

CS5740: Natural Language Processing

Introduction

Instructor: Yoav Artzi

TA: Max Grusky

Technicalities

- People:
 - Instructor: Yoav Artzi
 - Office hours: Monday 5pm, Baron
 - TA: Max Grusky:
 - Office hours: Thursday, 1pm, by Skype (coordinate)
- Webpage (everything is there):
 - <http://www.cs.cornell.edu/courses/cs5740/2017sp/>
- ~~• Discussion group on Piazza~~
- Chat on Slack
- Assignments on CMS
 - Repositories on Github Classroom

Technicalities

- Grading:
 - 40% assignments, 25% exam, and 30% class review quizzes, 5% participation
 - Participation = class + ~~Piazza~~ + Slack
- Enrollment and prerequisites:
 - At least B in CS 5785 (Applied ML) or equivalent Cornell Course
 - Or: instructor permission
 - Audit? Talk to me after class

Technicalities

- Quizzes:
 - First five minutes of every class, no extensions
 - Each quiz: 1.5% of the grade, up to 30%, only top 20 quizzes count
 - It is not possible to re-take a missed quiz
 - A missed quiz gets zero
 - Just like an exam: no copying, chatting, and not taking the quiz remotely → all AI violations
- Quiz practice
 - Phones and laptops
 - <http://socrative.com>
 - Use NetID to identify
 - Today's room: NLP5

Technicalities

- Collaboration:
 - All assignments must be done in pairs
- Use of external code/tools – specified in each assignment
 - If have doubt – ask!
- Late submissions:
 - 10% off for every 12 hours, rounded up
 - E.g., 25 hours late → grade starts at 70
 - No late submission for final exam
- All assignments should be implemented in Python

Technicalities

- Books (recommended, not required):
 - D. Jurafsky & James H. Martin, Speech and Language Processing
 - C.D. Manning & H. Schuetze, Foundations of Statistical Natural Language Processing
- Other material on the course website

Technicalities

- Come on time
 - Late? Enter quietly and sit at the back
 - Quiz starts on time
- No laptops or phones in class
 - Except during the quiz

WHY ARE YOU HERE?

What is this class?

- Depth-first technical NLP course
- Learn the language of natural language processing
- What this class is not?
 - It is not a tutorial to NLTK, TensorFlow, etc.
 - Stack Overflow already does this well

Class Goals

- Learn about the issues and techniques of modern NLP
- Be able to read current research papers
- Build realistic NLP tools
- Understand the limitation of current techniques

Main Themes

- Linguistic Issues
 - What are the range of language phenomena?
 - What are the knowledge sources that let us make decisions?
 - What representations are appropriate?
- 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 ...

Main Models

- Generative Models
- Discriminative Models
 - Neural Networks
- Graphical Models

What is NLP?



- **Fundamental goal:** deep understanding of broad language
 - Not just string processing or keyword matching!
- **End systems that we want to build:**
 - Simple:
 - Complex:

What is NLP?



- **Fundamental goal:** deep understanding 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, dialog interfaces, question answering...
 - Unknown: human-level comprehension (is this just NLP?)

Today

- Prominent applications
 - Try to imagine approaches
 - What's behind current limitations?
- Some history
- Key problems

Machine Translation



L'économie japonaise sort du rouge pour la première fois depuis Fukushima

Après avoir atteint un déficit record en 2014, le Japon dégage un excédent commercial pour la première fois depuis l'accident nucléaire de 2011.



Japan's economy turns red for the first time since Fukushima

After reaching a record deficit in 2014, Japan posted a trade surplus for the first time since the 2011 nuclear accident.

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
 - What fragments? How to combine? [learning to translate]
 - How to make efficient? [fast translation search]

Machine Translation

Le Monde.fr

La Bourse de Shanghai dégringolait de plus de 6 % mardi 25 août à l'ouverture, après s'être déjà effondrée de presque 8,5 % la veille, dans un marché affolé par l'affaiblissement persistant de l'économie chinoise et miné par des inquiétudes sur la conjoncture mondiale.

Dans les premiers échanges, l'indice composite chutait de 6,41 % soit 205,78 points à 3 004,13 points. La Bourse de Shenzhen plongeait quant à elle de

The Shanghai Stock Exchange tumbled more than 6% Tuesday, August 25 at the opening, having already collapsed by almost 8.5% yesterday, in a panicked market the persistent weakening of the Chinese economy and undermined by concerns about the global economy.

In early trade, the composite index fell by 6.41% or 205.78 points to 3 004.13 points. The Shenzhen Stock Exchange dived for its 6.97% to 1 751.28 points. The Hong Kong Stock Exchange, meanwhile, opened down 0.67%.

