CSE 517: Natural Language Processing

New Quals Course!

Instructor: Luke Zettlemoyer

Winter 2013

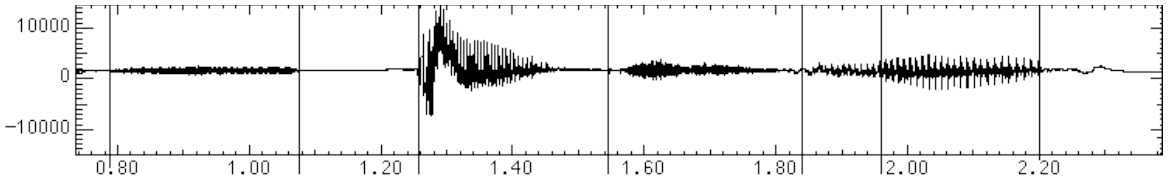
Slides adapted from Dan Klein

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.

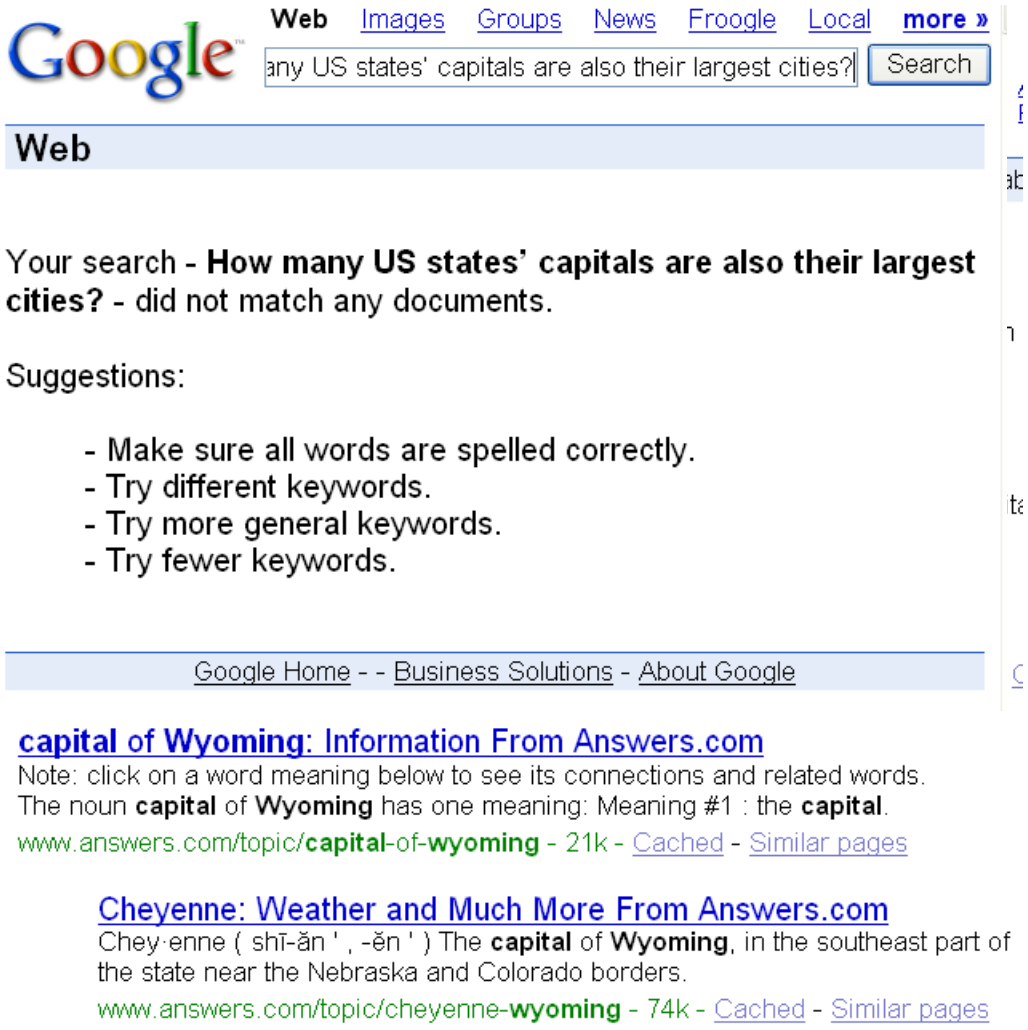
Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

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

New This Year!

