

# CSEP 517

# Natural Language Processing

Introduction  
Luke Zettlemoyer

Slides adapted from Dan Klein, Yejin Choi

# What is NLP?



- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching
- End systems that we want to build:
  - Simple: spelling correction, text categorization...
  - Complex: speech recognition, machine translation, information extraction, sentiment analysis, question answering...
  - Unknown: human-level comprehension (is this just NLP?)

# Why NLP

---

- To access information & knowledge

# Jeopardy! World Champion

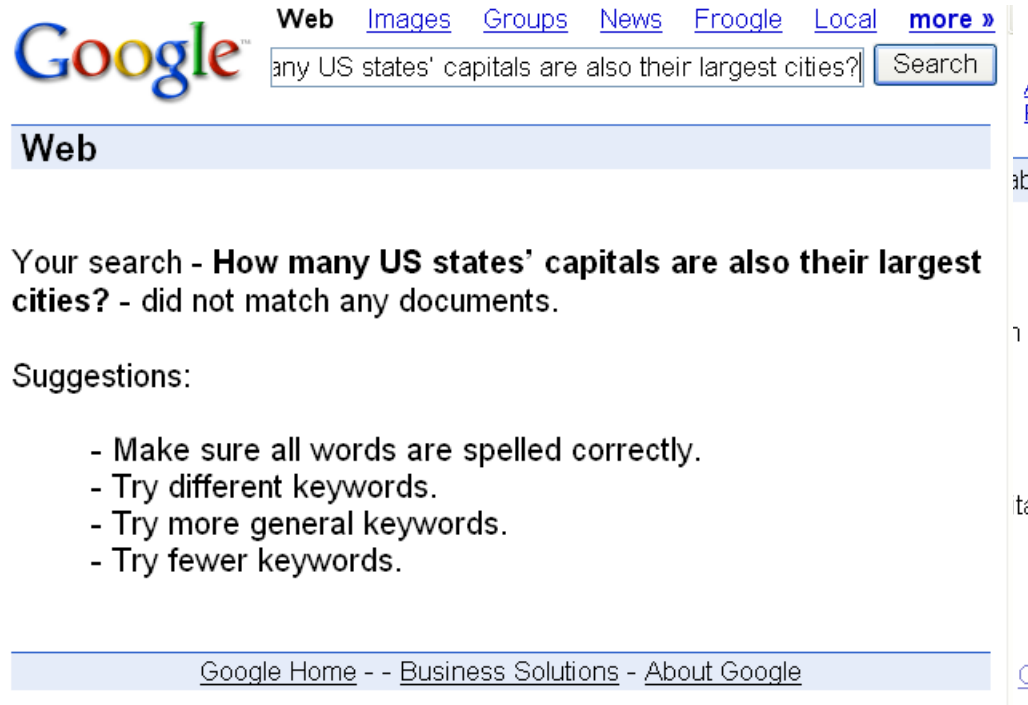


US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.



# Question Answering

- Question Answering:
  - More than search
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"



The screenshot shows a Google search interface. At the top, the Google logo is on the left, and navigation links for 'Web', 'Images', 'Groups', 'News', 'Froogle', 'Local', and 'more »' are on the right. The search bar contains the text 'any US states' capitals are also their largest cities?' and a 'Search' button. Below the search bar, a 'Web' tab is selected. The search results section displays the text: 'Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.' Below this, a 'Suggestions:' section lists four items: '- Make sure all words are spelled correctly.', '- Try different keywords.', '- Try more general keywords.', and '- Try fewer keywords.' At the bottom of the search results area, there are links for 'Google Home', 'Business Solutions', and 'About Google'.

## [capital of Wyoming: Information From Answers.com](#)

Note: click on a word meaning below to see its connections and related words.