Machine Translation

纽约时报中文网 国际纵览

The New York Times Beta

A股跌势蔓延全球

周一美股开盘大跌1000点

NATHANIEL POPPER, NEIL GOUGH 09:54

周一，A股市场下跌8.5%，回吐今年全部涨幅。投资者担心中国经济下滑失控，股市“黑色星期一”波及美欧和亚洲市场，道指开盘数分钟内下跌过千点。

A spread of global stocks decline

US stocks opened Monday fell 1,000 points

NATHANIEL POPPER, NEIL GOUGH 09:54

Monday, A-share market fell 8.5 percent, taking all the gains this year. Investors worried about the economic downturn runaway Chinese stock market "Black Monday" spread to the US and European and Asian markets, the Dow opened down over a thousand points within minutes.

Machine Translation

lrytas.lt

English Spanish French Lithuanian - detected ▾



English Spanish Arabic ▾

Translate

Kiek Lietuvoje kainuoja užsienyje vogtas dviratis? Kokiais keliais jie čia patenka ir kodėl policija pro pirštus žiūri į klestinčią prekybą vogtais daiktais? Atsakymų į šiuos bei kitus klausimus ieškojo Lietuvoje viešėjusi Danijos valstybinės televizijos „DR“...



As far as Lithuania free bike stolen abroad? In what ways are placed here and why the police connive at a thriving trade in stolen items? Answers to these and other questions put Lithuania who visited the Danish public television DR ...



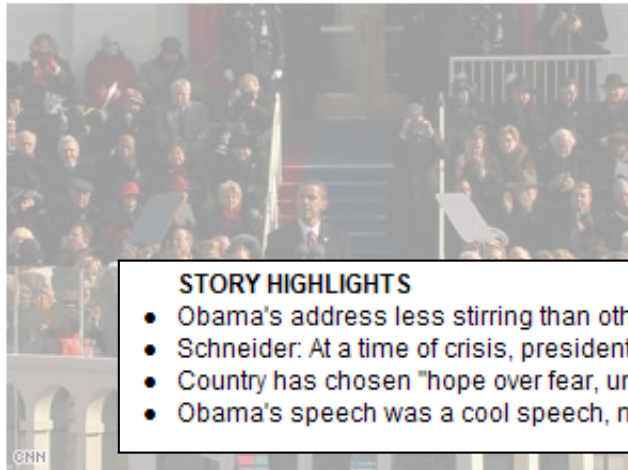
Wrong?



Summarization

- Condensing documents
 - Single or multiple docs
 - Extractive or abstractive
- Very context-dependent!

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 it was also more candid and down-to-earth.

"Starting today," the new president said, "we are beginning a new chapter in American history."

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

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

[more photos >](#)

his first inaugural in 1933, "The only thing wrong with America is fear. 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 arms. 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.

Information Extraction

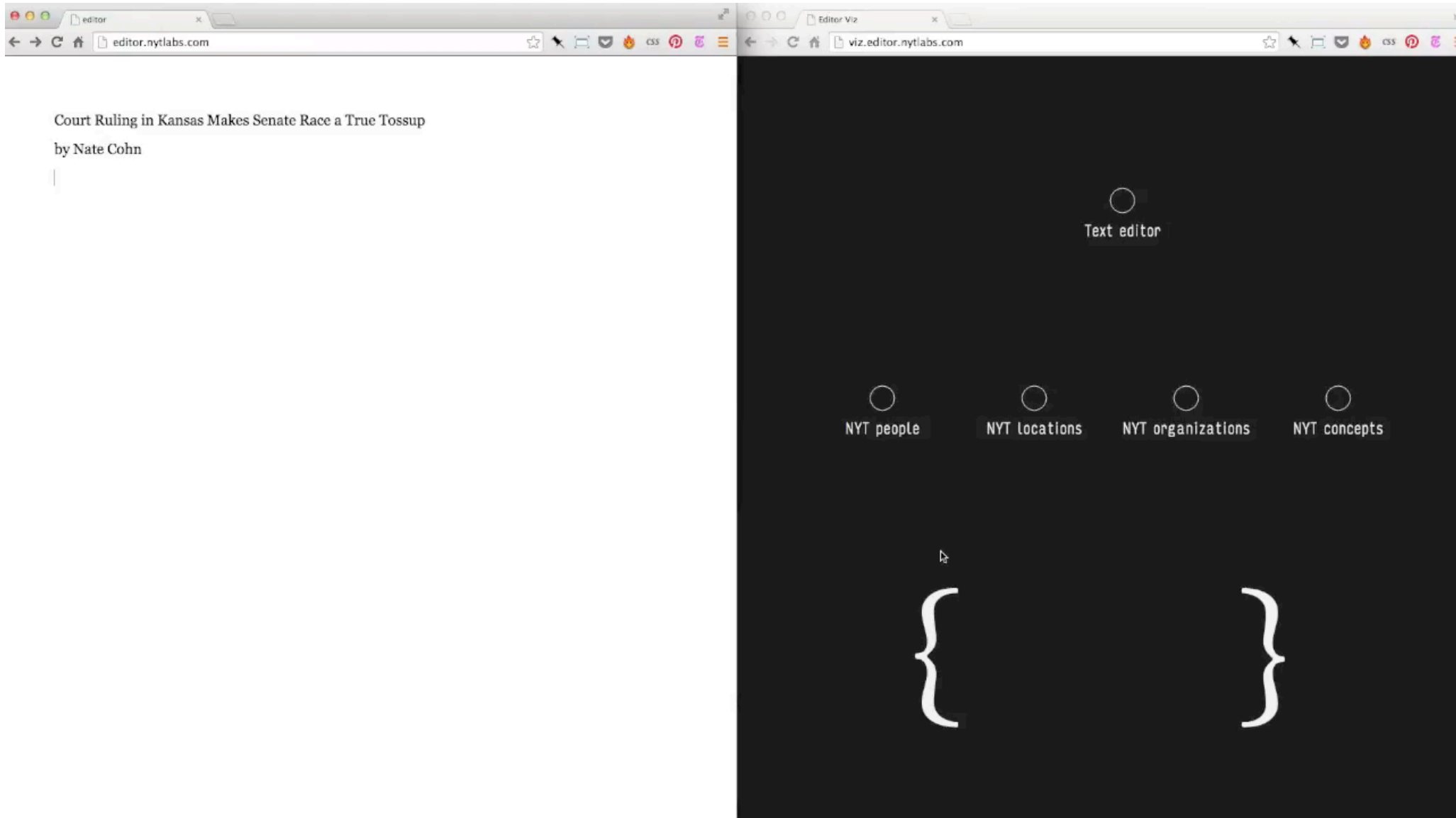
- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: good performance on simple templates (e.g., person-role)
- Harder without defining template