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''



Hot Area!

WolframAlpha computational knowledge engine

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 Raintree County | Deborah Kerr in Heaven Knows, Mr. Allison | Anna Magnani in Wild Is 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

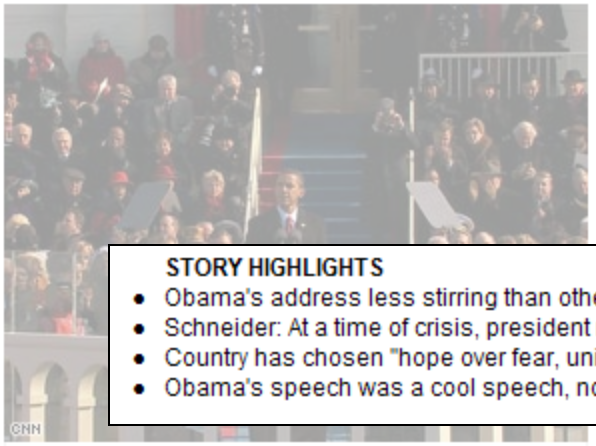
Academy Awards and nominations:

year	category	film

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.



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

more photos »

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

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

aid in 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.

Machine Translation

- 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 (next class) vs fidelity (later)

Machine Translation (French)

International – Le Monde.fr

http://www.lemonde.fr, RSS

Le Monde.fr

Mise à jour à 05h17 - Paris

"Il est impossible aux journal… rentrer dans les régions tibét…

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
Portfolio | **Reportage** | **Vidéo**

Accord sur la TVA : "Sar… de cause au pire momen…
Les ministres des finances europée… un compromis autorisant la réducti… certains secteurs, dont la restaurati…
Compte rendu Réactions mitigées a… une baisse de la TVA
Les faits Les taux réduits de TVA au…

Face aux déficits, la haus… paraît inéluctable
Le gouvernement exclut une augme… Philippe Séguin tire la sonnette d'al…
Infographie Finances publiques : les… gouvernementales
Les faits La crise avive le débat fisca…
Eclairage | **Compte rendu**

International – Le Monde.fr

Translated version of http://www.lemon…

http://translate.google.com/translate?prev=_t&hl=e

Google

Google™ This page was automatically translated from F…
View original web page or mouse over text to view orig…

"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
Portfolio | **Reportage** | **Video**

Agreement on the VAT: "Sarkozy wins t… case at the worst possible time"

The European finance ministers reached on Tuesday to a… compromise allowing the reduction of VAT rates in some… sectors, including catering.
Record Mixed reactions after **the** European agreement … reduction in VAT

Machine Translation (Japanese)

asahi.com（朝日新聞社）：ビジ…

http://www.asahi.com/business/

トップ ニュース スポーツ エンタメ ライフ

社会 ビジネス 政治 国際 文化 サイエンス

現在位置： asahi.com > ニュース > ビジネス

トップ ニュース 為替 株式 金利 トピックス

東洋経済ニュース ロイターニュース 宝くじ CSR

ビジネス

最新ニュース

- 東証は小幅安　金融株の下げ目立つ
  １２日の東京株式市場は、前日の大幅高の反動から売り注文が先行し、小幅に値を下げている。日経平均株価…… (11:13) [記事全文]
- 損保ジャパンと日本興亜が統合交渉　３大陣営に集約へ
  損害保険３位の損保ジャパンと５位の日本興亜損害保険が…を始めたことが１２日、分か…… (10:33) [記事全文]
- ＧＤＰ、１２．１%減に上方修正　１０-１２月期…
  内閣府が１２日発表した０８年１０～１２月期の国内総生…は、物価変動の影響を除いた…… (09:07) [記事全文]
- 金融サミット、気候変動も議論する可能性＝外交関…ター）
- 【株式・前引け】利益確定売りが先行、為替円高も…ＴＯＰＩＸとも小幅反落 (3/12)（東洋経済）
- 『今回の上昇は本物か』【森田レポート】 (3/11)（今…

