Introduction to Computational Linguistics

Spring 2016

Lecture 1: Introduction

Klinton Bicknell

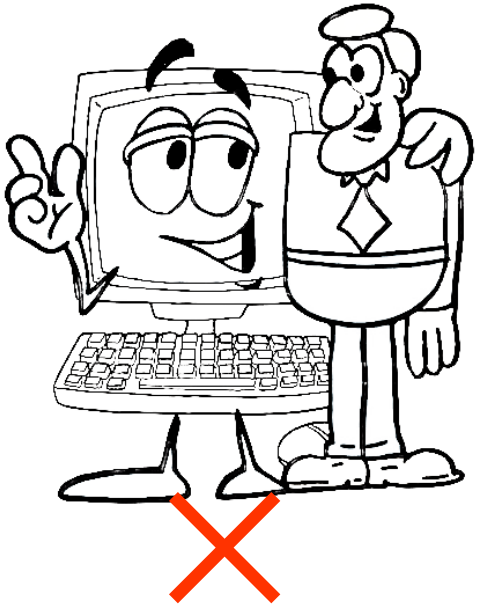
(borrowing from: Dan Klein, Roger Levy, Dan Jurafsky, and Jim Martin)

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

The mystery

- What's now impossible for computers (and any other species) to do is effortless for humans



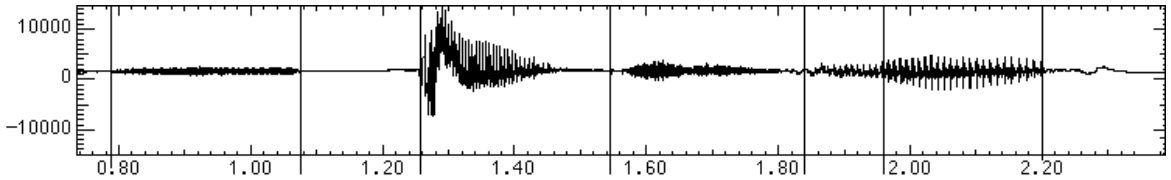
What is computational linguistics?

- 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…
- Theoretical goals: providing satisfactory accounts of human language acquisition and use

Speech Systems

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



“Speech Lab”

- Text to Speech (TTS)
  - Text in, audio out
  - State of the art: totally intelligible (if sometimes unnatural)

- Speech systems currently:
  - Model the speech signal
  - Model language
  - In practice, speech interfaces usually wired up to dialog systems

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
  - Something about how two languages correspond
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators

Information Extraction

- Information Extraction (IE)
  - Unstructured text to database entries

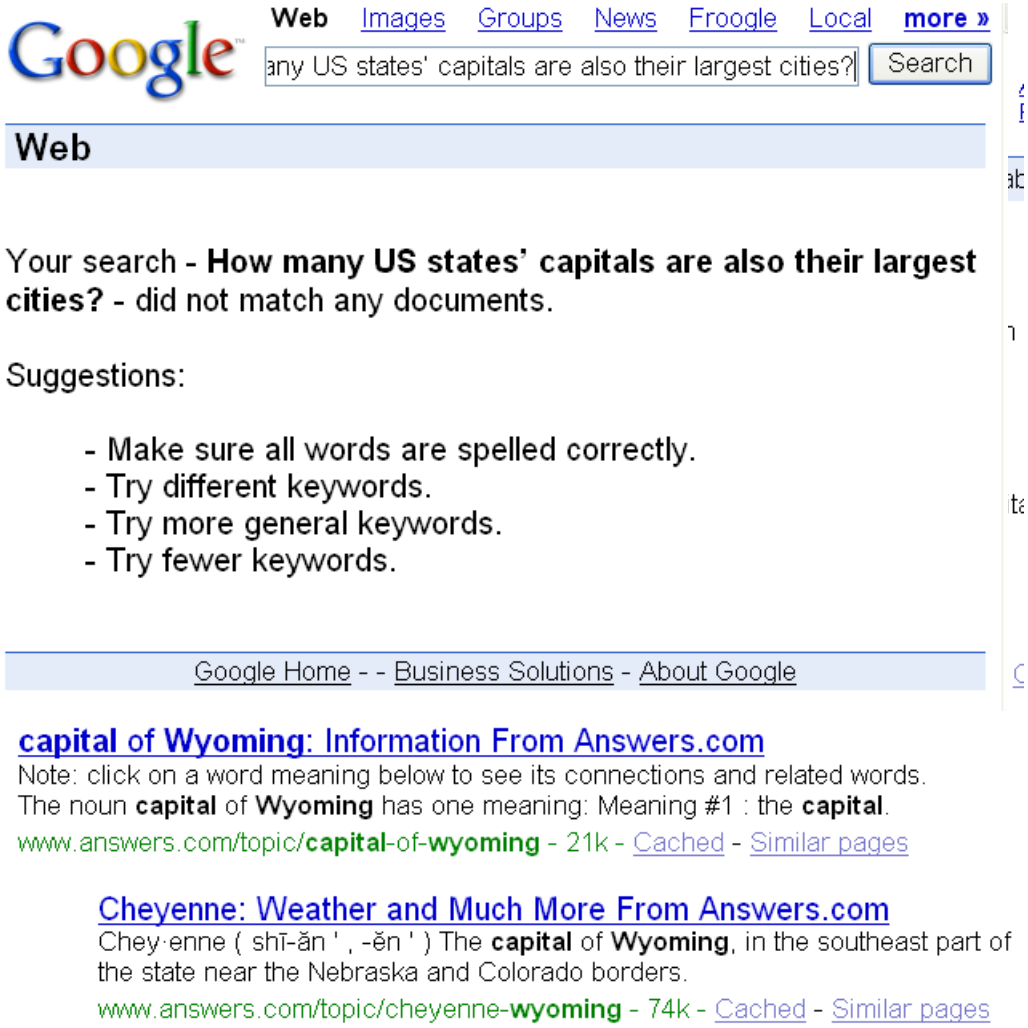
New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief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