Tagging: Back to Text



Question Answering



[>](#)

The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

[>](#)

Question Answering

- More than search

What's the capital of Wyoming?



Web

Maps

Shopping

Images

News

More ▾

Search tools

About 984,000 results (0.54 seconds)

Wyoming / Capital



Cheyenne

Question Answering

- More than search

How many US states' capitals are also their largest cities?



Web

Images

News

Shopping

Videos

More ▾

Search tools

About 982,000,000 results (0.67 seconds)

State Capitals and Largest Cities - Infoplease

www.infoplease.com › [United States](#) › [States](#) ▾

State Capitals and Largest Cities. The following table lists the **capital** and **largest city** of every **state** in the **United States**. **State, Capital, Largest city.**

State Capitals and Largest Cities - Fact Monster

www.factmonster.com › [United States](#) › [States](#) ▾ [Fact Monster](#) ▾

State Capitals and Largest Cities. The following table lists the **capital** and **largest city** of every **state** in the **United States**. **State, Capital, Largest city.**

List of capitals in the United States - Wikipedia, the free ...

https://en.wikipedia.org/.../List_of_capitals_in_the_United_Sta... ▾ [Wikipedia](#) ▾

Austin is the largest **state capital** that is not **also** the **state's largest city**. The Confederate **States** of **America** had two **capitals** during **its** existence. The first ... In **many cases**, former **capital** cities of **states** are outside the current **state** borders.

[State capitals](#) - [Insular area capitals](#) - [Former national capitals](#)

Question Answering

- More than search

What are the main issues in the global warming debate?



Web

News

Images

Videos

Shopping

More ▾

Search tools

About 79,300,000 results (0.36 seconds)

[Global warming controversy - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Global_warming_controversy)

https://en.wikipedia.org/wiki/Global_warming_controversy ▾ Wikipedia ▾

Jump to [Mainstream scientific position](#), and [challenges](#) to it - [edit]. **Main article:**

Scientific opinion on **climate change**. Summary of opinions from climate ...

[Climate Change ProCon.org](https://climatechange.procon.org/)

climatechange.procon.org/ ▾ ProCon.org ▾

The pro side argues rising levels of atmospheric greenhouse gases are a direct result of human activities such as burning fossil fuels, and that these increases are causing significant and increasingly severe **climate** changes including global warming, loss of sea ice, sea level rise, stronger storms, and more droughts.

[Is human activity a substantial](#) - [Footnotes & Sources](#) - [Carbon Dioxide \(CO2\)](#)

[Climate Change and Global Warming — Global Issues](http://www.globalissues.org/issue/178/climate-change-and-global-warming)

www.globalissues.org/issue/178/climate-change-and-global-warming ▾

Some of the **major** conferences in recent years are also discussed. 32 articles on "**Climate Change and Global Warming**" and 1 related **issue**: ...

