

CSEP 517: Natural Language Processing

New PMP Course!

Instructor: Luke Zettlemoyer

Autumn 2013

Slides adapted from Dan Klein

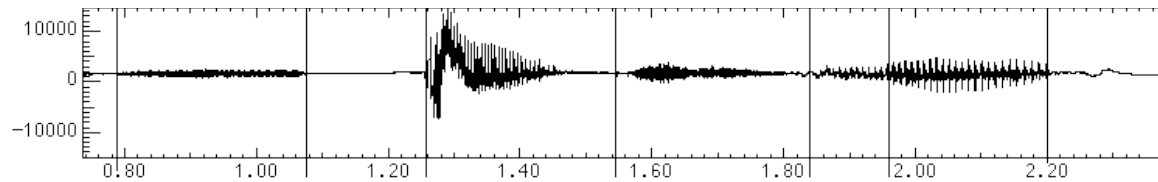
What is NLP?



- 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, dialog interfaces, question answering...
 - Unknown: human-level comprehension (is this just NLP?)

Speech Systems

- **Automatic Speech Recognition (ASR)**
 - Audio in, text out
 - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



“Speech Lab”

- **Text to Speech (TTS)**
 - Text in, audio out
 - SOTA: totally intelligible (if sometimes unnatural)



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: perhaps 80% accuracy for multi-sentence templates, 90%+ for single easy fields
- But remember: information is redundant!

New This Year!

The screenshot shows a search engine interface with a dark blue navigation bar at the top containing the following links: Home, Tips & Tricks, Features, Search Stories, Playground, Blog, and Help. The main content area is a 'Knowledge Graph' visualization on a black background with a starry pattern. It features a network of nodes and connecting lines. Two prominent nodes are circular portraits of Leonardo da Vinci, one in blue and one in white. Other nodes include a red maple leaf, a green globe, and various landscape and architectural images. A blue button with a white right-pointing arrow is positioned to the left of the text 'The Knowledge Graph'. Below this text is the subtext: 'Learn more about one of the key breakthroughs behind the future of search.' To the right, a search results card for 'Leonardo da Vinci' is displayed. The card includes a portrait of Leonardo da Vinci, a brief biographical description, and several key facts: 'Born: April 15, 1452, Anchiano', 'Died: May 2, 1519, Clos Lucé', 'Buried: Château d'Amboise', and 'Parents: Caterina da Vinci, Piero da Vinci'. Below these facts are links for 'Structures: Vejlem Sand Da Vinci Project' and 'leonardo.net'. A grey button with a white right-pointing arrow is positioned to the left of the text 'See it in action'. Below this text is the subtext: 'Discover answers to questions you never thought to ask, and explore collections and lists.'

Home Tips & Tricks **Features** Search Stories Playground Blog Help

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

See it in action
Discover answers to questions you never thought to ask, and explore collections and lists.

Leonardo da Vinci
Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. Wikipedia
Born: April 15, 1452, [Anchiano](#)
Died: May 2, 1519, [Clos Lucé](#)
Buried: [Château d'Amboise](#)
Parents: [Caterina da Vinci](#), [Piero da Vinci](#)
Structures: [Vejlem Sand Da Vinci Project](#)
[leonardo.net](#)



QA / NL Interaction

■ 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?”

■ Natural Language Interaction:

- Understand requests and act on them
- “Make me a reservation for two at Quinn’s tonight”

The screenshot shows a Google search interface. At the top, there are navigation links for Web, Images, Groups, News, Froogle, Local, and more. The search bar contains the text "any US states' capitals are also their largest cities?" and a "Search" button. Below the search bar, the word "Web" is displayed in a blue box. The main content area shows the search results: "Your search - **How many US states' capitals are also their largest cities?** - did not match any documents." Below this, there is a "Suggestions:" section with four bullet points: "- Make sure all words are spelled correctly.", "- Try different keywords.", "- Try more general keywords.", and "- Try fewer keywords." At the bottom of the page, 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 ' , -ĕn ') 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](#)

Oscar for best actress 1958



Examples Random

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 *Painted Faces*
Deborah Kerr in *Heaven Knows, Mr. Allison*
the Wind