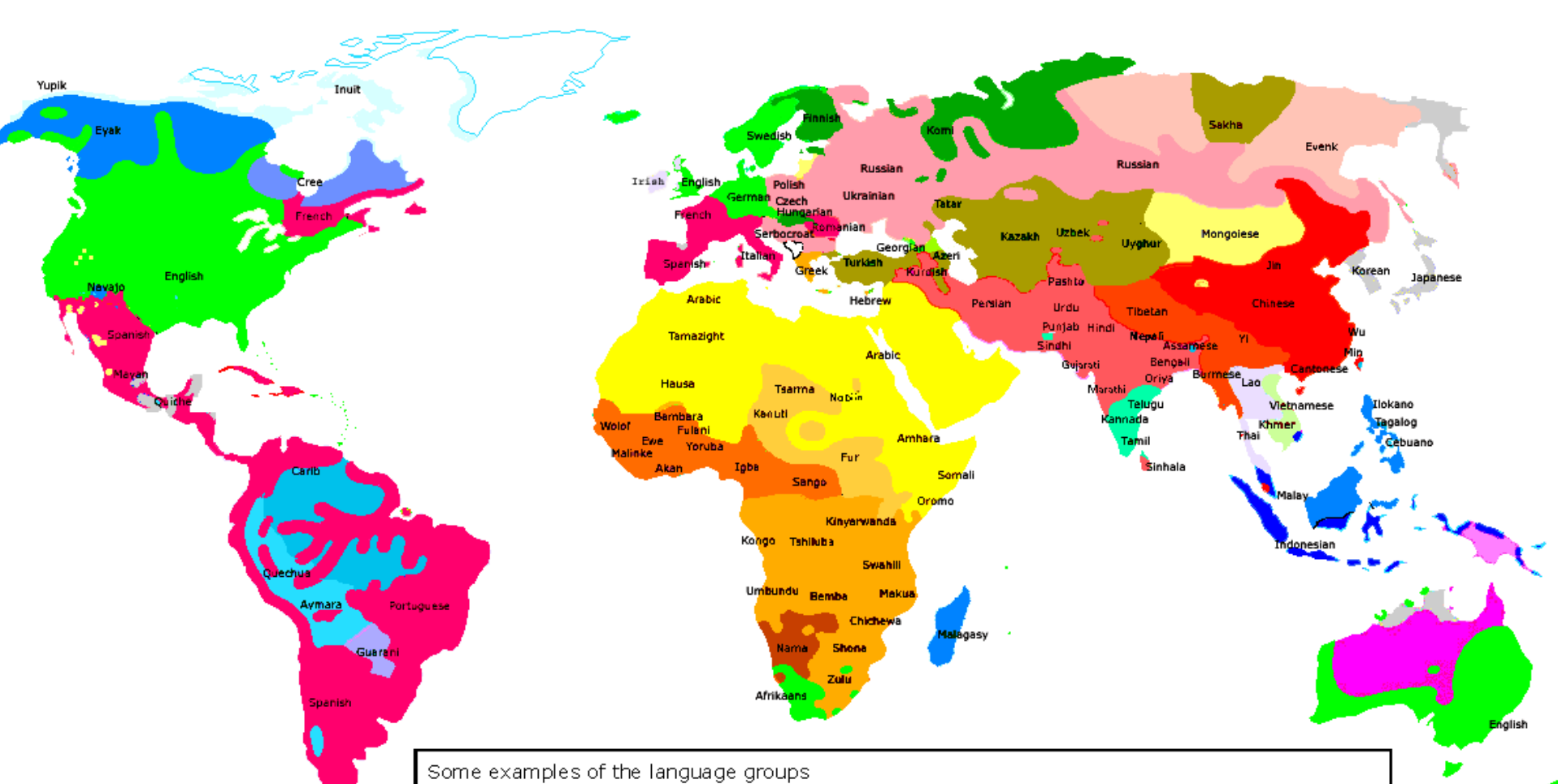
The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.

[www.answers.com/topic/capital-of-wyoming](#) - 21k - [Cached](#) - [Similar pages](#)

## [Cheyenne: Weather and Much More From Answers.com](#)

Chey·enne ( shī-ăn ' , -ěn ' ) The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.

[www.answers.com/topic/cheyenne-wyoming](#) - 74k - [Cached](#) - [Similar pages](#)



Some examples of the language groups

<ul style="list-style-type: none"> <li>■ Afro-Asiatic</li> <li>■ Niger-Congo</li> <li>■ Bantu</li> <li>■ Nilo-Saharan</li> <li>■ Khoisan</li> <li>■ Indo-European</li> <li>■ Germanic</li> <li>■ Albanic</li> <li>■ Romance</li> <li>■ Slavic</li> <li>■ Indo-Iranian</li> <li>■ Baltic</li> <li>■ Caucasian</li> </ul>	<ul style="list-style-type: none"> <li>■ Altaic</li> <li>■ Turkic</li> <li>■ Mongolic</li> <li>■ East Siberian languages</li> <li>■ Uralic</li> <li>■ Dravidian</li> <li>■ Sino-Tibetan</li> <li>■ Chinese</li> <li>■ Burmese-Tibetan</li> </ul>	<ul style="list-style-type: none"> <li>■ Austro-Asiatic</li> <li>■ Austronesian</li> <li>■ Borneo-Philippines/Formosan</li> <li>■ Nuclear Malayo-Polynesian</li> <li>■ Papan</li> <li>■ Pama-Ngyungan</li> <li>■ Tai-Kadal</li> <li>■ Isolate</li> </ul>	<ul style="list-style-type: none"> <li>■ Na-Déne</li> <li>■ Eskimo-Aleut</li> <li>■ American Indian</li> <li>■ Algonic</li> <li>■ Uto-Aztecan</li> <li>■ Mayan</li> <li>■ Andean</li> <li>■ Tupian</li> <li>■ Brazilian indigenous</li> </ul>
---	--	--	---

# 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 tibétaine : la Chine sur ses gardes



## "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
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (second half of this class) vs fidelity (later)

# 2013 Online Translation: French

EN CE MOMENT Impôts Kenya Syrie Pakistan Emploi Scandale Prism

## Impôt sur le revenu : vous en 2014 ?



Sélectionnez votre revenu et votre situation familiale pour bénéficier de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ? [VISUEL INTERACTIF](#)
- Un budget 2014 soumis aux critiques



**Le chômage baisse pour la première fois depuis avril 2011** [POST DE BLOG](#)

AT THIS MOMENT Taxes Kenya Syria Pakistan Use Prism scandal

## Income tax: how much do you pay in 2014?



Select your income and family situation to see if you get the tax break.