Question Answering

WolframAlpha computational... knowledge engine

Oscar for best actress 1958

Assuming year of award ceremony | Use year of film release instead

Input Interpretation:

Academy Awards | actress in a leading role | 1958 (year of award ceremony)

Result:

Joanne Woodward in *The Three Faces of Eve*

Other nominees:

Lana Turner in *Peyton Place* | Elizabeth Taylor in *Raintree County* | Deborah Kerr in *Heaven Knows, Mr. Allison* | Anna Lee in *The Wind*

Information about Joanne Woodward:

full name	Joanne Gignilliat Trimmier Woodward
date of birth	Thursday February 27, 1930 (age: 82 years)
place of birth	Thomasville, Georgia, United States

AT&T 3:06 PM

“What's the best movie to see this weekend”

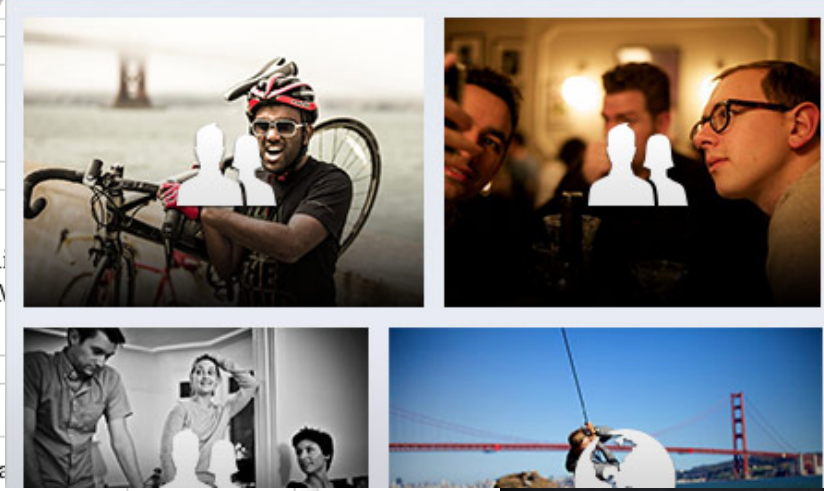
That would probably start an argument. But here's a list of highly-regarded movies:

MOVIES

Y NORTHWEST
y 17, 1959 100%

URE OF THE SIERRA...
January 6, 1948 100%

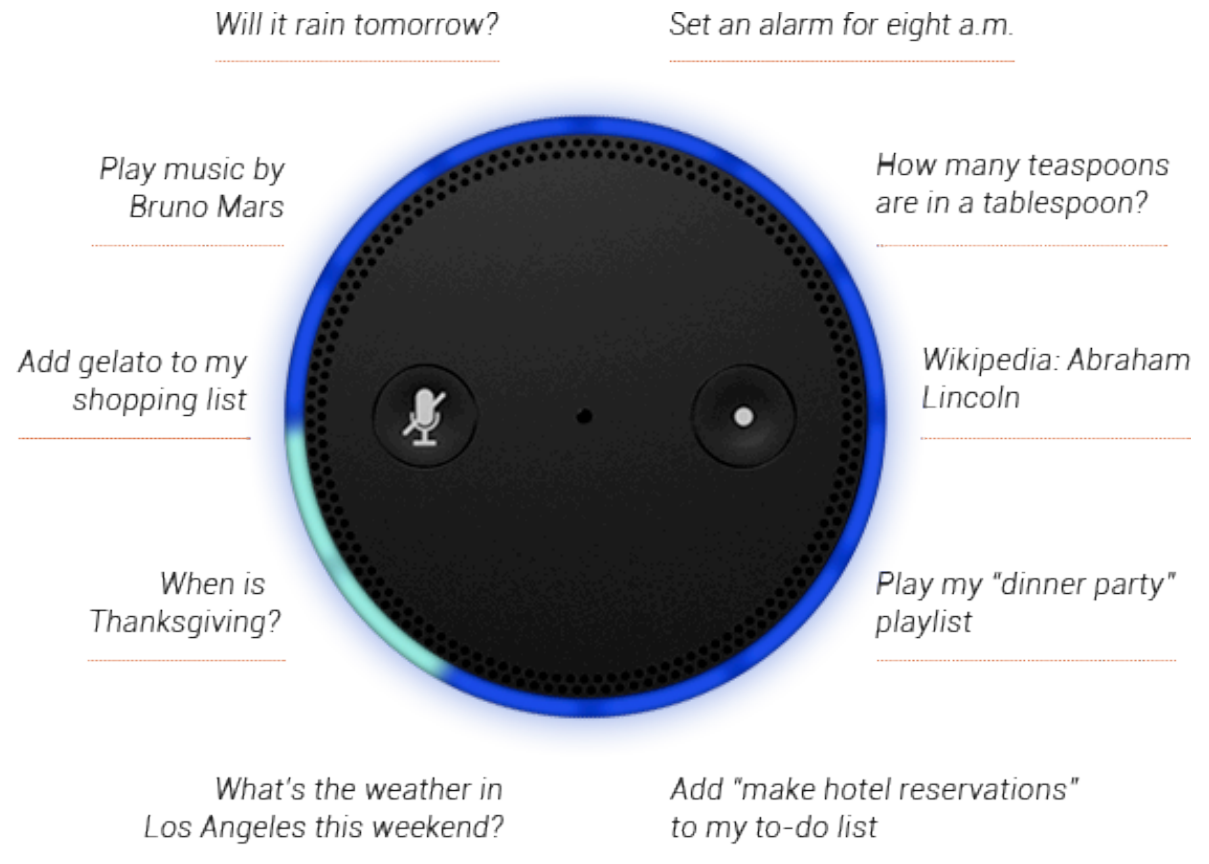
Photos of my friends



Natural Language Instruction



- What makes this possible?
- Limitations?



