Statistical NLP

Spring 2008

University of California
C A L
N L P
Berkeley

Lecture 1: Introduction

Dan Klein – UC Berkeley

Course Info

- Meeting times
  - Lectures: Tu/Th 11am-12:30pm, 405 Soda
  - Section: F 11am-12pm, location TBA
  - Dan's office hours: Th 12:30-2:30pm, 775 Soda + TBA
  - Aria's office hours: F 3pm-4pm, 525 Soda
- Communication
  - Web page: www.cs.berkeley.edu/~klein/cs294-19
  - Prof: Dan Klein: klein@cs.berkeley.edu
  - TA: Aria Haghighi: aria42@cs.berkeley.edu
  - Course newsgroup: ucb.class.cs294-19 (link to webnews on the web page)
- Enrollment

Access / Computation

- Accounts
  - Data and code will be available via the web
  - The class login and password will be emailed to the address you listed with the registrar
- Computing Resources
  - You will want more compute power than the instructional labs
  - Recommendation: start assignments early to find out whether what you have works

The Dream

- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge
- So:

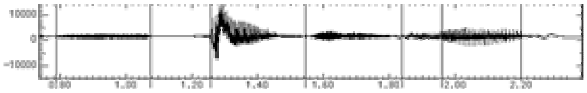
What is NLP?

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah blah blah...

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Modest: spelling correction, text categorization…

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

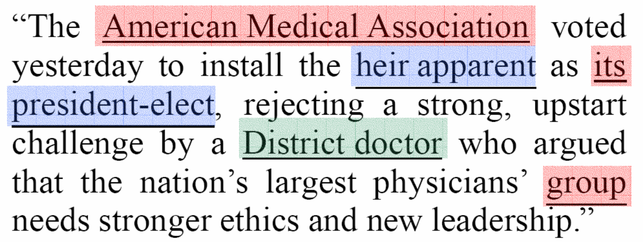
- Unstructured text to database entries

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 70% accuracy for multi-sentence temples, 90%+ for single easy fields

Coreference Resolution

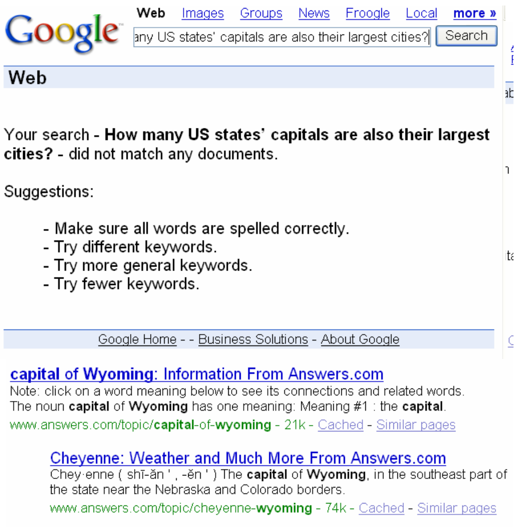


"The American Medical Association voted yesterday to install the heir apparent as its president-elect, rejecting a strong, upstart challenge by a District doctor who argued that the nation's largest physicians' group needs stronger ethics and new leadership."

- Can link repeated mentions of an entity both inside a document and across documents, model variation in spelling, transliteration, etc.
- SOTA: varies widely, 70-90%

Question Answering

- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - 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?”
- SOTA: Can do factoids, even when text isn’t a perfect match



Web

Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.

Suggestions:

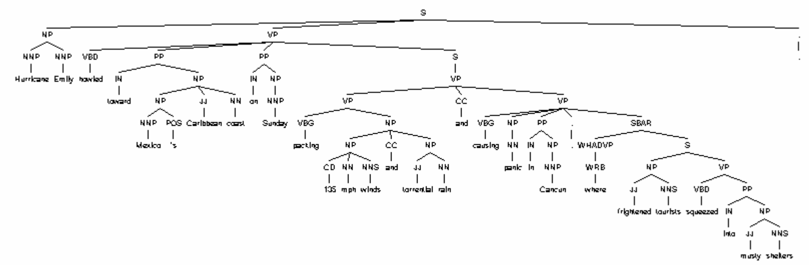
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Google Home - - Business Solutions - 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