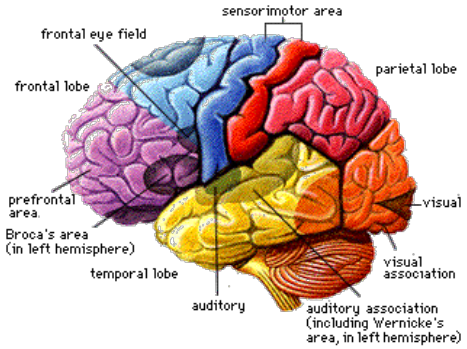
Question Answering

- 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



Closely related to CL

- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working CL prototype!
  - We'll cover a bit of this near the end of the course



Fast growth of CL

- an enormous amount of knowledge now available in machine readable form as natural language text
- conversational agents becoming an important for of human-computer interaction
- much of human-human communication is mediated by computers
- all this language data well used by ***statistical approaches***

Ambiguity

- computational linguists are obsessed with ambiguity
- ambiguity is a fundamental problem in CL
- resolving ambiguity is a crucial goal

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

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category:** "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** "make" can mean "create" or "cook"

Ambiguity is Pervasive

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

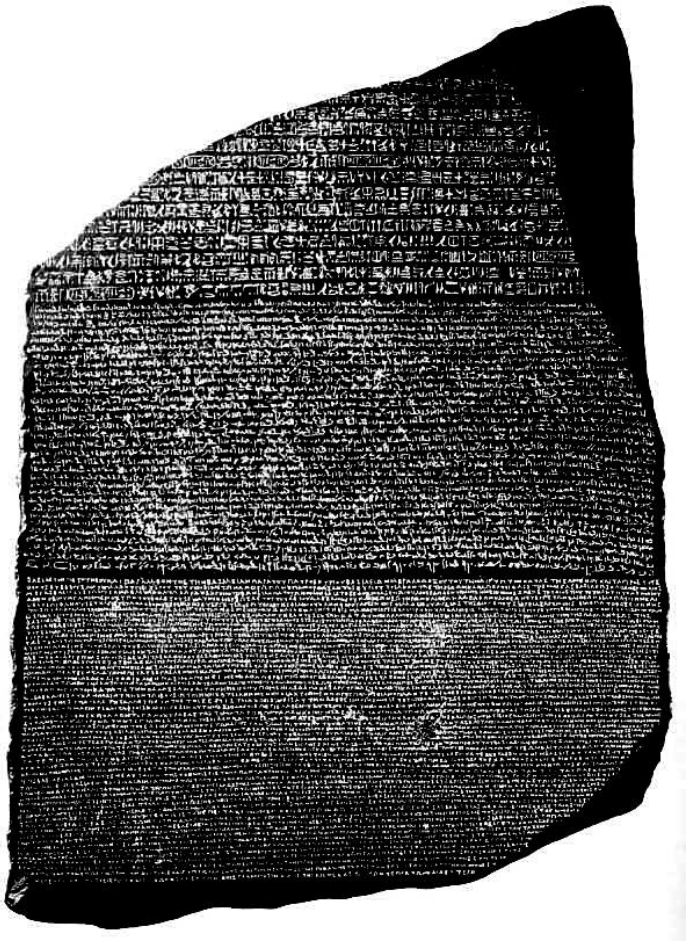
Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

Ambiguity: no single right answer

- many interpretations could be correct
- but most interpretations are very unlikely
- goal: we want to assign probabilities to interpretations
- solution: what linguistic analysis did similar inputs receive before?
    - use corpora to *train* models
    - how do we formalize *similar*