新型プリ

asahi.com：朝日新聞社の速報ニュースサイト

Translated version of http://ww

http://translate.google.com/translate?prev=hp&hl=

Google

Google™ This page was automatically translated from Japanese.
View original web page or mouse over text to view original lan

Business

Latest News

- **The exchange of financial stocks fell slightly prominent lower**

12 stocks in Tokyo, ahead of sell orders from the backlash of higher yesterday, with slightly lower values. Nikkei ... ... (11:13) [Full article]

New Prius

- **Negotiation and integration of Japan Sompo Japan興亜to aggregate in three large camps**

Sompo Japan Insurance and it's five to start the negotiations for the merger o…
NIPPONKOA Insurance Co., Ltd. No. 12, 2007, minutes ... ... (10:33) [Full …

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

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

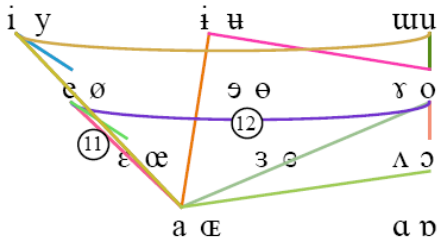
NLP: machine learning and empiricism

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

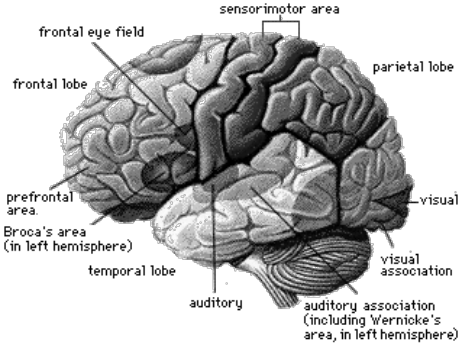
- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: you decide!

What is Nearby NLP?

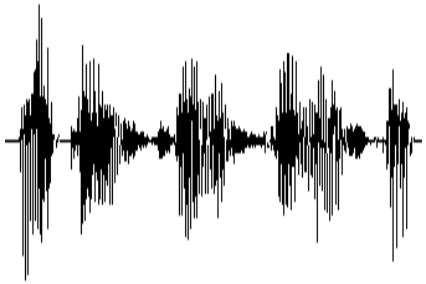
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- 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