Language Comprehension

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. 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.

Language Comprehension

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. 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 Xianguang'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 Comprehension

Bang, bang, his silver hammer came down upon her head

PIENSE
THINK
SMACHIS
ΣΚΕΨΟΥ
DENKE
PENSER

\$200
Ken

\$4,000
WATSON

\$600
BRAD

Maxwell's silver hammer

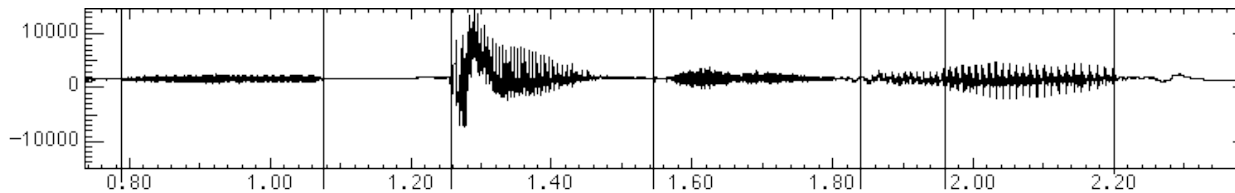
FRANK SINATRA	96%
Brown	11%
Brown	7%



Speech Systems



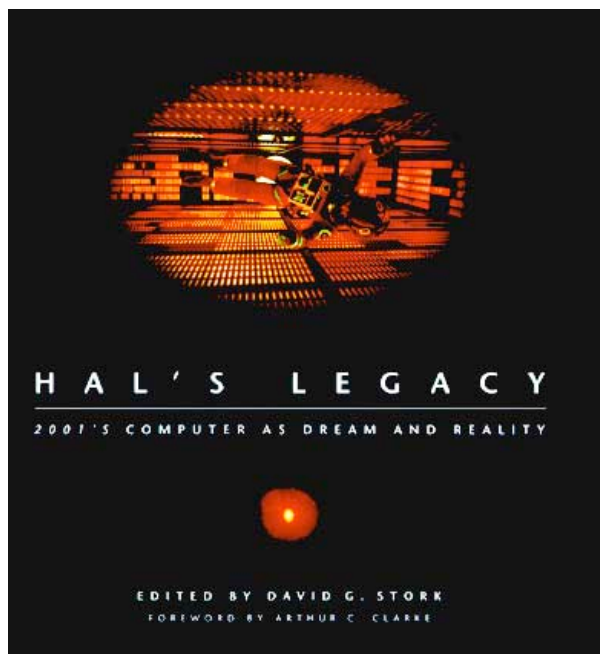
- Automatic Speech Recognition (ASR)
 - Audio in, text out
 - SOTA: 16% PER, Google claims 8% WER



“speech lab”

- Text to Speech (TTS)
 - Text in, audio out
 - SOTA: mechanical and monotone

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

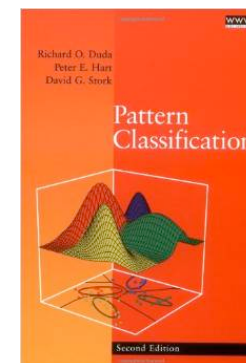
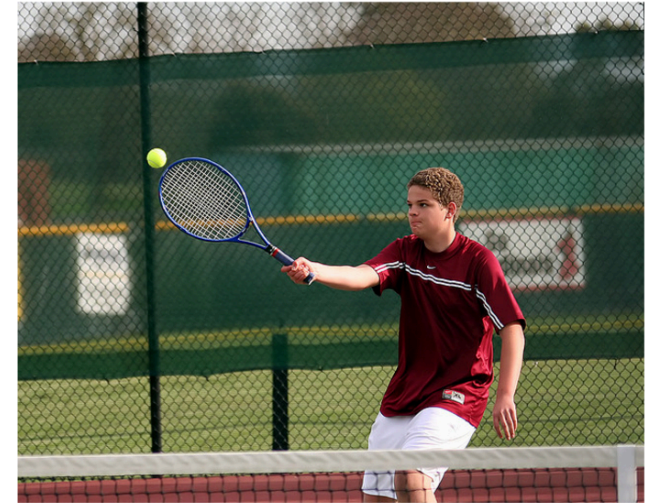