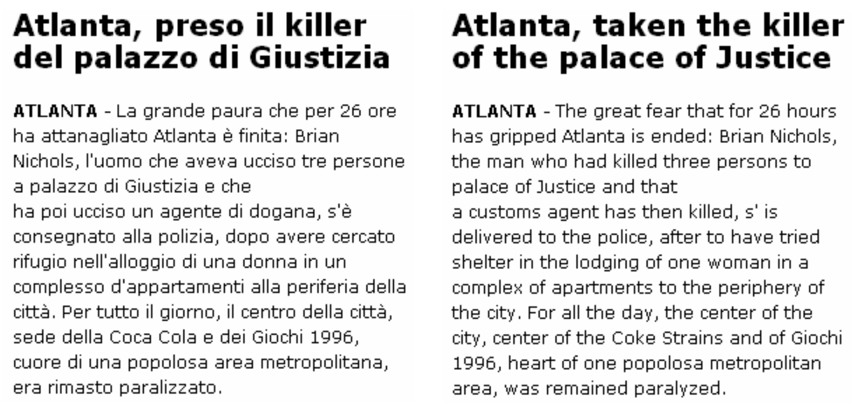
Syntactic Analysis



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: 80-90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

Machine Translation



Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che
ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

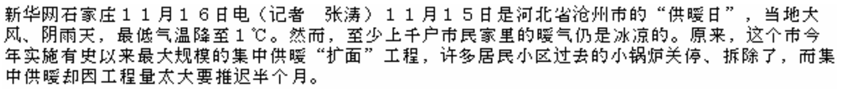
Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that
a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

- Translation systems encode:
  - Something about fluent language (next class)
  - Something about how two languages correspond (middle of term)
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators

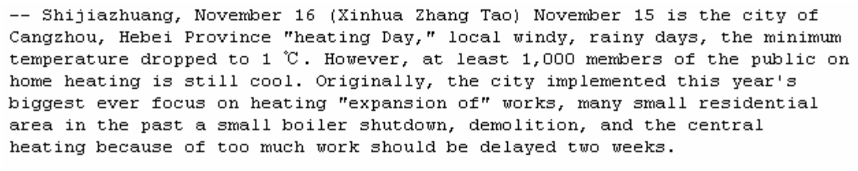
Machine Translation

Original Text



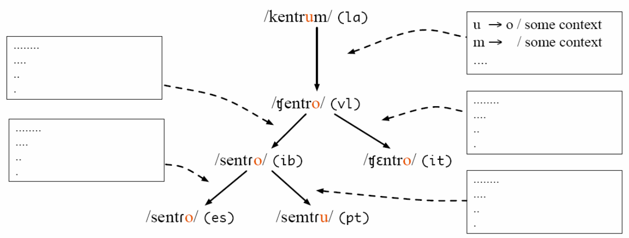
新华网石家庄１１月１６日电（记者　张涛）１１月１５日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至１℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

Translated Text



-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 ℃. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.

Etc: Historical Change

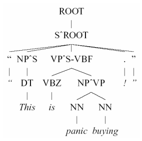


Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

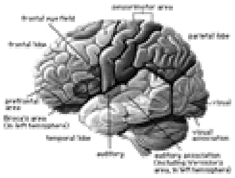
- Can model change in form over time, reconstruct ancient forms and phylogenies
- … just an example of the many other kinds of models we can build

What is nearby NLP?

