CS 294-5: Statistical Natural Language Processing

Course Introduction
Lecture I: 8/29/05

Course Info

- Meeting times
  - Lectures: Monday / Wednesday 1-2:30pm
  - Office hours: Thursday 4-5pm, Friday 1-2pm
- Communication
  - Web page: www.cs.berkeley.edu/~klein/cs294-5
  - My email: klein@cs.berkeley.edu
  - Course newsgroup: ucb.class.cs294-5 (link to webnews on the web page)
- Questionnaires!

Accounts and Access

- Accounts
  - Data and code available on:
    - CS instructional accounts
    - EECS research accounts
    - Millennium accounts
  - Make sure you have one of them working
  - More details on resources in assignment 1 (next class), but sort out access ASAP
- Computing Resources
  - Lab resources may not be enough
  - Recommendation: start assignments early to find out
  - NLP cluster on Millennium network, signups later in the term

The Dream

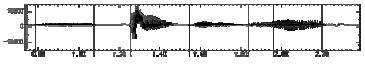
- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle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

What is NLP?

- 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% for digit strings, 5% dictation, 50%+ TV



"Speech Lab"

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)
- Speech systems currently:
  - Model the speech signal (later part of term)
  - Model language (next class)

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

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

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?
  - Technical Methods
    - Learning and parameter estimation
    - Increasingly complex model structures
    - Efficient algorithms: dynamic programming, search
  - Engineering Methods
    - Issues of scale
    - Sometimes, very ugly hacks
- 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:
    - Basic probability and statistics
    - Basic linguistics background
    - Decent coding skills (Java)
  - Most people are probably missing one of the above
  - We'll address some review concepts with sections, TBD
- 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 gaping holes in the field are!

Course Work

- Readings:
  - Texts
    - Manning and Shuetze (available online)
    - Jurafsky and Martin
    - Both on reserve in the Engineering library
  - Papers (on web page)
- Assignments
  - 5 individual coding assignments (60% of grade)
    - 7 late days, 3 per assignment
    - Lowest score dropped
    - Substantial programming in Java 1.5
    - Evaluated by write-ups
  - 1 group final project (40% of grade)

Some Early NLP History

- 1950's:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military (imagine that)
    - 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: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*

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

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*

What Made 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?

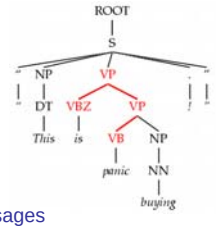
Fed raises interest rates 0.5 % in a measure against inflation

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

- 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

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

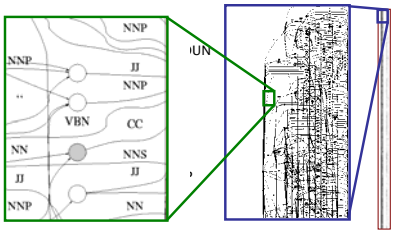
- Sometimes life is easy
  - Can do text classification pretty well just knowing the set of words used in the document, same for authorship attribution
  - Word-sense disambiguation not usually needed for web search because of majority effects or intersection effects ("jaguar habitat" isn't the car)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

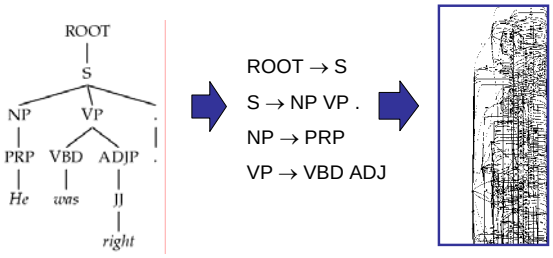


Corpora

- 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

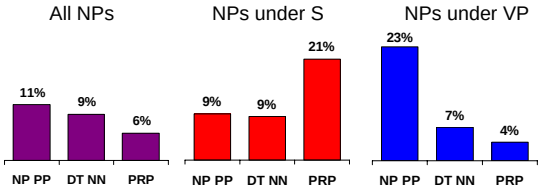
Corpus-Based Methods

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



Corpus-Based Methods

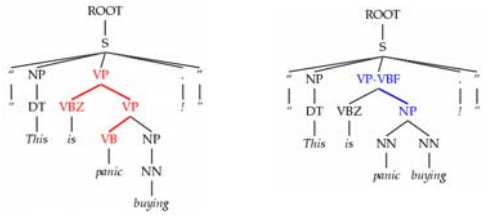
- It gives us statistical information



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

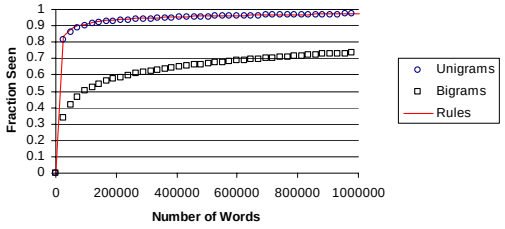
Corpus-Based Methods

- 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:
  - 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 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)