- How is the budget for 2014 is allocated? [INTERACTIVE VISUAL](#)
- Budget: these expenses no government can reduce
- A 2014 budget submitted to criticism
- Budget 2014: the retail savings [INTERACTIVE VISUAL](#)



**Unemployment fell for the first time since April 2011** [POST BLOG](#)



**Surviving in the Central time looting and anarchy**

DÉCOUVREZ TOUS LES **SERVICES ABONNÉS**

S'abonner au Monde à partir de 1 €



**CALL FOR EVIDENCE**

**Member (s) of Europe Ecology-Greens, do you share the finding of severe Christmas Mamère EELV?**

Share your experience

**Continuous**

- 7:53 Budget: the fixed expenses
- 7:36 Heard the "Fashion Week" in Paris
- 7:19 control giant Airbus
- 7:04 Complaint against "Actual Values"
- 7:01 Venezuela: 17 people arrested
- 6:59 Vidberg: the new budget came
- 6:50 The "noble mission" of the NSA
- 6:38 Roma: jousting between Brussels &

**DE  
FURSAC**

automne-hiver 13/14



# 2020 Online Translation: French

Le Monde

Le Monde



Consulter le journal



Consult the journal

Log in

Subscribe

ACTUALITÉS ÉCONOMIE VIDÉOS OPINIONS

NEWS ECONOMY VIDEOS REVIEWS CULTURE M THE MAG SERVICES

22:30

Netflix saisit la Cour suprême brésilienne

22:01

Dopage : le TAS saisit du dossier russe

21:18

Brexit : Johnson obtient le feu vert des députés

10:30 p.m.

Netflix seizes the Brazilian Supreme Court

10:01 p.m.

Doping: the CAS seizes the Russian file

9:18 p.m.

Brexit: Johnson gets green light from MEPs

9:07 p.m.

**Alerte**  
Boeing crash in Iran: Canada claims "plane shot down by Iranian missile"

7:46 p.m.

Retreats: story of the 4th day of mobilization

See more >

## Crash d'un Boeing en Iran : le Canada affirme que « l'avion a été abattu par un missile iranien »



Le premier ministre Justin Trudeau a affirmé disposer d'informations « de sources multiples ». « Ce n'était peut-être pas intentionnel », a-t-il ajouté.

- Le Canada réclame « une enquête approfondie » après le crash du Boeing à Téhéran
- Le crash d'un Boeing ukrainien peu après son décollage à Téhéran fait 176 morts



« Se mouve différe oppos. des rel nouve



Bol obtien député pour n



San Francisco Paris from \$ 189 - Lowest fare for direct flights

French bee, the cheapest international airline with a high standing on board us.frenchbee.com

TO OPEN

## Boeing crash in Iran: Canada claims "plane shot down by Iranian missile"



Prime Minister Justin Trudeau said he had information "from multiple sources". "It may not have been intentional," he added.

- Canada calls for "a thorough investigation" after the Boeing crash in Tehran
- Ukrainian Boeing crash shortly after takeoff in Tehran kills 176



"Only the duration of the movement will make a difference": opponents of pension reform once again on the street



Boris Johnson Gets Green Light From British MPs To Lead Brexit



SELECTION

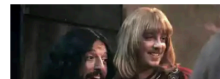
2020 on a set: the selection of shows from critics of the "World" Theater, dance, opera, humor: an overview of the most promising shows at the start of the year.

13 min read

"LE MONDE" AT 1 € FOR 3 MONTHS

Subscribe now and view the full articles.

Subscribe



# Why NLP

---

- To access information & knowledge
- To communicate

# Human-Machine Interactions



# Will this Be Part of All Our Home Devices?

amazon echo



FROM: AMAZON.COM

*Will it rain tomorrow?*

*Set an alarm for eight a.m.*

*Play music by  
Bruno Mars*

*How many teaspoons  
are in a tablespoon?*

*Add gelato to my  
shopping list*

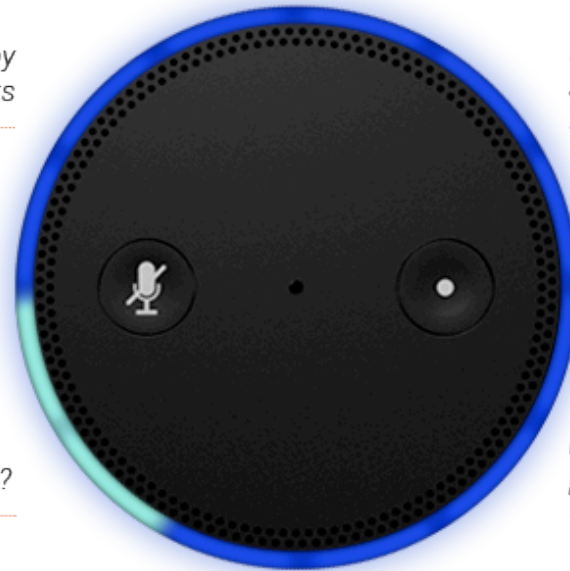
*Wikipedia: Abraham  
Lincoln*

*When is  
Thanksgiving?*

*Play my "dinner party"  
playlist*

*What's the weather in  
Los Angeles this weekend?*

*Add "make hotel reservations"  
to my to-do list*



[< PREV TEAM](#) | [VIEW ALL](#)