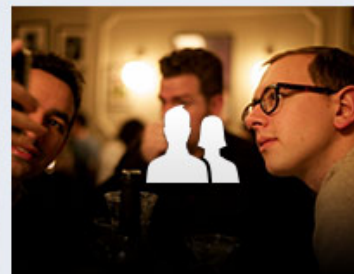
Information about Joanne Woodward:

full name	Joanne Gignilliat Trimmier
date of birth	Thursday February 27, 1930
place of birth	Thomasville, Georgia, United States

Academy Awards and nominations:

year	category
1991 (age: 61 years)	actress in a leading role
1974 (age: 44 years)	actress
1969 (age: 39 years)	actress

Photos of my friends |

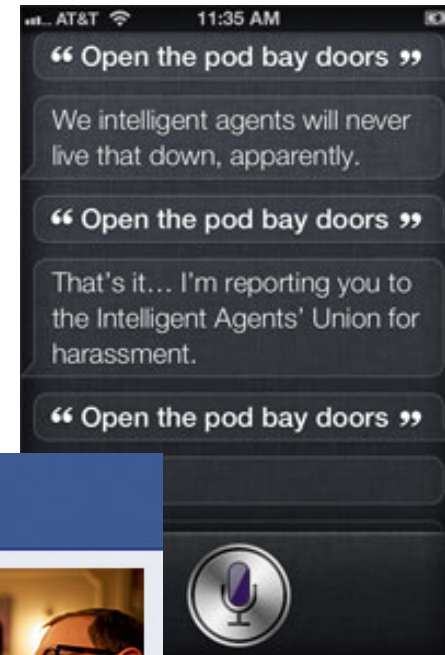


Winter Dreams



Rachel, Rachel

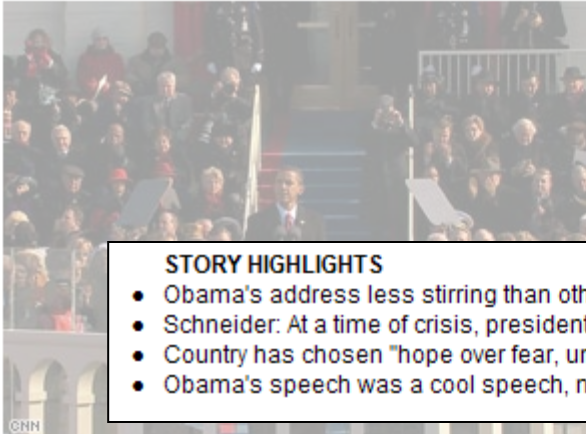
Hot Area!



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.

This year: 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

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 Google Translate: French

EN CE MOMENT

Impôts Kenya Syrie Pakistan Emploi Scandale Prism

Impôt sur le revenu : vous en 2014 ?



Sélectionnez votre revenu et votre situation et vous bénéficiez 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](#)
- A 2014 budget submitted to criticism
- Budget: these expenses no government can reduce
- 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

2013 Google Translate: Russian



pravda.ru
ENG RUS PT ITA

Поиск

Например: [Большой Кавказ](#)

Мир | Наука | Общество | Здоровье | Красота

■ Новости

20:09

[В Шри-Ланке хотели перевезти золото в желудках](#)

20:00

[Выходец из России может получить "Нобеля" по химии](#)

19:46

[В США установили стандарты торговли оружием](#)

19:35

[Директор Эрмитажа: Обыски нанесли ущерб музею](#)

19:25

[Мозгу ребенка полезен послеобеденный сон](#)

19:24

[Роликом с водителями-детьми заинтересовалась петербургская полиция](#)

19:15

[К Марсу приближается "комета века"](#)

18:55

[Выявлено более 160 нарушений на судостроительных предприятиях](#)

18:44

[Астахов назначен на новый срок в Европейской сети детских омбудсменов](#)

■ Главное

["Обиженные люди работают, а иностранцы к нам не поедут"](#)

25.09.2013 19:48



Ректор "Бауманки" Анатолий "Правде.Ру", какие шаги над чиновникам и ученым в связи реформе РАН.