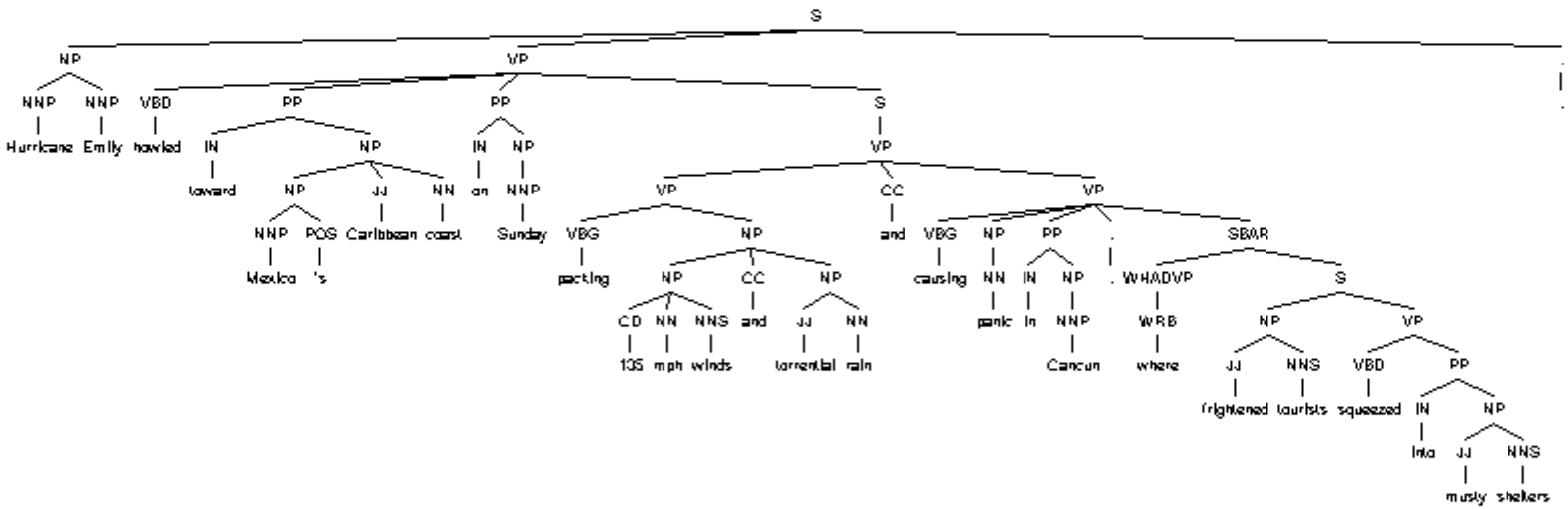


Problem: Ambiguities

- 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. ☺

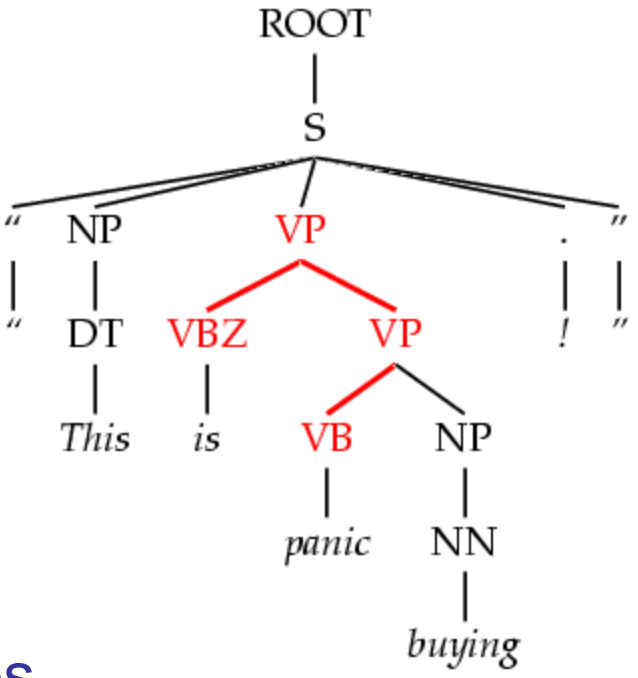
[Example from L. Lee]

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

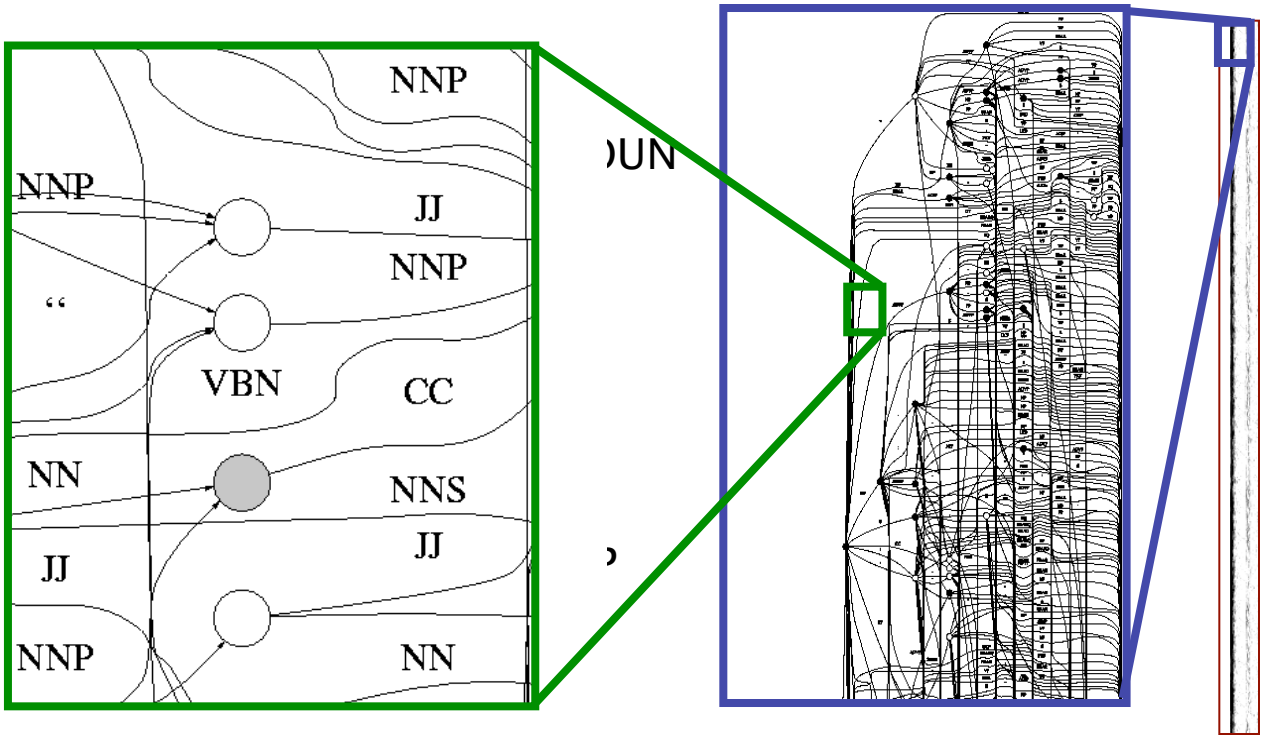
"This will panic buyers ! "