## University of Washington Sounding Board



**Sounding Board**



**Location:** Seattle, WA, USA  
**Faculty Advisor:** [Mari Ostendorf](#)

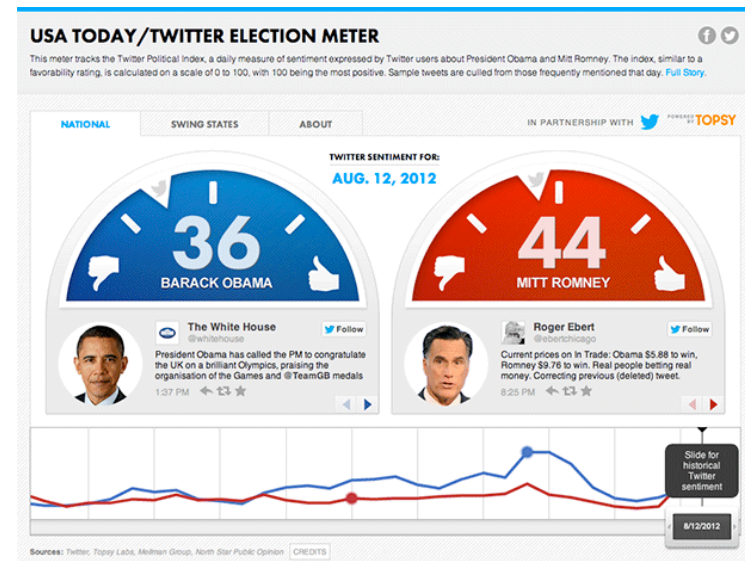
# Why NLP

---

- To access information & knowledge
- To communicate
- To understand our society

# Analyzing public opinion, making political forecasts

- 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)
- Challenge: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages



# Why NLP

---

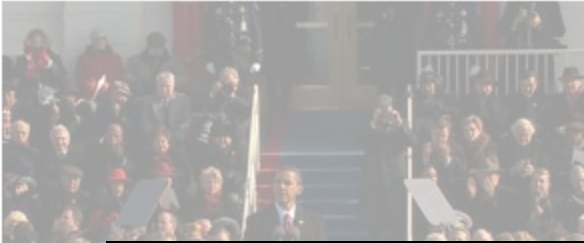
- To access information & knowledge
- To communicate
- To understand our society
- And to make our lives easier



# Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation

WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.



Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must begin

**STORY HIGHLIGHTS**

- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

aid in

President Obama renewed his call for a massive plan to stimulate economic growth.

his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

[Obama](#), too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their [inaugural address](#) to set out a bold agenda.

# Start-up Summly → Yahoo!

CEO Marissa Mayer announced an update to the app in a blog post, saying, "The new Yahoo! mobile app is also smarter, using Summly's natural-language algorithms and machine learning to deliver quick story summaries. We acquired Summly less than a month ago, and we're thrilled to introduce this game-changing technology in our first mobile application."



Launched 2011, Acquired 2013 for \$30M

# OpenAI has published the text-generating AI it said was too dangerous to share

*The lab says it's seen 'no strong evidence of misuse so far'*

By [James Vincent](#) | Nov 7, 2019, 7:24am EST

[f](#) [t](#) [SHARE](#)