■ Фотосессия

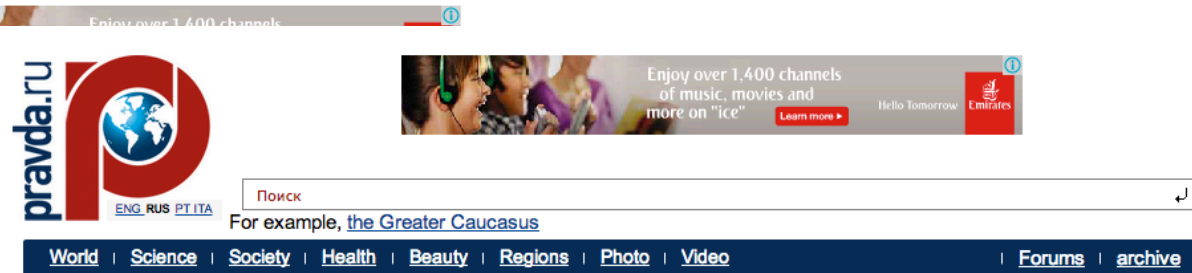


[Наводнение в Индии: 40 жителей эвакуированы](#)

Найроби. Газета The Independent "Уэстгейт" во время захвата.

■ Мир

Израиль на востоке



pravda.ru
ENG RUS PT ITA

Поиск

For example, [the Greater Caucasus](#)

World | Science | Society | Health | Beauty | Regions | Photo | Video

Forums | archive

■ News

20:09

[In Sri Lanka, wanted to carry the gold in the stomachs](#)

20:00

[A native of Russia can get the "Nobel" in Chemistry](#)

19:46

[In the United States set the standard arms trade](#)

19:35

[Director of the Hermitage: The searches have damaged the museum](#)

19:25

[The child's brain is useful afternoon nap](#)

19:24

[The roller with the drivers, children become interested in the St. Petersburg Police](#)

19:15

[To Mars is approaching "comet of the century"](#)

18:55

[There are over 160 violations at shipyards](#)

18:44

[Astakhov appointed for a new term in the European Network of Ombudsmen for children](#)

18:24

■ Point

["Mentally ill people are working, and foreign scholars to us will not go"](#)

25/09/2013 19:48



The Rector, "Bauman" Anatoly Alexandrov told with "Pravda.Ru" what steps need to be taken to officials and scientists in connection with the adoption of the law on the reform of the RAS.

■ Photoshoot



[World through the lens: September 25.](#)

In Kenya - mourning for the victims of the terrorist attack in Nairobi. The newspaper The Independent said about the people who were at the mall, "Westgate" during capture.



[Expert: The poorer the society is, the more scandals due to copyright](#)

09/25/2013 20:04

Why Russians are greedy for free, and do not like to pay for downloading movies and music, with "Pravda.Ru" said the head of Liveinternet German Klimenko.



[Putin met environmentalists "Greenpeace" trying to grab the platform](#)

25/09/2013 14:39

President of Russia, speaking at the International Arctic Forum in Salekhard, spoke about the ecology of Greenpeace, staged on a platform of "Prirazlomnaja."



[Expert: It is necessary to encourage participation in the election, rather than returning the column](#)

["against all"](#)

09/25/2013 13:27

Political scientist and philosopher, Professor Oleg Matveychev HSE commented with "Pravda.Ru" Valentina Matviyenko offer to return to the ballot line "against all."



[The British newspaper described the heroes and victims in Nairobi](#)

25/09/2013 10:27

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

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.



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!
- 2010s: you decide!

What is Nearby NLP?