Image Captioning: The Good



a bunch of bananas sitting on top of a wooden table
logprob: -8.52



a man is playing tennis on a tennis court
logprob: -6.77

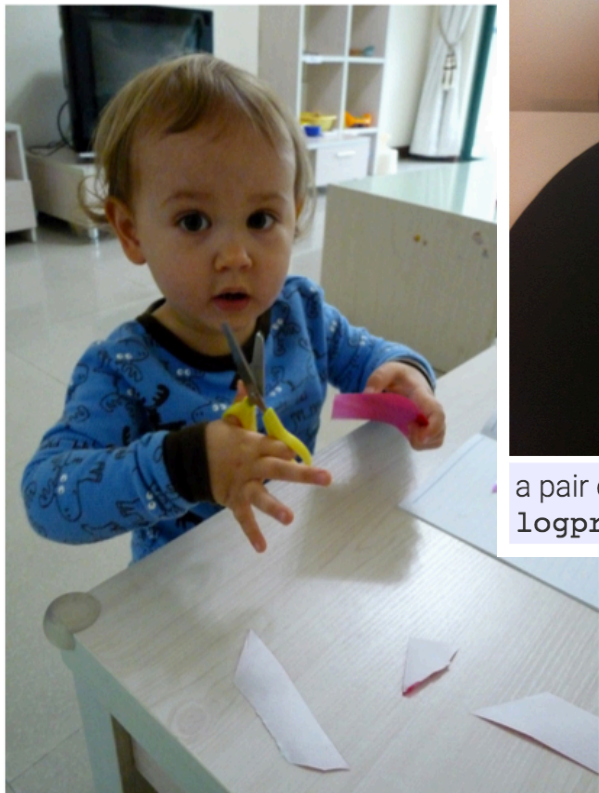


a train traveling down tracks next to a lush green field
logprob: -7.65



a pizza with toppings on a white plate
logprob: -7.40

Image Captioning: The Bad



a young boy is eating a
piece of cake
logprob: -7.75



a pair of scissors with a pair of scissors
logprob: -9.07



a man is holding a cell
phone in his hand
logprob: -8.90



a large jetliner flying through a blue sky
logprob: -5.79

Image Captioning: The Sitting



a cat is sitting on a toilet seat
logprob: -7.95



a pizza **sitting** on top of a white plate
logprob: -6.15



a laptop computer sitting on top of a wooden desk
logprob: -6.38



a bunch of luggage **sitting** on top of a hard wood floor
logprob: -10.50



a large airplane **sitting** on top of an airport runway
logprob: -6.70



a group of people sitting around a table with a cake
logprob: -8.83

NLP History: Pre-statistics

- (1) Colorless green ideas sleep furiously.
- (2) Furiously sleep ideas green colorless

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 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)

NLP History: Pre-statistics

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

NLP History: ML 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

NLP History: ML 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

“Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, “Every time I fire a linguist, my performance goes up”). The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high-performance broad-domain language understanding systems; for instance, in the speech recognition setting described above, a better understanding of language structure can lead to better language models.”

- Lillian Lee (2001) <http://www.cs.cornell.edu/home/llee/papers/cstb/index.html>

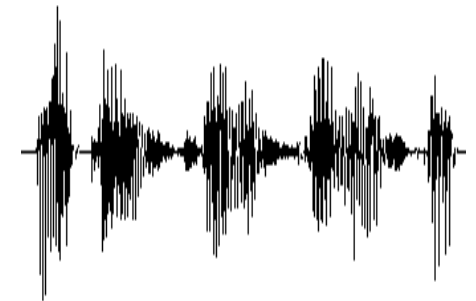
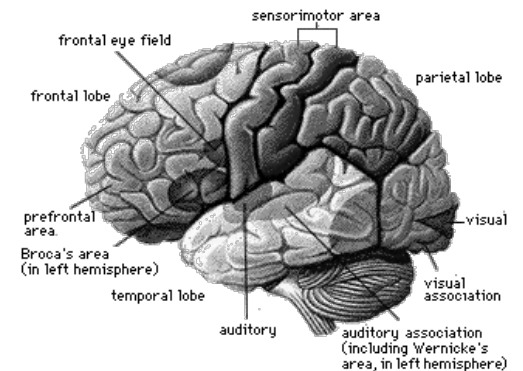
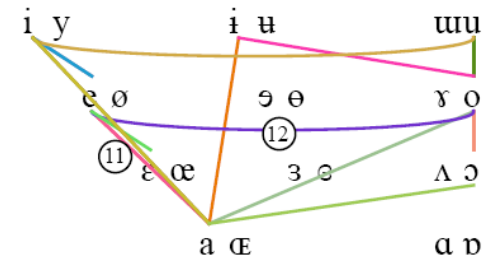
NLP History: ML and Empiricism

“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!
- 2010s: you decide!

Related Fields

- 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



Key Problems

We can understand programming languages.
Why is NLP not solved?

Key Problems

We can understand programming languages.
Why is NLP not solved?