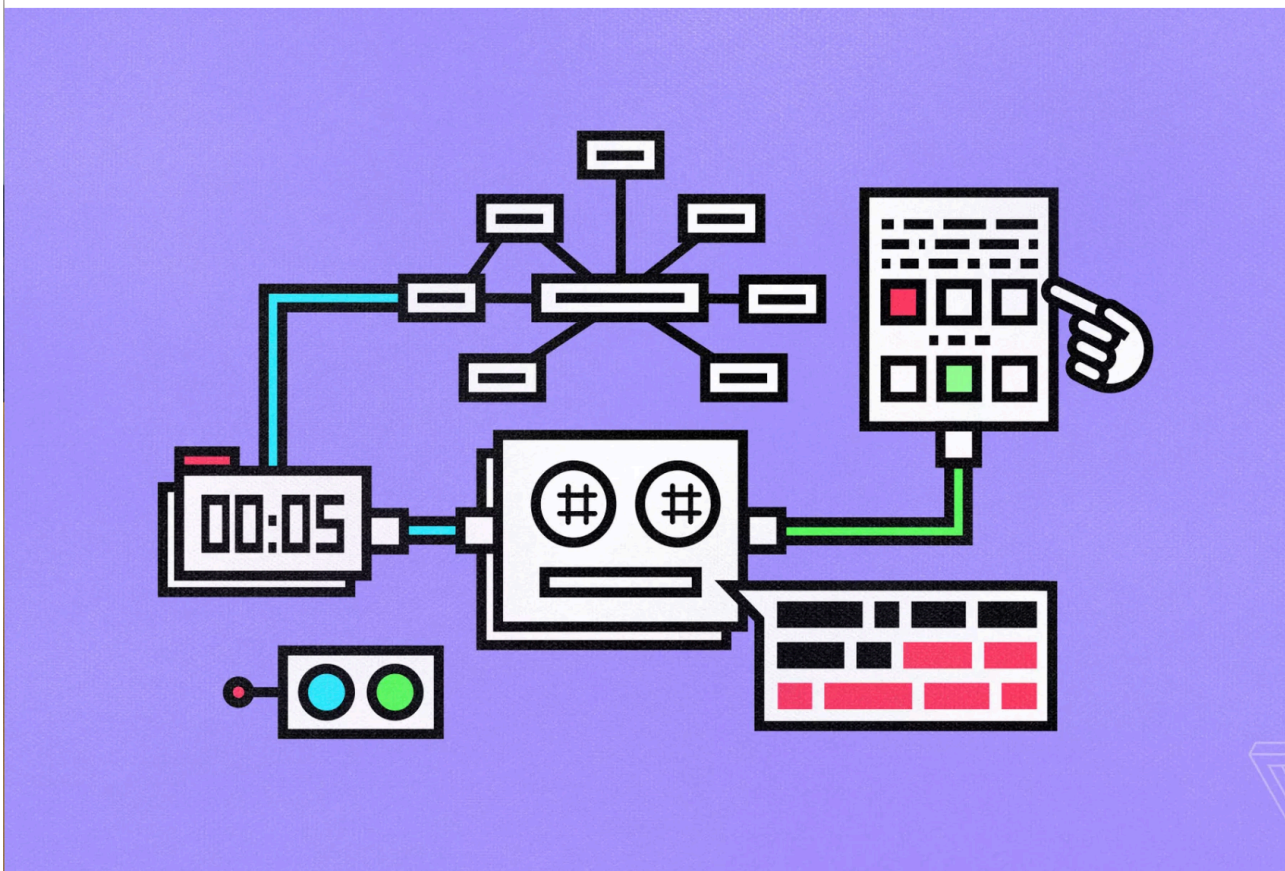


Illustration by Alex Castro / The Verge

# Why NLP

---

- To access information & knowledge
- To communicate
- To understand our society
- To make our lives easier
- NLP and AI

# Language Comprehension?

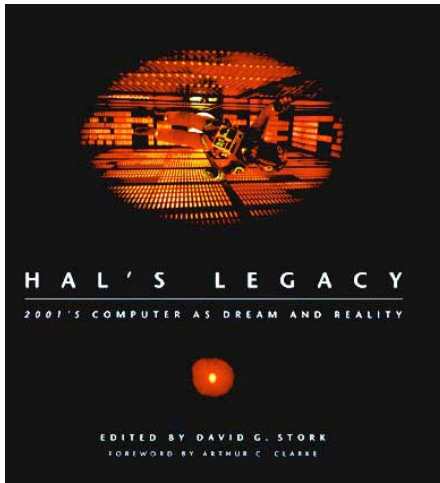
---

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xiangang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a *Naraoia* like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xiangang's "hands began to shake", because he was:

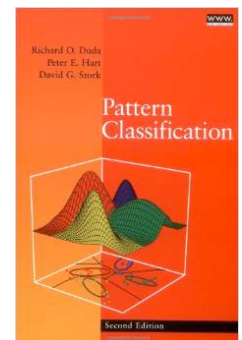
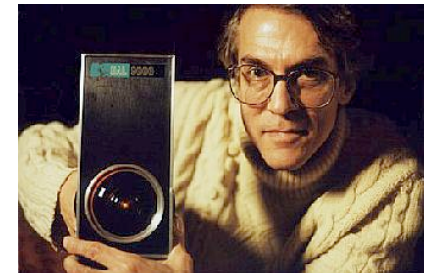
- (A) afraid that he might lose the fossil
- (B) worried about the implications of his finding
- (C) concerned that he might not get credit for his work
- (D) uncertain about the authenticity of the fossil
- (E) excited about the magnitude of his discovery

# 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



# What begins to work (e.g., Kuznetsova et al. 2014)



The flower was so  
**vivid and attractive.**



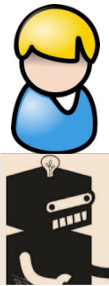
Blue flowers are **running rampant** in my garden.

We sometimes do well: 1 out of 4 times, machine captions were preferred over the original Flickr captions:



Spring in a white dress.

**Blue flowers have no scent. Small white flowers have no idea what they are.**



Scenes around the lake on my bike ride.

**This horse walking along the road as we drove by.**



# Table of Content

---

- Definition of NLP
- Historical account of NLP



# NLP History: pre-statistics

---

(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 grammaticality, 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)
- 70s and 80s: more linguistic focus
  - Emphasis on deeper models, syntax and semantics
  - Toy domains / manually engineered systems
  - Weak empirical evaluation

# NLP: machine learning and empiricism

---

“Whenever I fire a linguist our system performance improves.” –Jelinek, 1988

- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!

# NLP: deep learning / neural networks

---

“The idea of what an internal representation would look like was it would be some kind of symbolic structure. That has completely changed with these big neural nets.”

–Hinton, 2016

- ~2014-now: Neural networks
  - Big models, more data, less and less linguistic bias
  - Can be brittle to adversarial inputs
  - Can be difficult to interpret
- 2020s: What comes next?
  - Hybrid models? Just deeper networks?
  - You decide!!!