■ Computational Linguistics

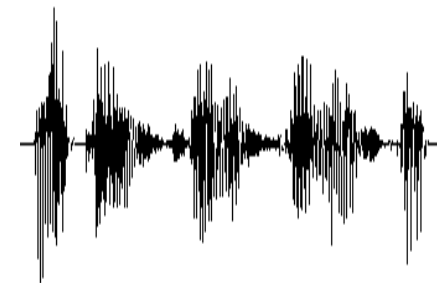
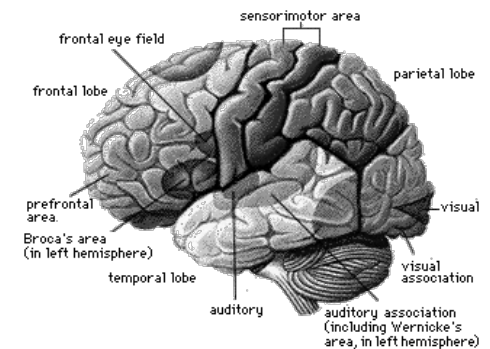
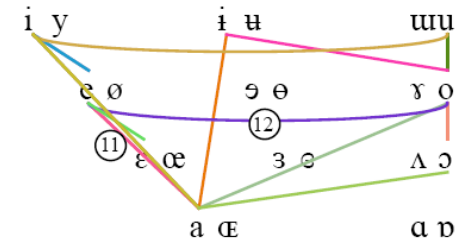
- Using computational methods to learn more about how language works
- We end up doing this and using it

■ 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

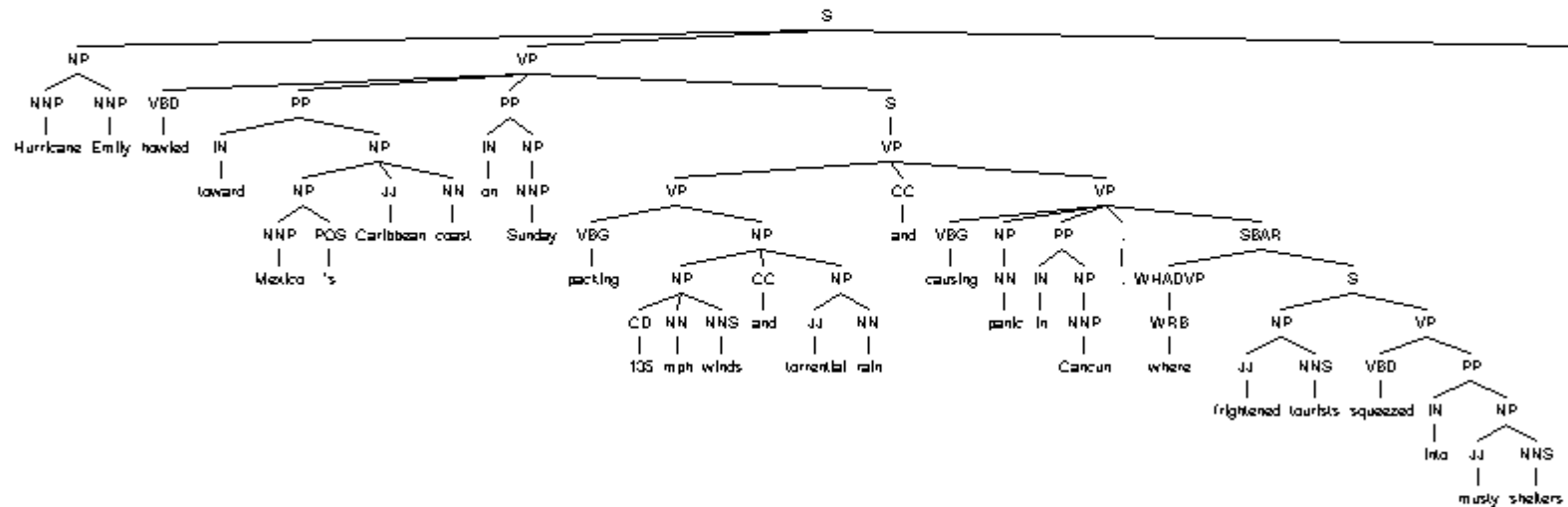


Problem: Ambiguities

- 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:** ~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.