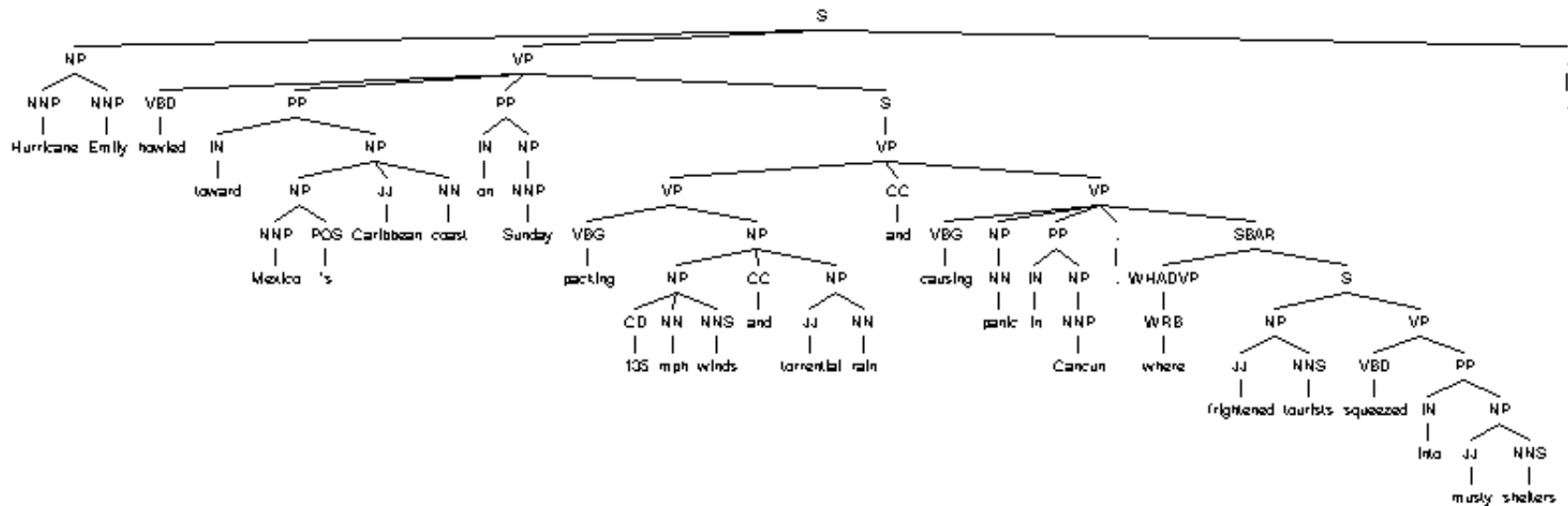
- Ambiguity
- Scale
- Sparsity

Key Problem: Ambiguity

- Some 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

Syntactic Ambiguity

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: ~90% 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.

Semantic Ambiguity

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

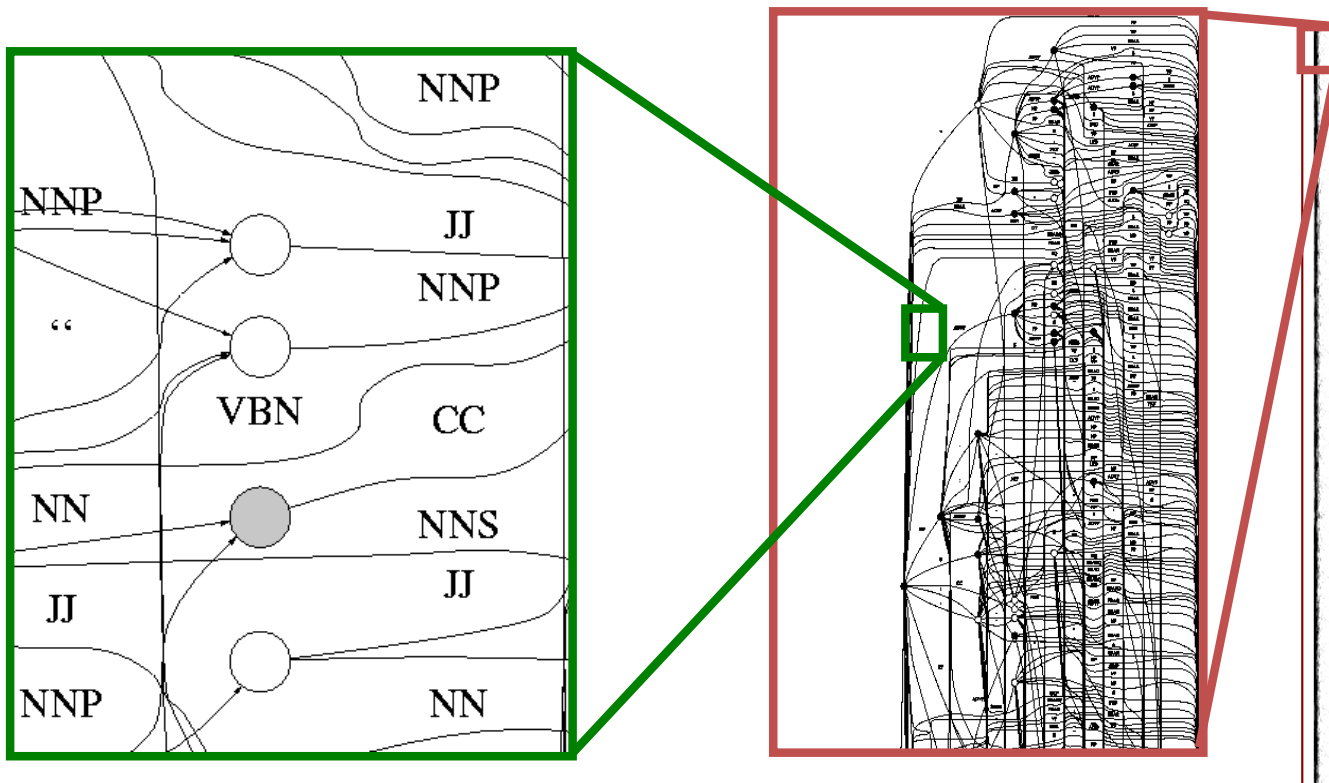
- Direct Meanings:
 - It understands you like your mother (does) [presumably well]
 - It understands (that) you like your mother
 - It understands you like (it understands) your mother
- But there are other possibilities, e.g. mother could mean:
 - a woman who has given birth to a child
 - a stringy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

