Natural Language Processing CS224N/Ling237

Christopher Manning
Spring 2003
Lecture 1
Handout #3

Course logistics

- Instructor: Christopher Manning
- TA: Rajat Raina
- Time: MW 1:15–2:30
- Handouts:
  - Course info, syllabus, lecture 1, info sheet
- Information sheet (fill it in!)
  - Section time?
- Programming language: officially Java?
- 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
  - Some ability to learn missing knowledge
- Teaches key theory and methods for parsing, statistical NLP, semantics, etc.
- But it's something like an "AI Systems" class:
  - A lot of it is hands on
  - Often practical issues dominate over theoretical niceties
  - We often combine a bunch of ideas

Natural language understanding

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

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

Natural Language Processing

- Natural Language Processing (NLP) is the study of the computational treatment of natural language:
  - most commonly Natural Language Understanding
  - The complementary task is Natural Language Generation
- NLP draws on research in Linguistics, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, etc.

What involved in NLP?

Morphology
Expectation Maximization
Lexicon
Information Content
Parsing
Syntax
Grammar
Corpora
Semantics
Ontologies
Recall & Precision
Discourse
Phonetics/Phonology
Pragmatics
Planning & Realization
Zipf's Law
Dynamic Programming

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 …
  - Searching 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

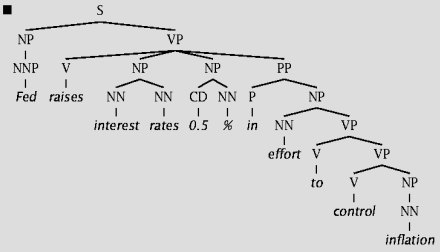
- We're going beneath the surface…
  - Not just string processing
  - Not just keyword matching in a search engine
    - (Search Google on "tennis racquet" and "tennis racquets" and the results you get are completely different!)
  - Not just converting a sound stream to a string of words
    - Like IBM, Dragon, Philips, (L&H) 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.

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?

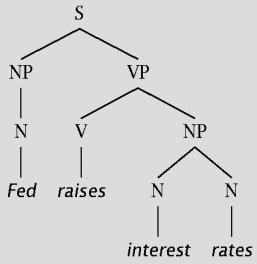
Part of speech ambiguities

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

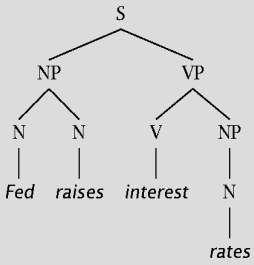
Word sense ambiguities: Fed → *"federal agent"*
interest → *a feeling of wanting to know or learn more*

Semantic interpretation ambiguities above the word level

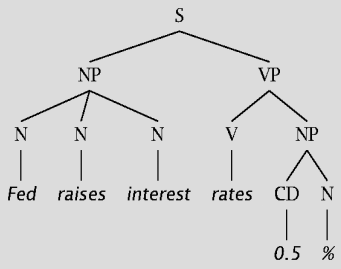
The bad effects of V/N ambiguities (1)



The bad effects of V/N ambiguities (2)



The bad effects of V/N ambiguities (3)



Newspaper headlines

- Ban on Nude Dancing on Governor's Desk – *from a Georgia newspaper column discussing current legislation*
- Lebanese chief limits access to private parts – *talking about an Army General's initiative*
- Death may ease tension – *an article about the death of Colonel Jean-Claude Paul in Haiti*
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree

Newspaper headlines

- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Man Struck by Lightning Faces Battery Charge
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

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

We learn and understand a changing system

- Neologisms:
  - mouse potato, nutraceutical, petabyte,
  - cuddle puddle, weaponize, squagel
  - afrosaxon, arm candy, bioterrorism
  - cybersquatting, cyberstalker
  - energy bar, fashionista, flexidoxy, hospitalist
  - keypal, infotainment, magalogue,
  - pop under (ad), desk rage
- Sentence structure too, though it's subtler

But there is syntax: Jabberwocky (Lewis Carroll)

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

Linguistic data is ubiquitous

- Most of the information in most companies, organizations, etc. is material in human languages (reports, customer email, web pages, discussion papers, text, sound, video) – not stuff in traditional databases
  - Estimates: 70%, 90% ?? [all depends how you measure]. Most of it.
- Most of that information is now available in digital form:
  - Estimate for companies in 1998: about 60% [CAP Ventures/Fuji Xerox]. More like 90% now?