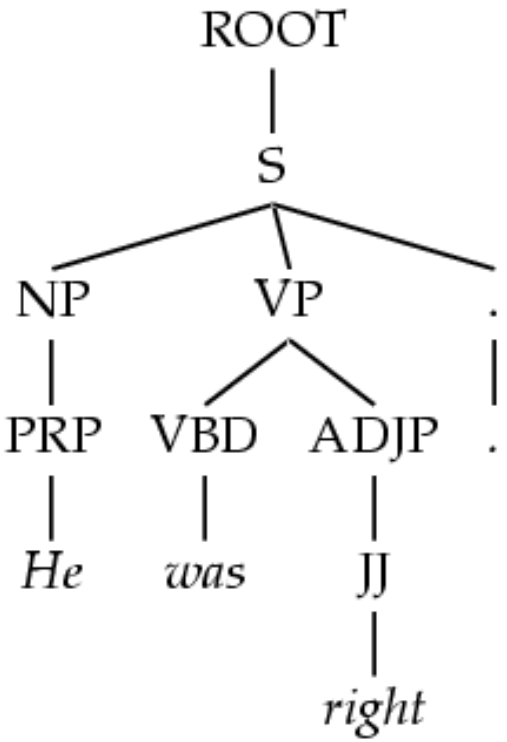
Corpora



- 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

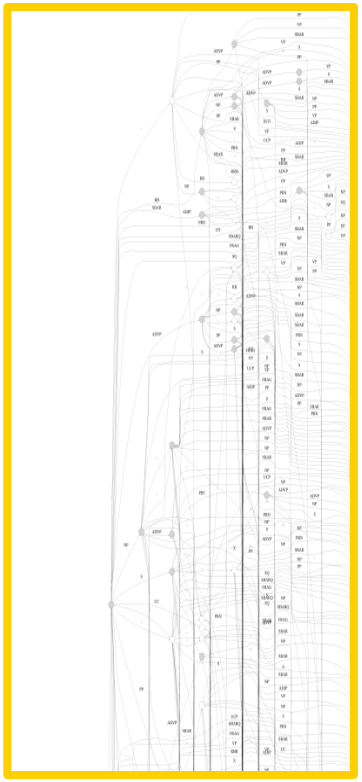


ROOT → S

S → NP VP .

