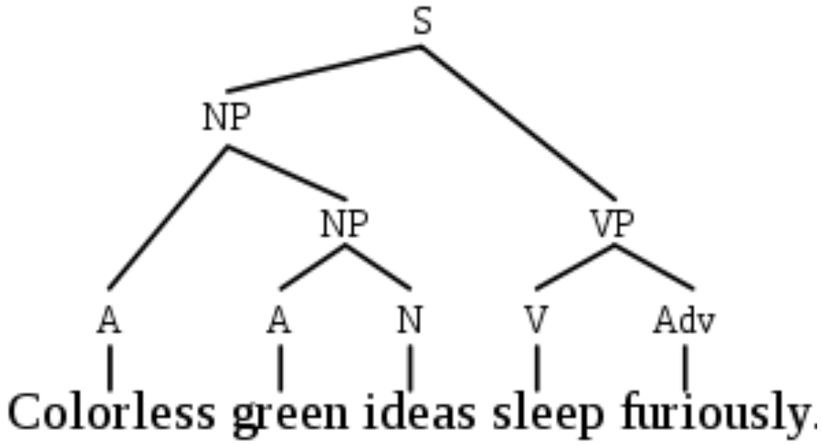
- 1940-50s: introducing probability
- 1950-80s: expert hand-written rules
- 1990s: statistical model coming back

Second Attempt: Writing Rules

(1) Colorless green ideas sleep furiously.

(2) Furiously sleep ideas green colorless.

It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not.” (Chomsky 1957)



Brief History of NLP

- 1940-50s: introducing probability
- 1950-80s: expert hand-written rules
- 1990s: statistical model coming back

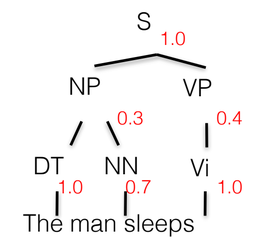
Annotating Data: Penn Treebank (1993)

```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (NP
        (NP (DT a) (NN round))
        (PP (IN of)
          (NP
            (NP (JJ similar) (NNS increases))
            (PP (IN by)
              (NP (JJ other) (NNS lenders)))
            (PP (IN against)
              (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
      (, ,)
      (S-ADV
        (NP-SBJ (-NONE- *))
        (VP (VBG reflecting)
          (NP
            (NP (DT a) (VBG continuing) (NN decline))
            (PP-LOC (IN in)
              (NP (DT that) (NN market)))))))
    (. .)))
```

- 50,000 annotated sentences!
- Usual set-up:
  - 40,000 training
  - 2,400 test

S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	IN	NP	1.0

t_1=



$p(t_1)$=1.0*0.3*1.0*0.7*0.4*1.0

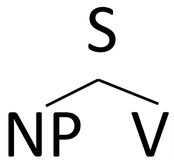
S

A brief history of statistical NLP

"AI winter"
rule-based,
expert systems

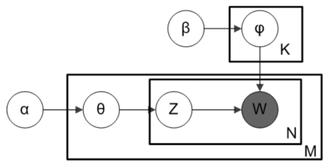
earliest stat MT
work at IBM

Penn
treebank



Collins vs.
Charniak
parsers

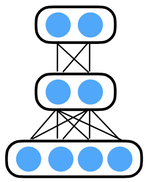
Unsup: topic
models,
grammar induction



Sup: SVMs,
CRFs, NER,
Sentiment

Semi-sup,
structured
prediction

Neural



1980 1990 2000 2010 2020

Today

- Why study NLP?
- Why is NLP hard?
- Little bit of history
- **Current state of the field**

Where are we?

- Exciting Time!
  - Rapid progress in multiple benchmark tasks (machine translation, question answering)
  - Active interaction with other disciplines (vision, robotics, speech)

Where are we?