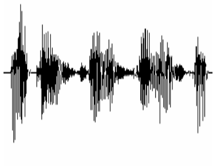
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- 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

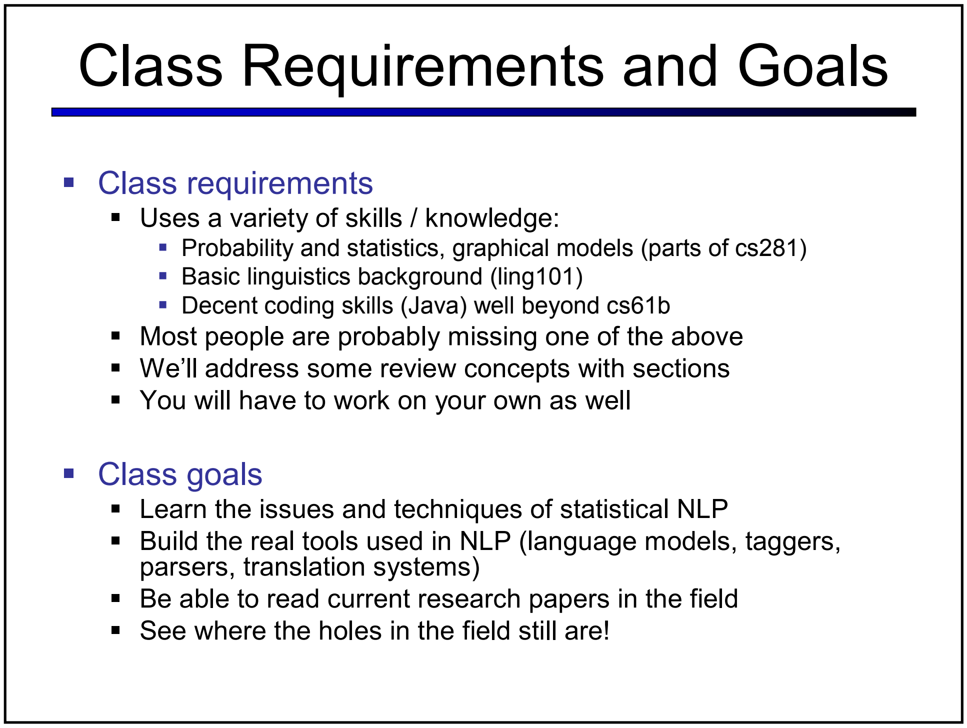


What is this Class?

- 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
  - Engineering Methods
    - Issues of scale
    - Sometimes, very ugly (but important) hacks
- We'll focus on what makes the problems hard, and what works in practice…

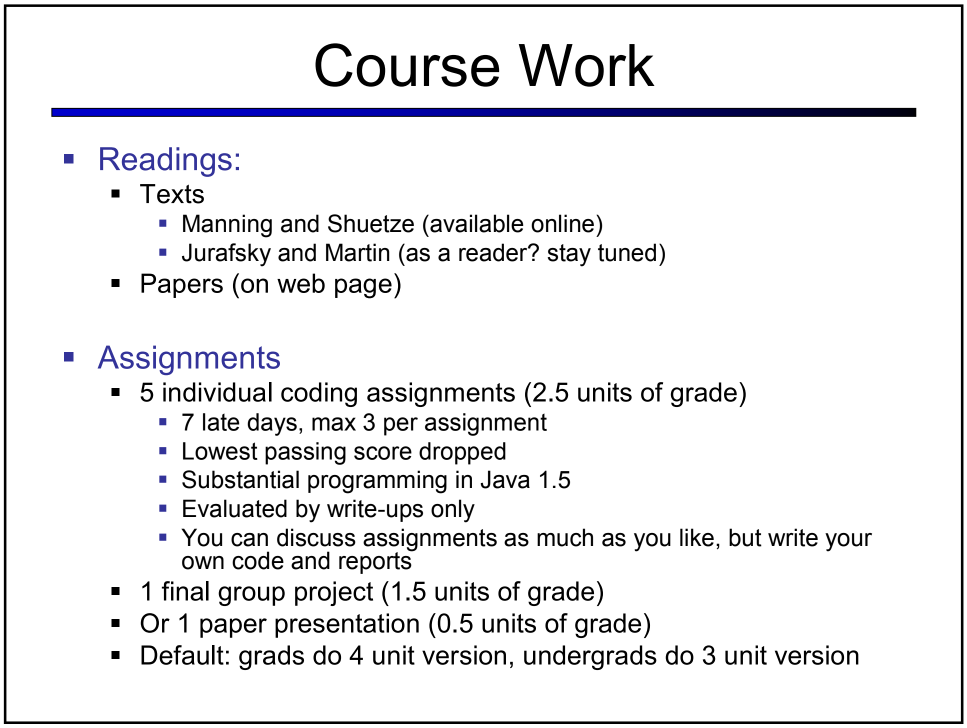
Outline of Topics

- Word level models
  - N-gram models and smoothing
  - Classification and clustering
- Sequences
  - Part-of-speech tagging
  - Information extraction
  - Speech recognition / synthesis
- Trees
  - Syntactic parsing
  - Grammar induction
  - Machine translation
  - Question answering



