Natural Language Processing
CS224N/Ling237

Christopher Manning
Spring 2006
Lecture 1

Course logistics in brief

- Instructor: Christopher Manning
- TA: Bill MacCartney
- Time: MW 11:00-12:15. (Section F 11:00-12:15)
- Handouts:
  - Course info, lecture 1, assignment 1, info sheet
- Programming language: Java 1.5
- Other information: see the webpage.
  - http://cs224n.stanford.edu/

This class

- Assumes you come with some skills...
  - Some basic linear algebra, probability, and statistics, decent programming skills
  - But not everyone has the same skills
    - Assume some ability to learn missing knowledge
- Teaches key theory and methods for statistical NLP, parsing, semantics, etc.
- But it's something like an "AI Systems" class:
  - A lot of it is hands on problem solving
  - Often practical issues are as important as theoretical niceties
  - We often combine a bunch of ideas

Natural language: the earliest UI

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave. I'm afraid I can't do that.

(cf. also false Maria in Metropolis – 1926)

Goals of the field of NLP

- Computers would be a lot more useful if they could handle our email, do our library research, chat to us ...
- But they are fazed by natural human languages.
  - Or at least their programmers are ... most people just avoid the problem and get into XML, or menus, drop boxes, and mice, or ...
- But someone has to work on the hard problems ...
  - How can we tell computers about language? (Or help them learn it as kids do?)

Course coverage

- Understand what NLP is about
- Particular subproblems:
  - word sense disambiguation, part-of-speech tagging, parsing, semantic representations
- Levels of analysis:
  - morphology, syntax, semantics, discourse
- Different approaches:
  - knowledge-based categorical grammars and statistical machine learning approaches
- Applications:
  - machine translation, information extraction, ...

Course goals

- Learn the basic principles and theoretical approaches underlying natural language processing
- Learn techniques and tools which can be used to develop practical, robust systems that can (partly) understand text or communicate with users in one or more languages
- Gain insight into many of the open research problems in natural language

What/where is NLP?

- Goals can be very far reaching …
  - True text understanding
  - Reasoning about texts
  - Real-time participation in spoken dialogs
- Or very down-to-earth …
  - Finding the price of products on the web
  - Context sensitive spell-checking
  - Analyzing reading level or authorship statistically
  - Extracting facts or relations from documents
- These days, the latter predominate (as NLP becomes increasingly practical, it is increasingly engineering-oriented – also related to changes in approach in AI/NLP)

The hidden structure of language

- We're going beneath the surface…
  - Not just string processing
  - Not just keyword matching in a search engine
    - ~~(Search Google on "tennis racquet" and "tennis racquets" and the results you get are completely different!)~~
    - Search Google on "laptop" and "notebook" and the results you get are completely different!
    - Not just converting a sound stream to a string of words
      - Like IBM, Dragon, Philips speech recognition
- We want to recover and manipulate at least *some* aspects of structure and meaning

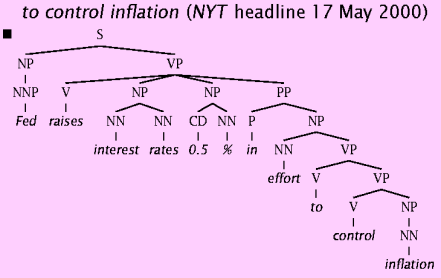
Is the problem just cycles?

- Bill Gates, Remarks to Gartner Symposium, October 6, 1997:
  - Applications always become more demanding. Until the computer can speak to you in perfect English and understand everything you say to it and learn in the same way that an assistant would learn -- until it has the power to do that -- we need all the cycles. We need to be optimized to do the best we can. Right now linguistics are right on the edge of what the processor can do. As we get another factor of two, then speech will start to be on the edge of what it can do.