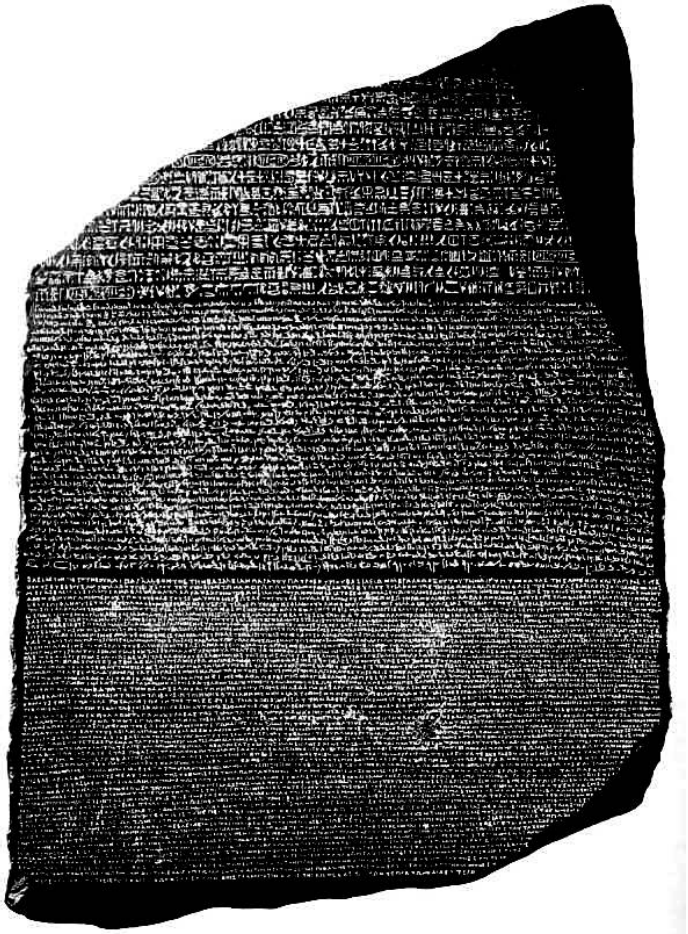
- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

