CS 294-5: Statistical Natural Language Processing

Dan Klein
MF 1-2:30pm
Soda Hall 310

Course Info

- Meeting times
  - Lectures: MF 1-2:30pm
  - 1:00-1:45, break, 1:50-2:30
  - Office hours: W 3-5pm (but negotiable)
- Communication
  - Web page: www.cs.berkeley.edu/~klein/cs294-5
  - My email: klein@cs.berkeley.edu
  - Course newsgroup: ucb.class.cs294-5
- Assignments
  - Readings (M+S, J+M, papers)
  - 5 Small Projects, 1 Final Project
- Questionnaires!

The Dream

- Language is our UI!
- It'd be great if machines could:
  - Process our email (usefully)
  - Translate for us
  - Write up our research
  - Talk to us / listen to us
- But they can't:
  - Language is ambiguous
  - Language is flexible
  - Language is complex
  - Language is subtle
- So we use their UIs.

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- Applications:
  - Ambitious: machine translation, information extraction, dialog systems, question answering…
  - Modest: spelling correction, text categorization

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 Recognition
  - 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
    - Memory limitations
    - Sometimes, very ugly hacks
- We'll visit a series of language problems

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!
- 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!

An Example

John bought a blue car

Language is Ambiguous

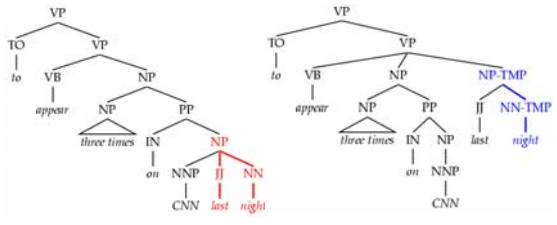
- Headlines:
  - Iraqi Head Seeks Arms
  - Ban on Nude Dancing on Governor's Desk
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
  - British Left Waffles on Falkland Islands
  - Clinton Wins on Budget, but More Lies Ahead
  - Hospitals Are Sued by 7 Foot Doctors
- Why are these funny?

Ambiguities Everywhere

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

Fed raises interest rates 0.5 % in a measure against inflation

More Attachment Ambiguities



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?
- Pragmatics: what intent is expressed by the literal meaning, how to react?

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)

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
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*

Classical NLP: Parsing

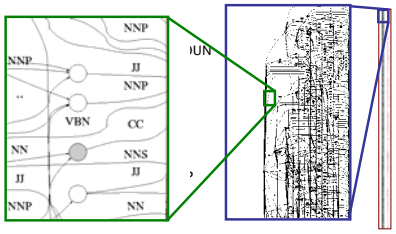
- Write symbolic or logical rules:

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

What Were They Thinking?

- 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

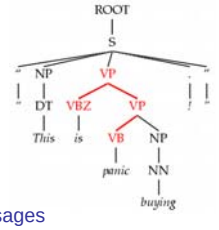


Problems (and Solutions?)

- *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 ! "



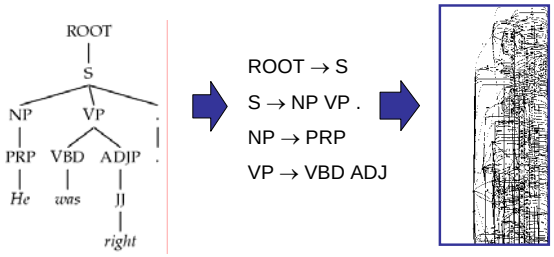
- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

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

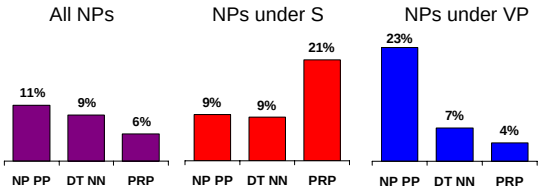
Corpus-Based Methods

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



Corpus-Based Methods

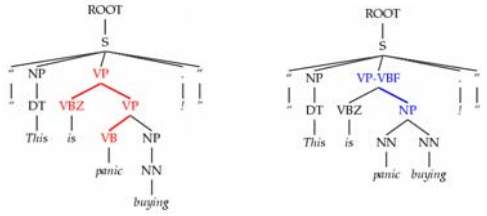
- It gives us distributional information



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

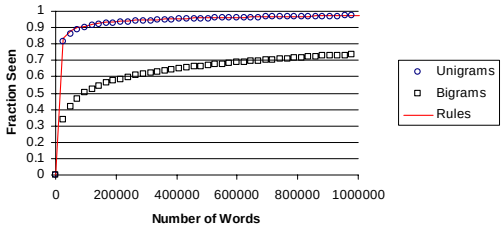
Corpus-Based Methods

- It lets us check our answers!



Corpus-Based Methods

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



The (Effective) NLP Cycle

- Pick a problem (usually 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

Example: POS Tagging

Decision Point

Local Context

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Decision Point

-3	-2	-1	0	+1
VBD	IN	DT	???	???
went	for	a	21-mile	hike

Example: NER Features

Decision Point:
State for *Grace*

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Weights

Feature Type	Feature	PERS	LOC
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	*<G*	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Language isn't Adversarial

- 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:
  - 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?

- One more class on classical NLP (parsing, semantic translation)
  - Sets the stage for statistical processing
  - Introduction to key ideas we'll need later
- Increasingly complex problems and models
  - Dealing with scale and sparsity
  - Sequence tasks (POS tagging, entity recognition)
  - Tree tasks (parsing, semantic interpretation)
  - Applications (translation, information extraction)
- Reading: M+S 3, J+M 1-3,10
- Assignment 0 will be distributed on Friday