The early history: 1950s

- Early NLP (Machine Translation) on machines less powerful than pocket calculators
- Foundational work on automata, formal languages, probabilities, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military, but basically just word substitution programs
- Little understanding of natural language syntax, semantics, pragmatics
- Problem soon appeared intractable

Why is NLU difficult? The hidden structure of language is hugely ambiguous

- Structures for: *Fed raises interest rates 0.5% in effort to control inflation* (*NYT* headline 17 May 2000)
- 

Where are the ambiguities?

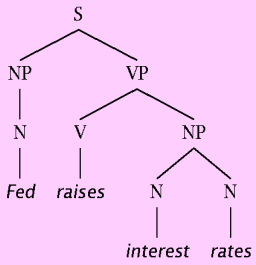
Part of speech ambiguities

									Syntactic attachment ambiguities
		VB							
	VBZ	VBP	VBZ						
NNP	NNS	NN	NNS	CD	NN				
Fed	*raises*	*interest*	*rates*	*0.5*	*%*	*in*	*effort*		
						to	*control*		
							inflation		

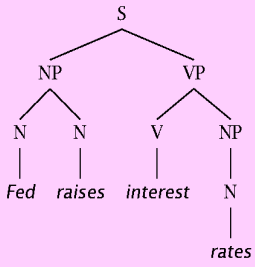
Word sense ambiguities: Fed → *"federal agent"*
interest → *a feeling of wanting to know or learn more*

Semantic interpretation ambiguities above the word level

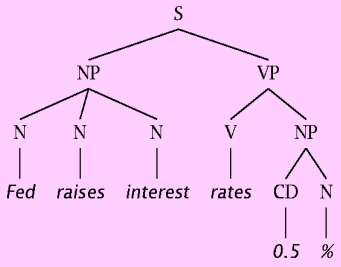
The bad effects of V/N ambiguities (1)



The bad effects of V/N ambiguities (2)



The bad effects of V/N ambiguities (3)



Why NLP is difficult: Newspaper headlines

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

LSAT / (former) GRE Analytic Section Questions

- Six sculptures – C, D, E, F, G, H – are to be exhibited in rooms 1, 2, and 3 of an art gallery.
  - Sculptures C and E may not be exhibited in the same room.
  - Sculptures D and G must be exhibited in the same room.
  - If sculptures E and F are exhibited in the same room, no other sculpture may be exhibited in that room.
  - At least one sculpture must be exhibited in each room, and no more than three sculptures may be exhibited in any room.
- If sculpture D is exhibited in room 3 and sculptures E and F are exhibited in room 1, which of the following may be true?
  - A. Sculpture C is exhibited in room 1
  - B. Sculpture H is exhibited in room 1
  - C. Sculpture G is exhibited in room 2
  - D. Sculptures C and H are exhibited in the same room
  - E. Sculptures G and F are exhibited in the same room

Reference Resolution

U: Where is A Bug's Life playing in Mountain View?
S: A Bug's Life is playing at the Century 16 theater.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show.
How much would that cost?

- Knowledge sources:
  - Domain knowledge
  - Discourse knowledge
  - World knowledge

Why is natural language computing hard?

- Natural language is:
  - highly ambiguous at all levels
  - complex and subtle
  - fuzzy, probabilistic
  - involves reasoning about the world
  - a key part of people interacting with other people (a social system):
    - persuading, insulting and amusing them
- But NLP can also be surprisingly easy sometimes:
  - rough text features can often do half the job

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- The answer that's been getting traction:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
- Some computer scientists think this is a new "A.I." idea
  - But really it's an old idea that was stolen from the electrical engineers….

CSLI Witas dialog system

Some demos

- Bell Labs Text-To-Speech
- Babelfish
- OneAcross
- AskJeeves (?)
- QA (AnswerBus)

Where do we head?

- Statistical machine translation
- Statistical NLP: classification and sequence models (part-of-speech tagging, named entity recognition, information extraction)
- Syntactic (probabilistic) parsing
- Building semantic representations from text
- Applications
- *(Unfortunately left out: natural language generation, phonology/morphology, speech dialogue systems, more on natural language understanding, …. There are other classes for some!)*