- **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. 😊

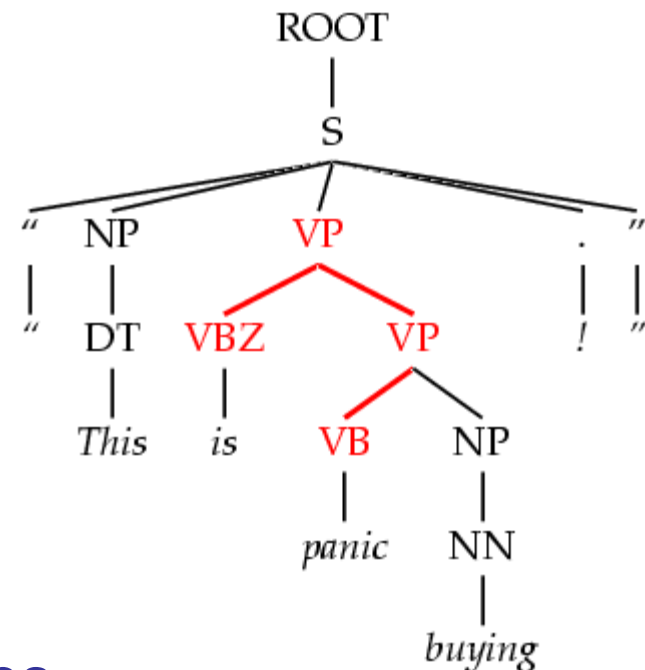
[Example from L. Lee]

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

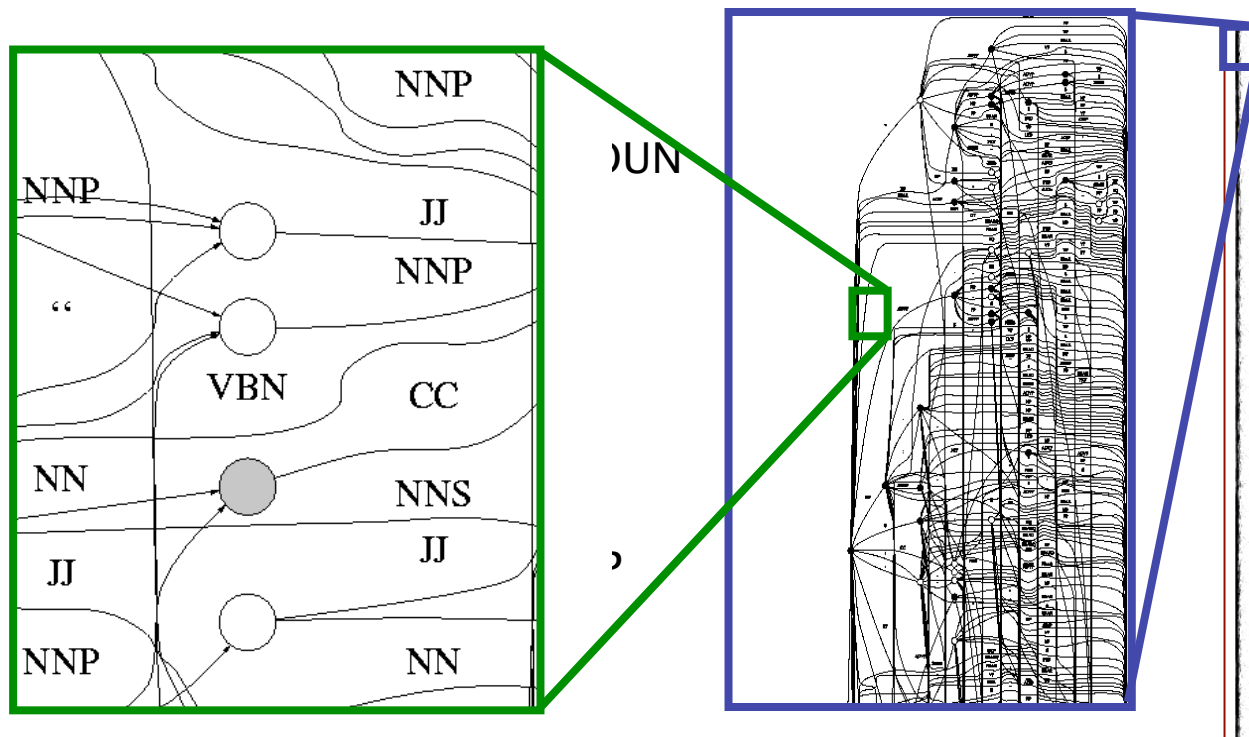
“This will panic buyers ! ”