NLP is applicable and used

- Many new applications are now feasible
  - Speech user interfaces/Voice portals
  - Rough machine translation for the web
  - Multisource news categorization and summarization
- Increasingly integrated as part of other applications
  - "Smart Tags" (Microsoft) inside documents via information extraction technology
  - Natural language query to databases (Microsoft e.g., again: English Query in MS SQL Server (improved wizards in new version!)

Some brief history. The 1940s: The Turing Test

- Alan Turing: the *Turing test* (language as test for intelligence)
- Three participants: a computer and two humans (one is an interrogator)
- Interrogator's goal: to tell the machine and human apart
- Machine's goal: to fool the interrogator into believing that a person is responding
- Other human's goal: to help the interrogator reach his goal

Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: 105621 (after a pause)

Some brief history: 1950s

- Early CL 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
- Transformational grammar: Chomsky 1957

Some brief history: 1960s

- Alvey report (1966) ends funding for MT in America – the lack of real results realized
- ELIZA (MIT): Fraudulent NLP in a simple pattern matcher psychotherapist
  - It's true, I am unhappy
  - *Do you think coming here will make you not to be unhappy?*
  - I need some help; that much seems certain
  - *What would it mean to you if you got some help?*
  - Perhaps I could learn to get along with my mother
  - *Tell me more about your family.*
- Early corpora: Brown Corpus (Kučera and Francis)

Some brief history: 1970s

- Winograd's SHRDLU (1971) existence proof of NLP (in tangled LISP code)
- Could interpret questions, statements, commands:
  - Which cube is sitting on the table?
  - *The large green one which supports the red pyramid*
  - Is there a large block behind the pyramid?
  - *Yes, three of them. A large red one, a large green cube, and the blue one.*
  - Put a small one onto the green cube which supports a pyramid
  - *OK*

Some brief history: 1980s

- Procedural → Declarative (including logic programming)
- Separation of processing (parser) from description of linguistic knowledge
- Common ground with linguists: new theories
- Representations of meaning: *procedural semantics* (SHRDLU), *semantic nets* (Schank), *logic* (perceived as answer; finally applicable to real languages (Montague)
- Perceived need for KR (Lenat and Cyc)
- Working MT in limited domains (METEO)

Some brief history: 1990s

- Resurgence of finite-state methods for NLP: in practice they are incredibly effective
- Speech recognition becomes widely usable
- Large amounts of digital text become widely available and reorient the field. The web.
- Resurgence of probabilistic/statistical methods, led by a few centers, especially IBM (speech, parsing, Candide MT system), often replacing logic for reasoning
- Recognition of *ambiguity* as key problem
- Emphasis on machine learning methods

Some brief history: 2000s

- A bit early to tell! But maybe:
  - Emphasis on meaning and knowledge representation
  - Emphasis on discourse and dialog
  - Strong integration of techniques, and levels: bringing together statistical NLP and sophisticated linguistic representations
  - Increased emphasis on unsupervised learning
  - More integration of NLP components into larger systems

Some demos

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

CSLI Witas dialog system

Where do we head?

- Sentence parsing
- Look at working with large text corpora and machine learning methods: word sense disambiguation
- Part of Speech tagging and Information extraction
- Building semantic representations from text
- Applications
- *(Unfortunately left out: NLG, phonology/morphology, speech dialogue systems, …)*

The End

Thank you!