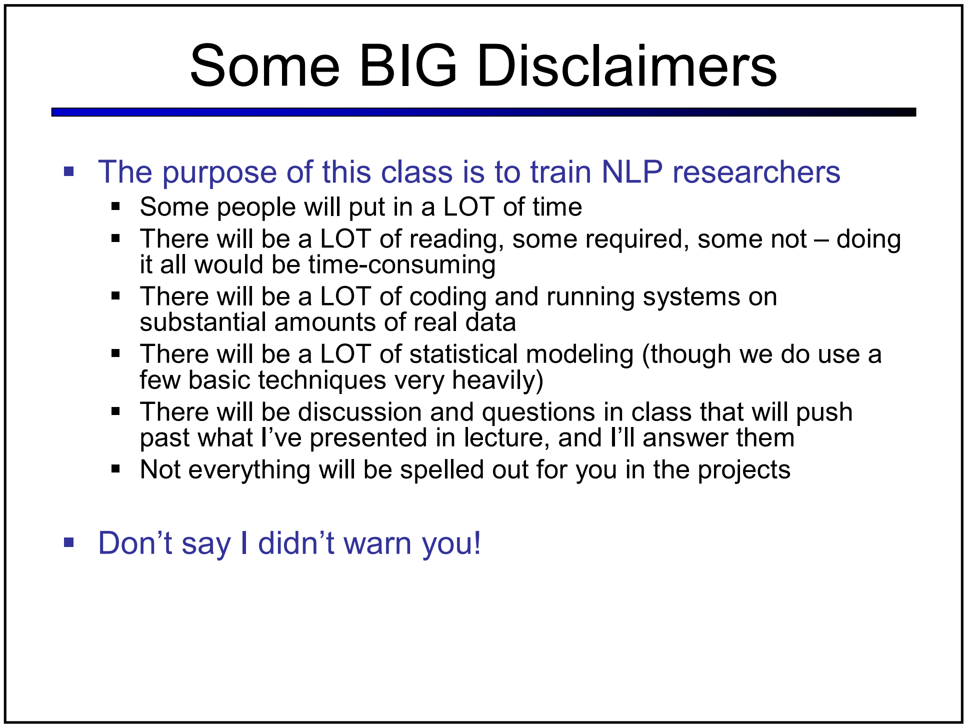
Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of cs281)
    - Basic linguistics background (ling101)
    - Decent coding skills (Java) well beyond cs61b
  - Most people are probably missing one of the above
  - We’ll address some review concepts with sections
  - You will have to work on your own as well
- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build the real tools used in NLP (language models, taggers, parsers, translation systems)
  - Be able to read current research papers in the field
  - See where the holes in the field still are!