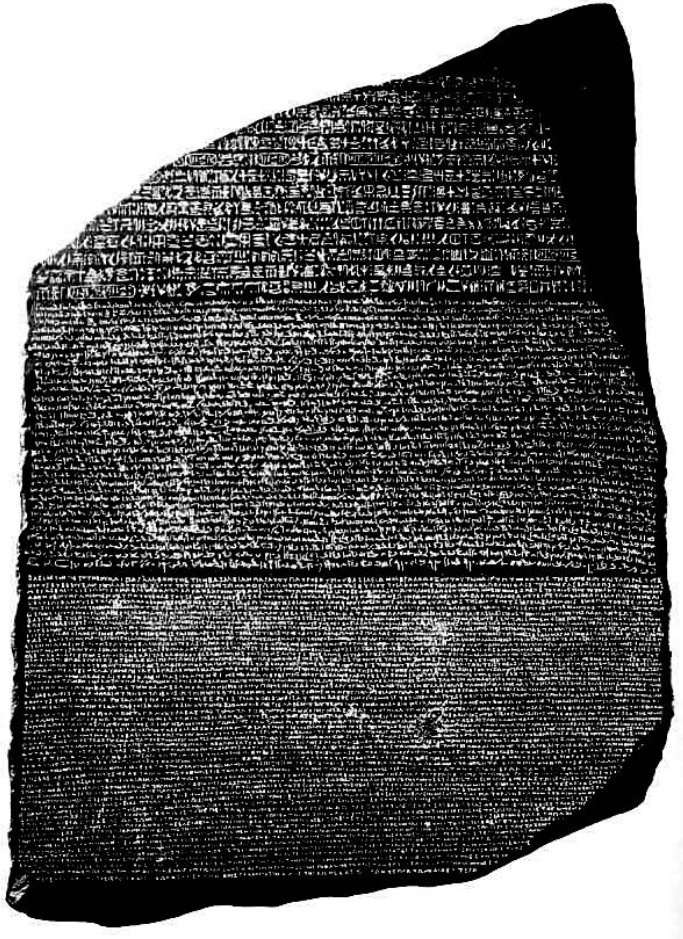
- Unknown words and new usages
- **Solution**: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