# 2019, the year of BERT....

---

- Train a big NN as a masked language model on \*lots\* of unlabeled data

**Input:** The man went to the [MASK]<sub>1</sub> . He bought a [MASK]<sub>2</sub> of milk .

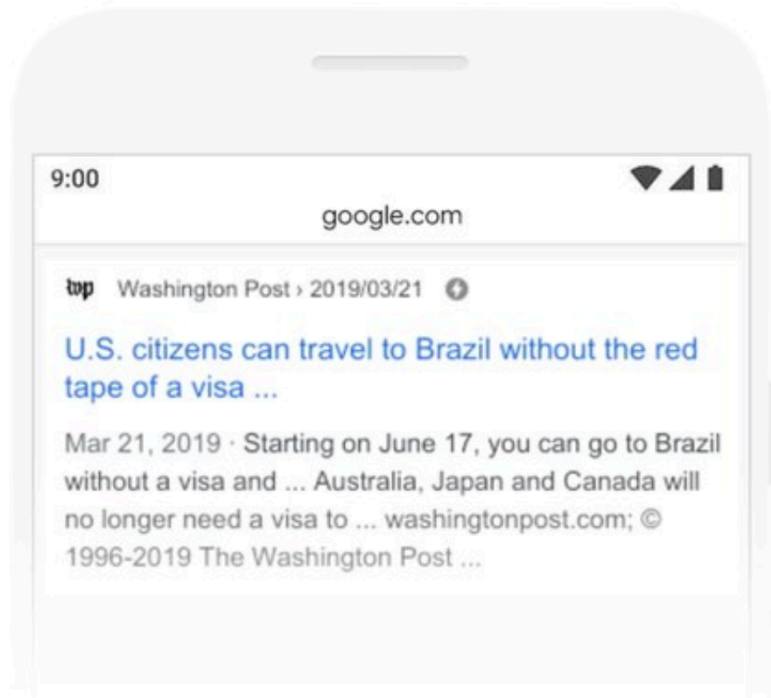
**Labels:** [MASK]<sub>1</sub> = store, [MASK]<sub>2</sub>=gallon

- Fine tune for end task with labeled data
- Over 3,000 citations in first year alone...

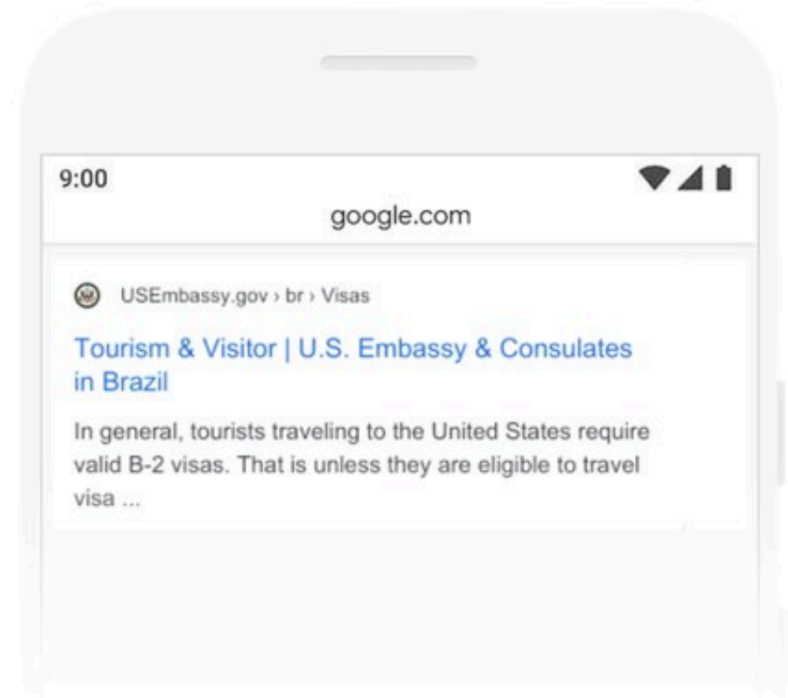
# BERT is in Google Search!

🔍 2019 brazil traveler to usa need a visa

BEFORE



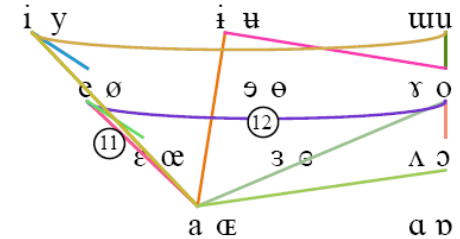
AFTER



# What is Nearby NLP?

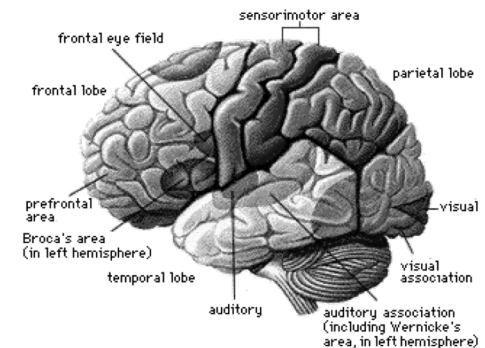
## ■ Computational Linguistics

- Using computational methods to learn more about how language works
- We end up doing this and using it



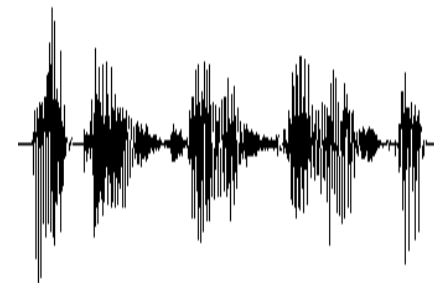
## ■ Cognitive Science

- Figuring out how the human brain works
- Includes the bits that do language
- Humans: the only working NLP prototype!