Key Problem: Scale

- People *did* know that language was ambiguous!
 - ...but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)

Key Problem: Scale

- People *did* know that language was ambiguous!
 - ...but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
 - ...they didn’t realize how bad it would be



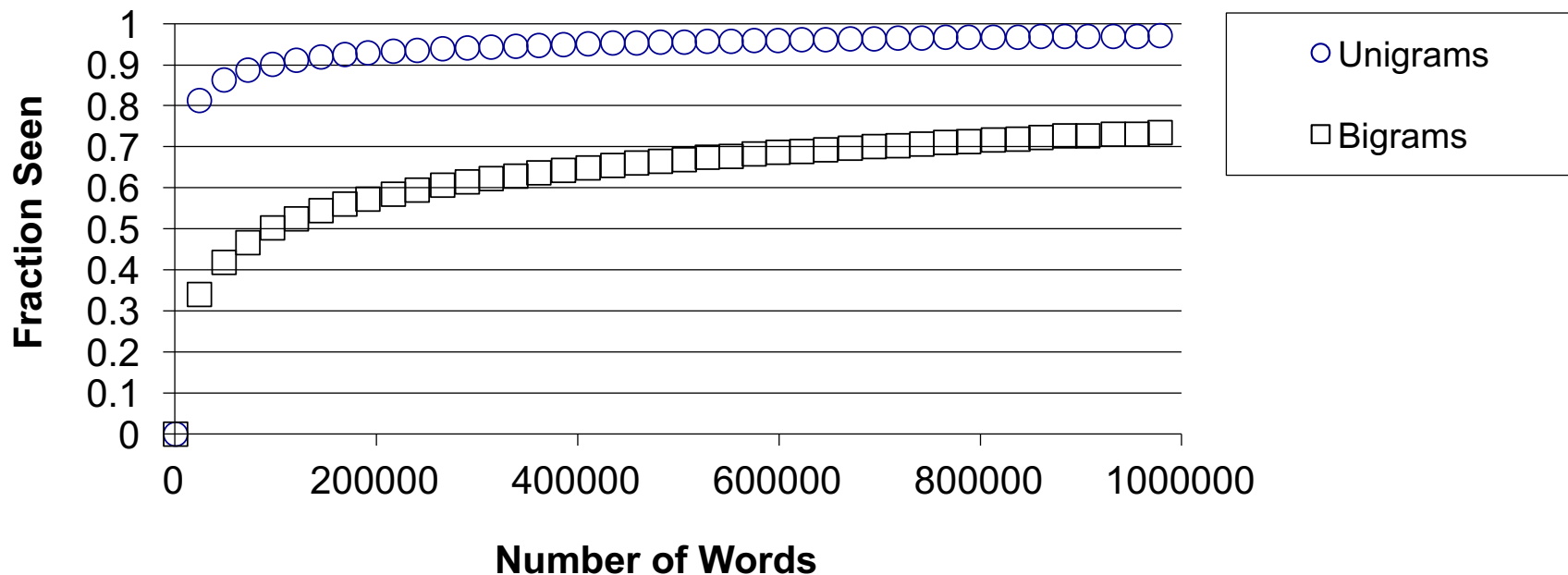
Key Problem: Sparsity



- 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

Key Problem: Sparsity

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



The NLP Community

- Conferences: **ACL**, **NAACL**, **EMNLP**, EACL, CoNLL, COLING, *SEM, LREC, CICLing, ...
- Journals: CL, **TACL**, ...
- Also in AI and ML conferences: AAAI, IJCAI, ICML, NIPS