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



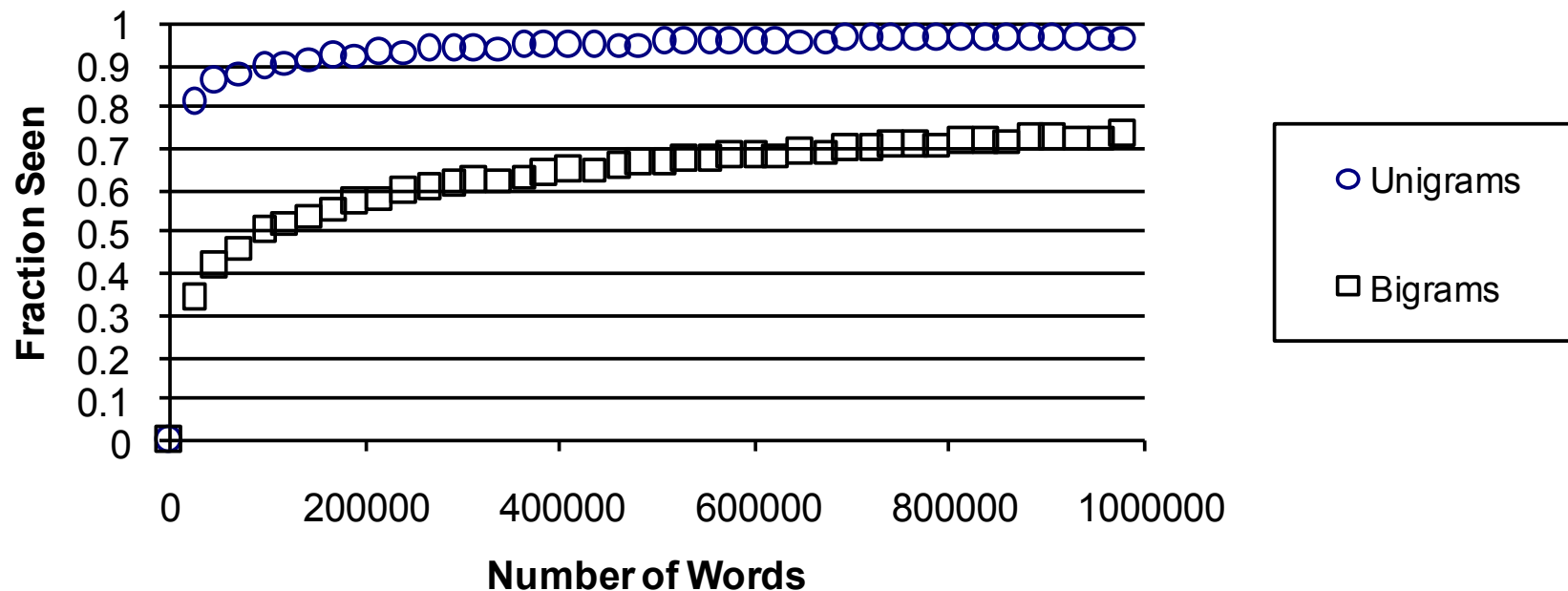
Corpora



- 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)



Outline of Topics

- Will be continually updated on website

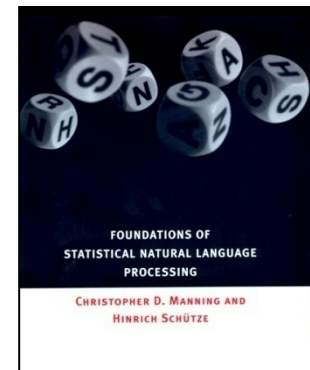
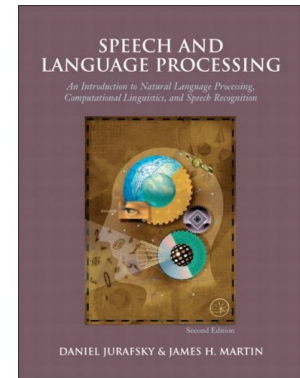
Schedule [*WARNING: topics subject to change!*]

Week	Dates	Topics & Lecture Slides	Notes (required reading)
1	Sep 25	Introduction; Language Modeling (LM)	LM Notes
2	Oct 2	Text Classification: Naive Bayes, Linear Models, EM	Naive Bayes (Sec. 1-4) , Log-linear models
3	Oct 9	Hidden Markov Models (HMMs) and Tagging	HMM Notes , CRF Notes
4	Oct 16	POS; PCFGs and Parsing	PCFG Notes
5	Oct 23	Parsing (cont.)	Lexicalized PCFGs
6	Oct 30	Intro to MT & Word Alignment	IBM Models 1 and 2
7	Nov 6	Phrase-based MT	Phrase-based Notes
8	Nov 13	Syntax-based MT; Information Extraction (IE)	
9	Nov 20	IE cont.; Discourse and Co-reference	
10	Nov 27	No class. Happy Thanksgiving!	
11	Dec 4	Compositional Semantics	

Textbooks

Course Details

- **Books (recommended but required):**
 - Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
 - Manning and Schuetze, Foundations of Statistical NLP
- **Prerequisites:**
 - CSE 421 (Algorithms) or equivalent
 - Some exposure to dynamic programming and probability helpful
 - Strong programming
 - **There will be a lot of math and programming**
- **Work and Grading:**
 - 100% - Four assignments (individual, submit code + write-ups)
- **Contact: see website for details**
 - Class participation is expected and appreciated!!!
 - Email is great, but please use the message board when possible (we monitor it closely)



Possible Assignments

- **Build a language model**
 - Sentence → Probability
- **Build a POS Tagger**
 - Sentence → Part of Speech (POS) for each word
- **Build a parser**
 - Sentence → Tree (encoding grammatical structure)
- **Build a word aligner**
 - Parallel sentences → Word/Phrase Translation Tables
- **Build a machine translation decoder**
 - Sentence in one language → sentence in another language

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...

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!

Questions for Class

- Office Hours
 - When? Daytime, evening, weekend?
 - Where? Should be try online chat rooms?
- Why NLP?
 - What are you iterests?
 - Any topics you want to see that aren't on the list?