## ■ 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



# Table of Content

---

- Definition of NLP
- Historical account of NLP
- Unique challenges of NLP

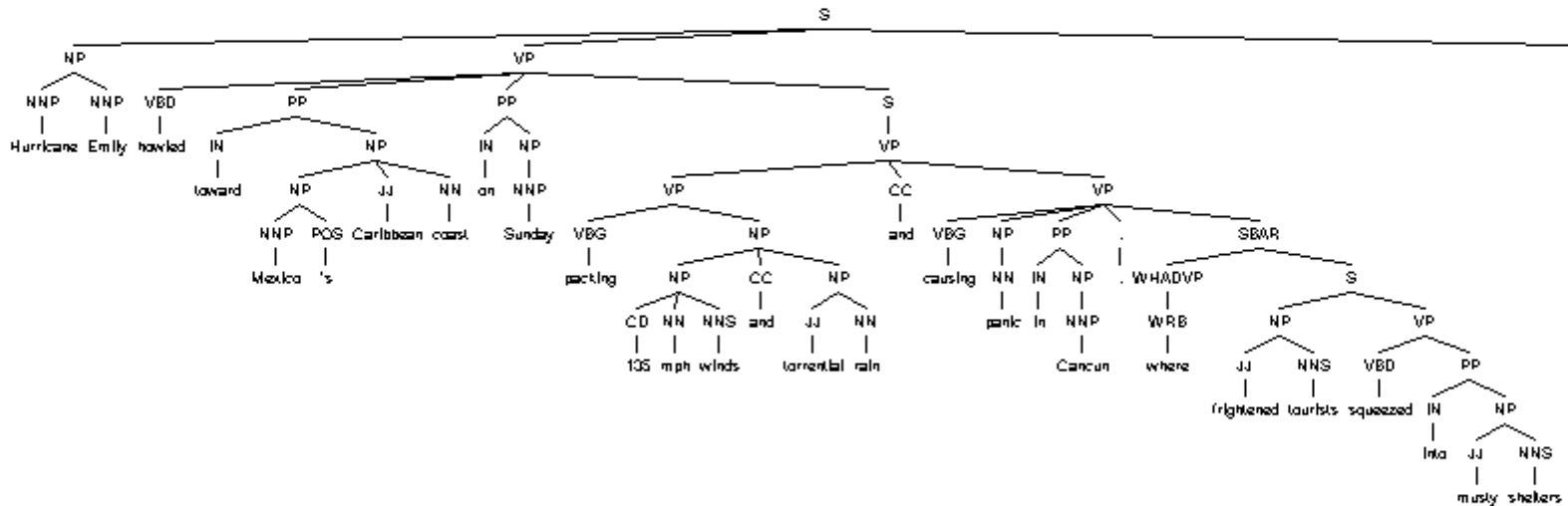
# Problem: Ambiguities

---

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Ban on Nude Dancing on Governor's Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?



# Syntactic Analysis



Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- **SOTA:** ~95% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

# Semantic Ambiguity

---

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

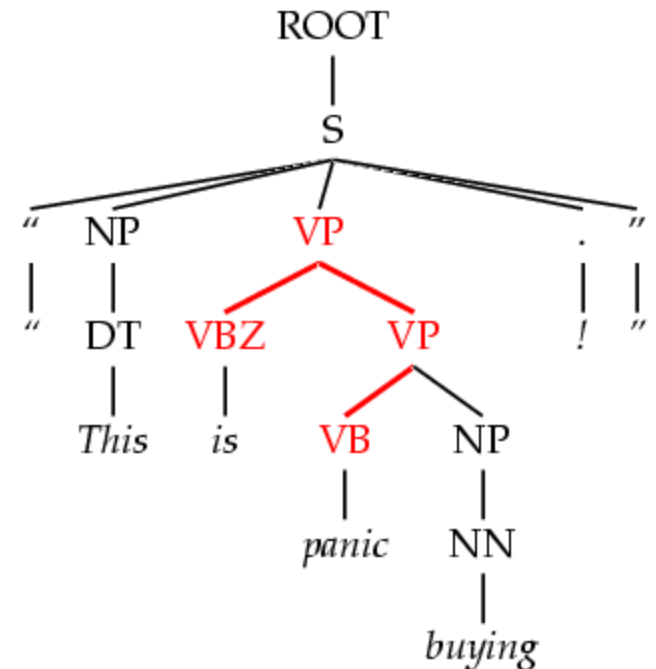
- **Direct Meanings:**
    - It understands you like your mother (does) [presumably well]
    - It understands (that) you like your mother
    - It understands you like (it understands) your mother
  - **But there are other possibilities, e.g. mother could mean:**
    - a woman who has given birth to 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 L. Lee]

# Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds to the correct parse of

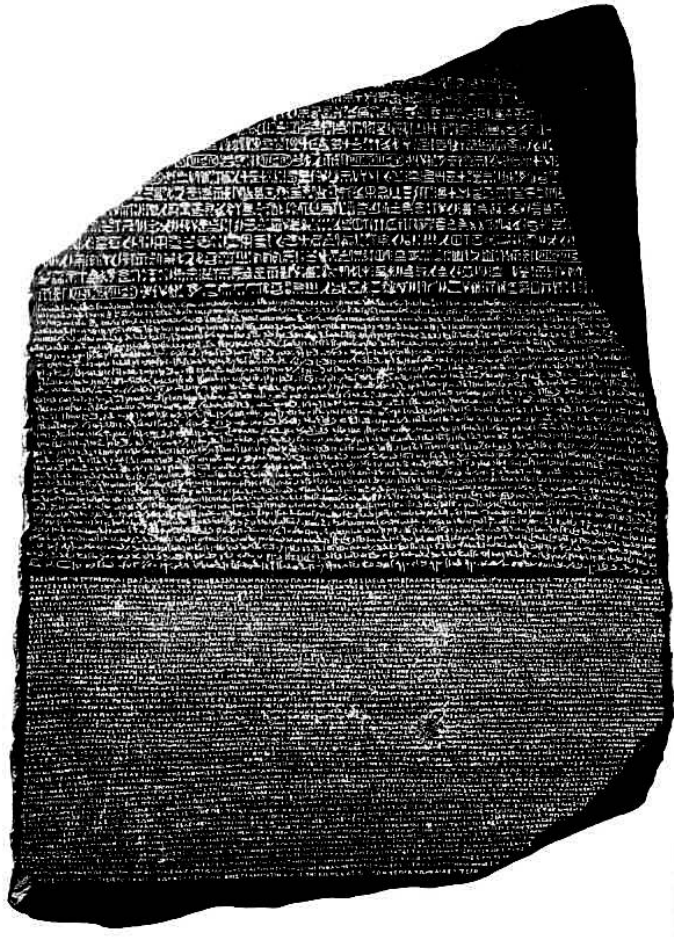
*“This will panic buyers ! ”*



- Unknown words and new usages
- **Solution**: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

# Corpora

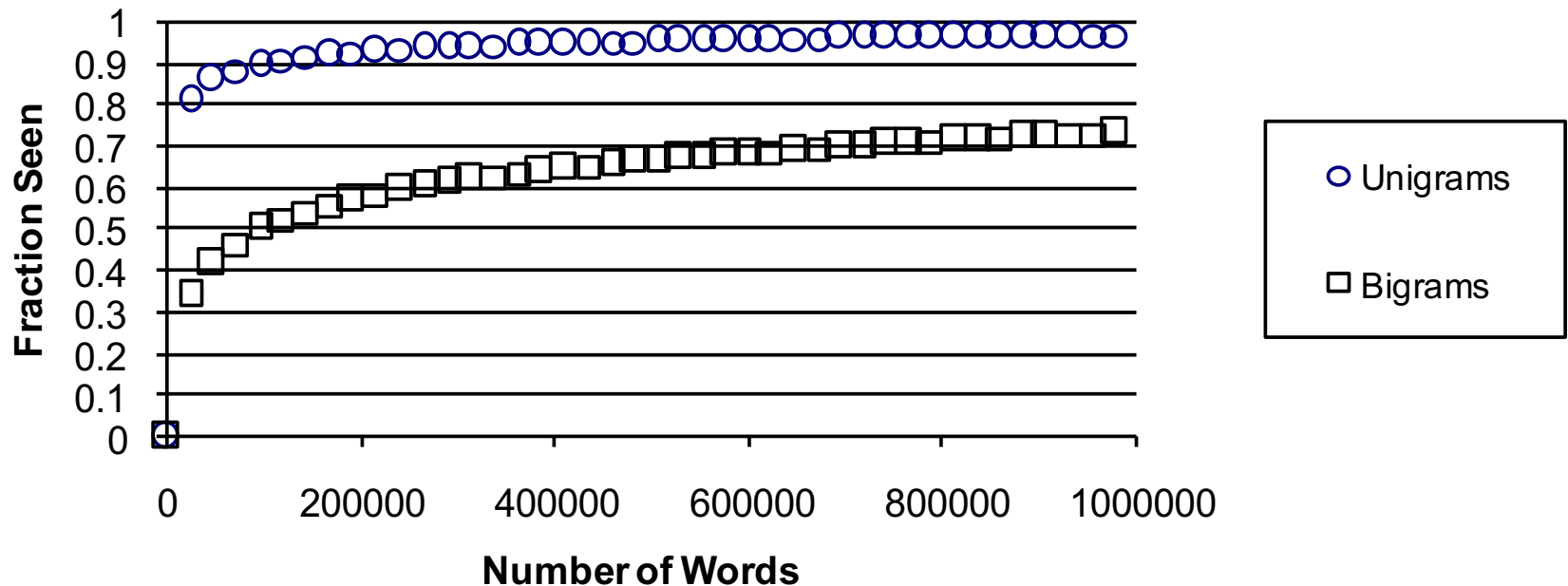
---



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

# Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair)



# Table of Content

---

- Definition of NLP
- Historical account of NLP
- Unique challenges of NLP
- Class administrivia / discussion

# What is this Class?

---

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice...

# Comparisons with Other Classes

---

- Compared to ML
  - Typically multivariate, dynamic programming everywhere
  - Structural Learning & Inference
  - Insights into language matters (a lot!)
  - DL: RNNs, LSTMs, Seq-to-seq, Attention, ...
- Compared to undergrad NLP
  - Faster paced
  - Stronger engineering skills & higher degree of independence assumed
- Compared to CompLing classes
  - More focus on core algorithm design, technically more demanding in terms of math, algorithms, and programming



# Class Requirements and Goals

---

## ■ Class requirements

- Uses a variety of skills / knowledge:
  - Probability and statistics
  - Basic linguistics background
  - Decent coding skills
- Most people are probably missing one of the above
- You will often have to work to fill the gaps

## ■ Class goals

- Learn the issues and techniques of modern NLP
- Build realistic NLP tools
- Be able to read current research papers in the field
- See where the holes in the field still are!