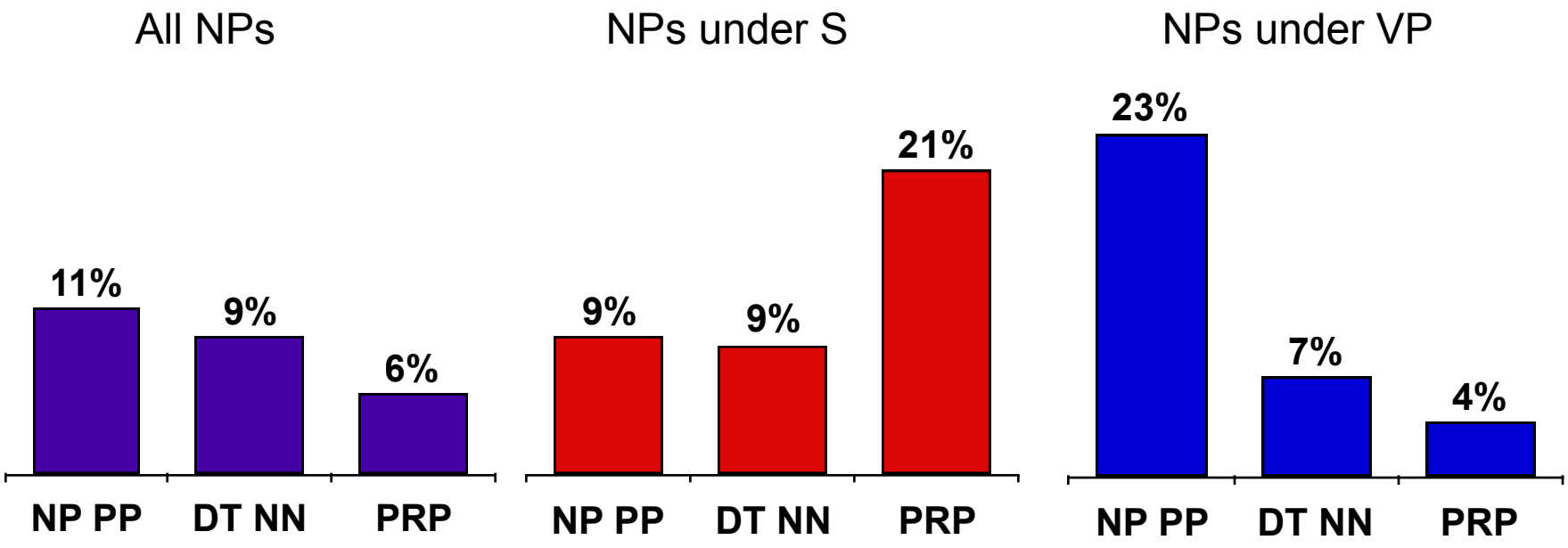
NP → PRP

VP → VBD ADJ



Corpus-Based Methods

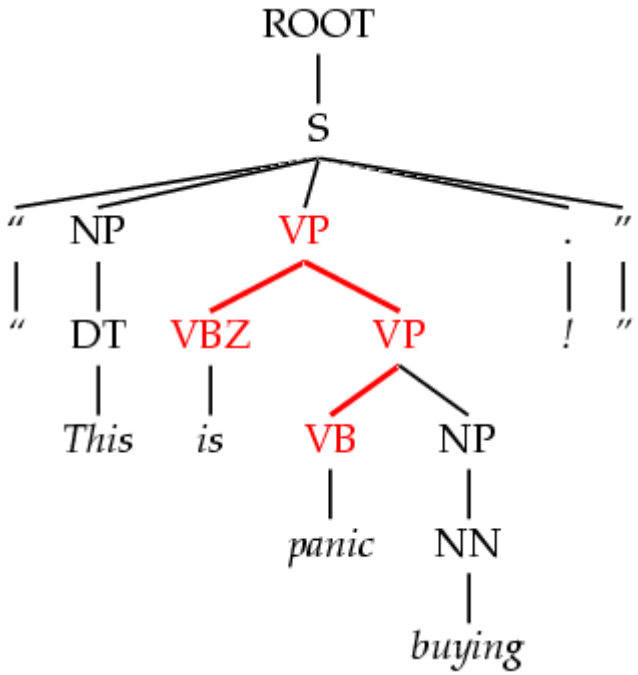
- It gives us statistical information
- "Subject-object asymmetry":

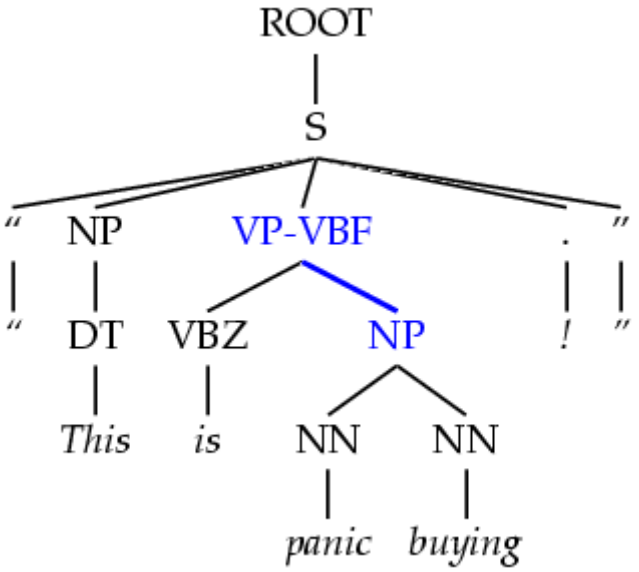


- *This is a very different kind of subject/object asymmetry than the traditional domain of interest for linguists*
- *However, there are connections to recent work with quantitative methods* (e.g., Bresnan, Dingare, Manning 2003)

Corpus-Based Methods

- It lets us check our answers!





The (Effective) CL 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:
  - See what the most common errors are
  - Figure out what information a human would use
  - Modify the system to exploit that information
    - Feature engineering
    - Representation design
    - Machine learning/statistics

Course goals

- Three aspects to the course:
  - How are computational linguistic tasks currently done?
    - automatic speech recognition
    - autocorrect
    - machine translation
    - syntactic parsing
    - identifying parts-of-speech
  - Probabilistic modeling skills
    - how to specify a (good) probabilistic model
    - how to make inferences given a model
    - how to train a probabilistic model
    - Efficient algorithms: dynamic programming, search
  - Programming and computing skills
    - how to program: using python
    - how to run programs on remote compute servers using linux
    - how to use existing off-the-shelf CL tools

Course outline

- **weeks 1–3: foundations**
  - unix/linux, programming, python (this is a lot to learn!)
  - regular expressions and finite-state automata
  - probability theory (can't be scared of math!)
- **week 4: n-gram models for word prediction**
  - inference in simple probabilistic models
- **weeks 5–6: hidden Markov models for part-of-speech tagging**
  - simplest case of inferring hidden linguistic structure
- **weeks 7–9: more advanced models**
  - noisy-channel models for spelling correction / autocorrect
  - context-free grammars for syntax, parsing
  - automatic speech recognition
  - machine translation
- **final topic**
  - computational psycholinguistics: how CL models are used to understand how people do language