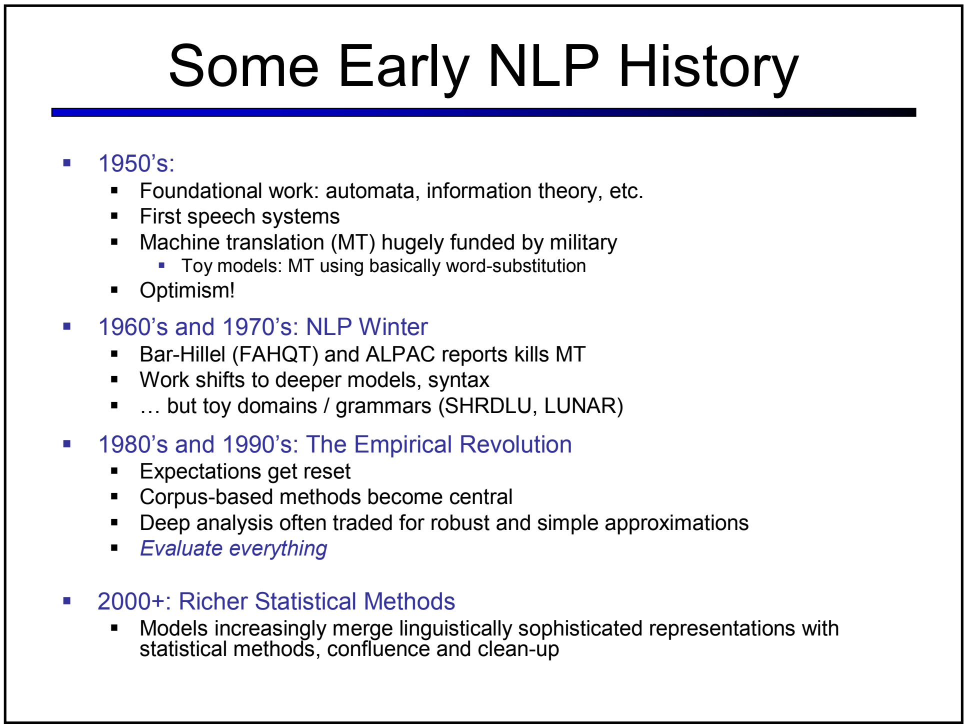
Course Work

- Readings:
  - Texts
    - Manning and Shuetze (available online)
    - Jurafsky and Martin (as a reader? stay tuned)
  - Papers (on web page)
- Assignments
  - 5 individual coding assignments (2.5 units of grade)
    - 7 late days, max 3 per assignment
    - Lowest passing score dropped
    - Substantial programming in Java 1.5
    - Evaluated by write-ups only
    - You can discuss assignments as much as you like, but write your own code and reports
  - 1 final group project (1.5 units of grade)
  - Or 1 paper presentation (0.5 units of grade)
  - Default: grads do 4 unit version, undergrads do 3 unit version



Some BIG Disclaimers

- The purpose of this class is to train NLP researchers
  - Some people will put in a LOT of time
  - There will be a LOT of reading, some required, some not – doing it all would be time-consuming
  - There will be a LOT of coding and running systems on substantial amounts of real data
  - There will be a LOT of statistical modeling (though we do use a few basic techniques very heavily)
  - There will be discussion and questions in class that will push past what I’ve presented in lecture, and I’ll answer them
  - Not everything will be spelled out for you in the projects
- Don’t say I didn’t warn you!



Some Early NLP History

- 1950’s:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960’s and 1970’s: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)
- 1980’s and 1990’s: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*
- 2000+: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up

NLP: Annotation

- Much of NLP is annotating text with structure which specifies how it's assembled.
  - Syntax: grammatical structure
  - Semantics: "meaning," either lexical or compositional

John bought a blue car

Why is NLP Hard?

- The core problems:
  - Ambiguity
  - Sparsity
  - Scale
  - Unmodeled Variables

Problem: Ambiguities

- Headlines:
  - Iraqi Head Seeks Arms
  - Ban on Nude Dancing on Governor’s Desk
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
  - Hospitals Are Sued by 7 Foot Doctors
- Why are these funny?

Syntactic Ambiguities

- Maybe we’re sunk on funny headlines, but normal, boring sentences are unambiguous?