- Interesting time — Working “formula” for many tasks

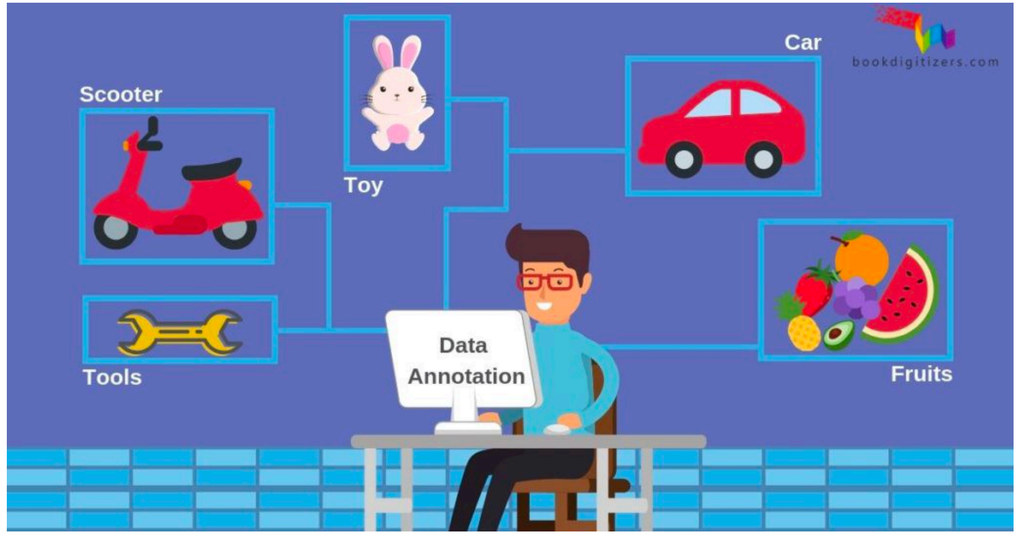
Self-supervision on a lot of text

Randomly masked: A quick [MASK] fox jumps over the [MASK] dog

↓ ↓

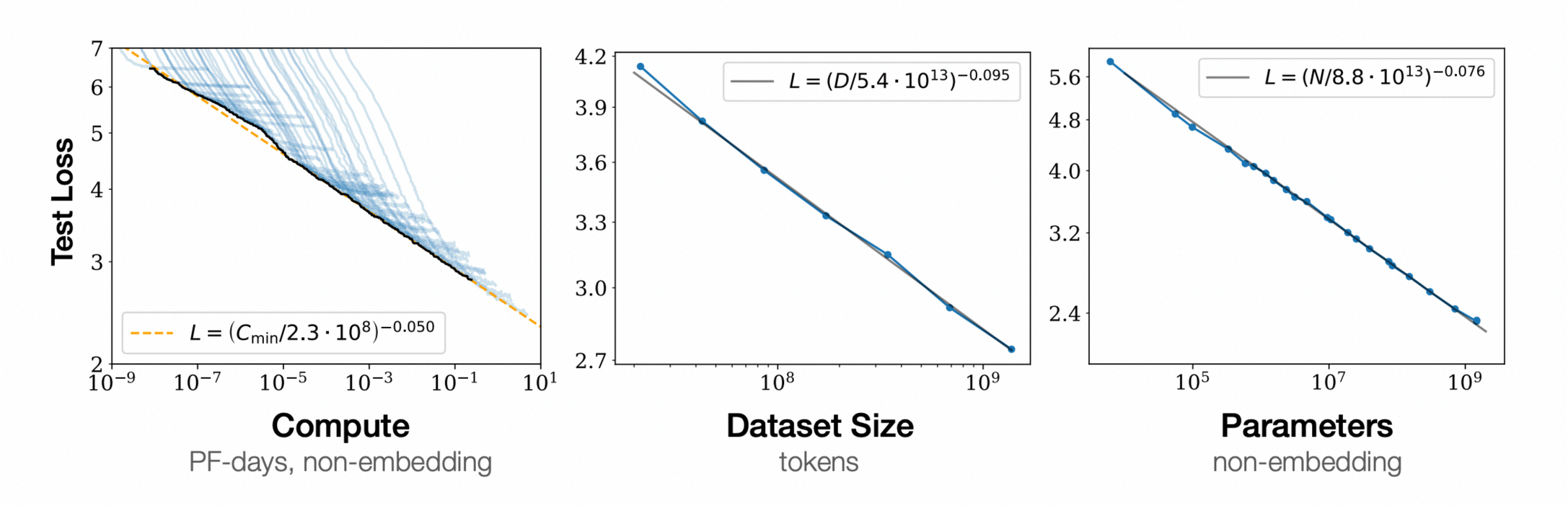
Predict: A quick brown fox jumps over the lazy dog

Fine-tuning on supervised data



Where are we?

- **A lot** of recent progress came from **simple and expensive**, and really powerful scaling



https://arxiv.org/pdf/2001.08361.pdf

Where are we?

- Yet, we all seem to agree scaling alone won't solve all the problems..
  - Limited success in integrating with perception, interaction, social grounding
  - Symbolic reasoning, common sense reasoning, compositionally
  - Ethical concerns: Models learn biases in available data

man

woman

Related Fields

- Computational Linguistics:
  - use computational tools to study language
  - Closely related to NLP

- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype (for now)

- Speech
  - Mapping audio ***signals*** to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Conduct

A climate conducive to learning and creating knowledge is the right of every person in our community. Bias, harassment and discrimination of any sort have no place here. If you notice an incident that causes concern, please contact the Campus Climate Response Team: **diversity.utexas.edu/ccrt**

The University of Texas at Austin
College of Natural Sciences

The College of Natural Sciences is steadfastly committed to enriching and transformative educational and research experiences for every member of our community. Find more resources to support a diverse, equitable and welcoming community within Texas Science and share your experiences at ***cns.utexas.edu/diversity***

Demo!

AllenNLP

https://demo.allennlp.org/masked-lm