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



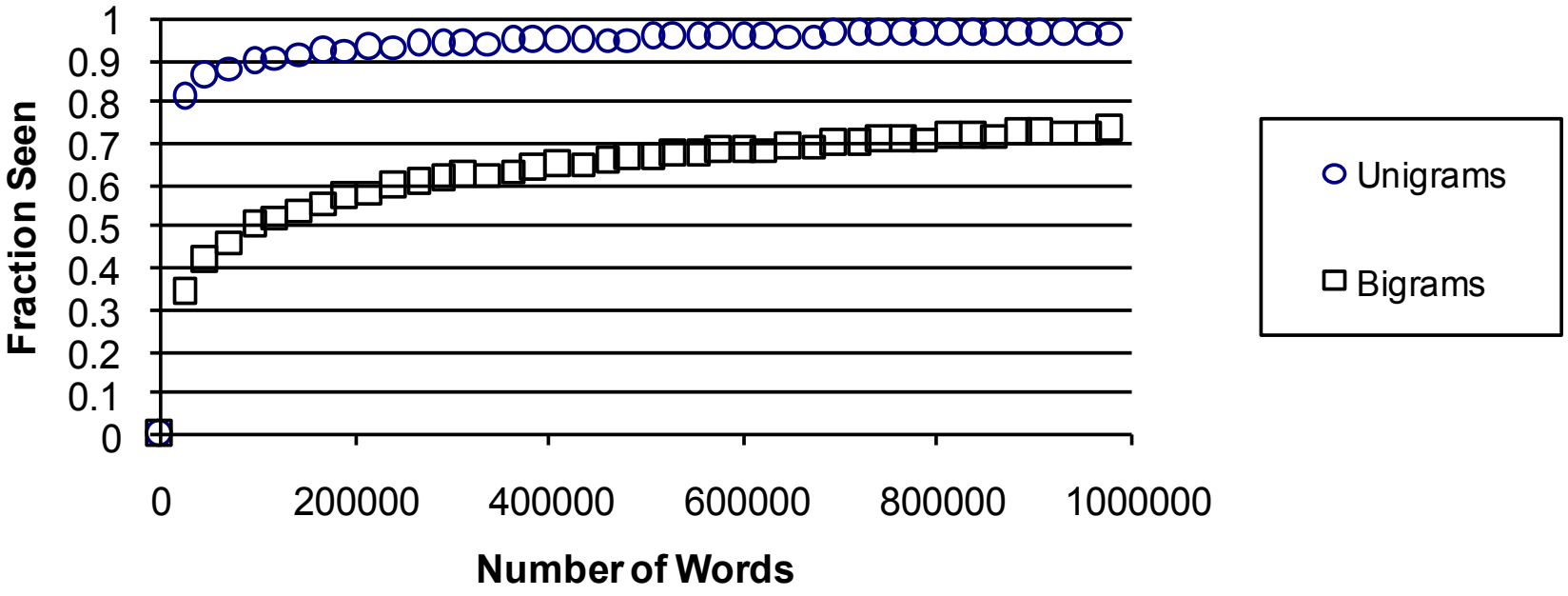
Corpora



- 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), and rule rates in newswire



Outline of Topics

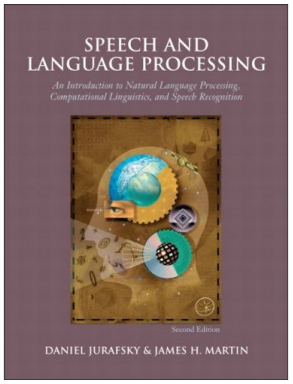
- Will be continually updated on website

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

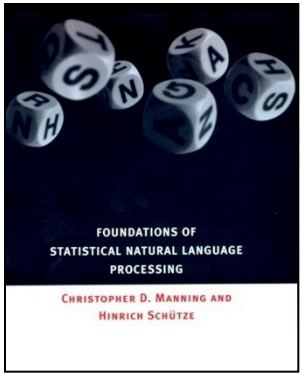
Dates	Topics & Lecture Slides	Notes	Textbook	Links
Jan 7, 9	Introduction; Language Modeling (LM)	LM Notes	J&M 4; M&S 6	[Smoothing] [Pitman-Yor]
Jan 14, 16	Hidden Markov Models (HMMs) and Tagging	HMM Notes	J&M 5.1-5.3, 6.1-6.4; M&S 9, 10.1-10.3	[TnT Tagger] [Stanford Ta
Jan 23	PCFGs and Parsing			
Jan 28, 30	PCFGs and Parsing (cont'd)			
Feb 4, 6	Machine Translation (MT) Intro.; Word Alignment			
Feb 11, 13	Phrase-based MT; Syntax-based MT			
Feb 20	Log-linear Models; Perceptron			
Feb 25, 27	Conditional Random Fields; Discriminative Parsing			
Mar 4, 6	Unsupervised Learning and EM			
Mar 11, 13	Compositional Semantics			

Course Details

- Books:
  - 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:
  - 60% - Four assignments (individual, submit code + write-ups)
  - 40% - Final project (individual or small group)

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!