Fed raises interest rates 0.5 % in a measure against inflation

Classical NLP: Parsing

- Write symbolic or logical rules:

Grammar (CFG)		Lexicon
ROOT → S	NP → NP PP	NN → interest
S → NP VP	VP → VBP NP	NNS → raises
NP → DT NN	VP → VBP NP PP	VBP → interest
NP → NN NNS	PP → IN NP	VBZ → raises
		…

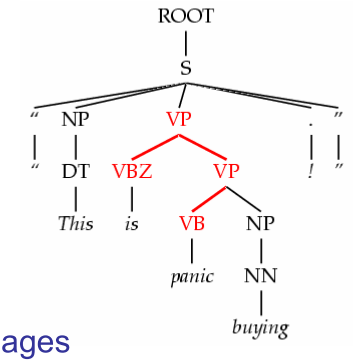
- Use deduction systems to prove parses from words
  - Minimal grammar on “Fed raises” sentence: 36 parses
  - Simple 10-rule grammar: 592 parses
  - Real-size grammar: many millions of parses
- This scaled very badly, didn’t yield broad coverage tools

Dark Ambiguities

- *Dark ambiguities*: most analyses are shockingly bad (meaning, they don’t have an interpretation you can get your mind around)

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

Semantic Ambiguities

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don’t always nail down the meaning

Every morning someone’s alarm clock wakes me up

John’s boss said he was doing better

- In general, every level of linguistic structure comes with its own ambiguities…

Other Levels of Language

- Tokenization/morphology:
  - What are the words, what is the sub-word structure?
  - Often simple rules work (period after “Mr.” isn’t sentence break)
  - Relatively easy in English, other languages are harder:
    - Segementation

哲学家维特根斯坦出生于维也纳

    - Morphology

sarà	*andata*
be+fut+3sg	go+ppt+fem

“she will have gone”

- Discourse: how do sentences relate to each other?
- Pragmatics: what intent is expressed by the literal meaning, how to react to an utterance?
- Phonetics: acoustics and physical production of sounds
- Phonology: how sounds pattern in a language

Disambiguation for Applications

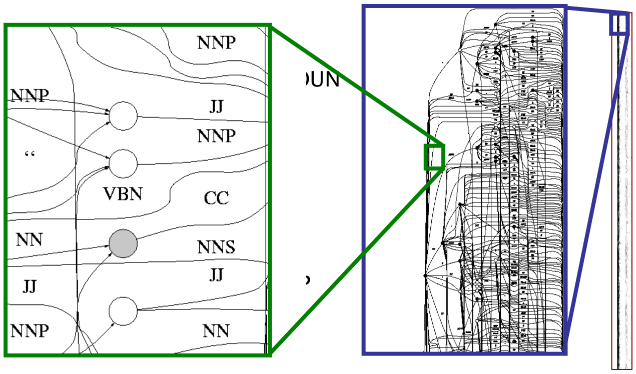
- Sometimes life is easy
  - Can do text classification pretty well just knowing the set of words used in the document, same for authorship attribution
  - Word-sense disambiguation not usually needed for web search because of majority effects or intersection effects (“jaguar habitat” isn’t the car)
- Sometimes only certain ambiguities are relevant

he hoped to record a world record

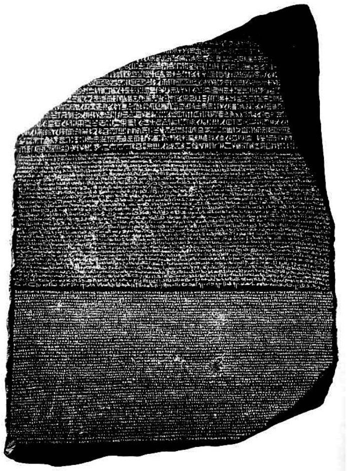
- Other times, all levels can be relevant (e.g., translation)

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



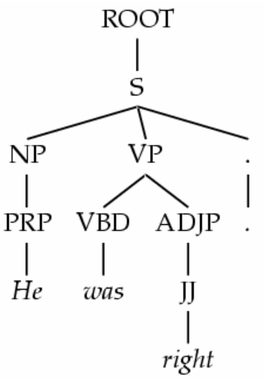
Corpora



- 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

Corpus-Based Methods

- A corpus like a treebank gives us three important tools:
  - It gives us broad coverage

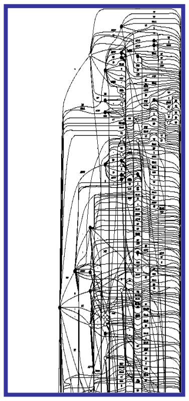


ROOT → S

S → NP VP .

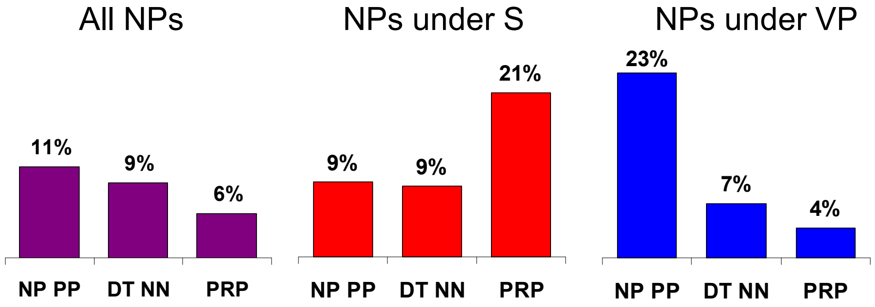
NP → PRP

VP → VBD ADJ



Corpus-Based Methods

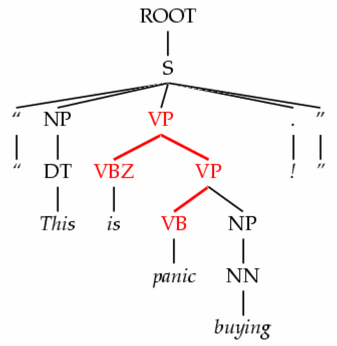
- It gives us statistical information

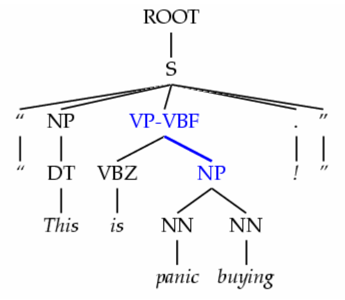


This is a very different kind of subject/object asymmetry than what many linguists are interested in.

Corpus-Based Methods

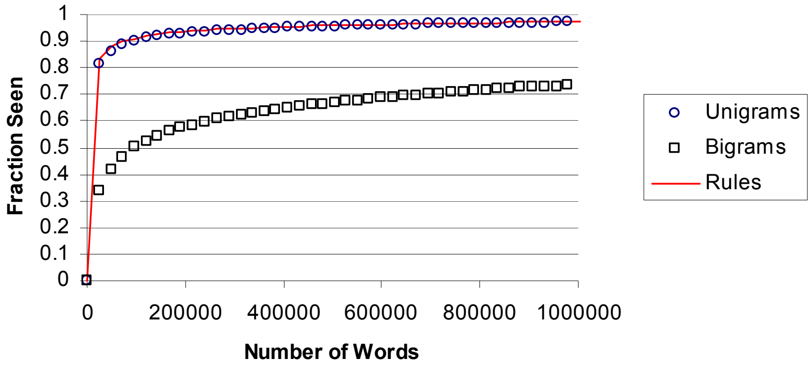
- It lets us check our answers





Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire



The (Effective) NLP Cycle

- Pick a problem (usually some disambiguation)
- Get a lot of data (usually a labeled corpus)
- Build the simplest thing that could possibly work
- Repeat:
  - Examine the most common errors are
  - Figure out what information a human might use to avoid them
  - Modify the system to exploit that information
    - Feature engineering
    - Representation redesign
    - Different machine learning methods
- We’re going to do this over and over again

Language isn't Adversarial

- One nice thing: we know NLP can be done!
- Language isn't adversarial:
  - It's produced with the intent of being understood
  - With some understanding of language, you can often tell what knowledge sources are relevant
- But most variables go unmodeled
  - Some knowledge sources aren't easily available (real-world knowledge, complex models of other people's plans)
  - Some kinds of features are beyond our technical ability to model (especially cross-sentence correlations)

What's Next?

- Next class: noisy-channel models and language modeling
  - Introduction to machine translation and speech recognition
  - Start with very simple models of language, work our way up
  - Some basic statistics concepts that will keep showing up
- If you don't know what conditional probabilities and maximum likelihood estimators are, read up!
- Reading on the web
- Assignment 1 will be out this week!