Course Overview

Lecture #1

Introduction to Natural Language Processing
CMPSCI 585, Fall 2004

University of Massachusetts Amherst

Andrew McCallum

Today's Main Points

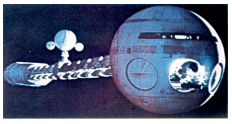
- Why is NLP interesting and difficult, complex and ambiguous.
  - Why? How to humans resolve this ambiguity?
- The six "layers" of NLP.
- NLP history, an overview, current successes.
- Course mechanics; what you can expect

1967

Stanley Kubrick, filmmaker 1928 - 1999

Arthur C. Clarke, author, futurist, 1917 -

HAL



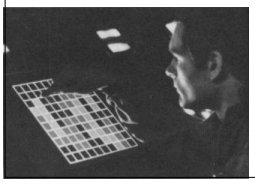
HAL's Capabilities

- Display graphics
- Play chess
- *Natural language production and understanding*

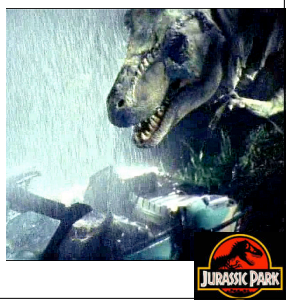
- Vision
- Planning
- Learning
- …

Graphics

HAL

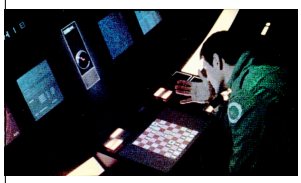


Now

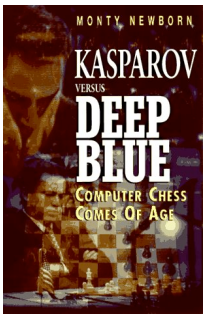


Chess

HAL



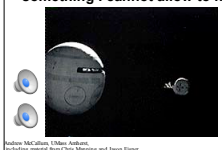
Now



Natural Language Understanding

HAL

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
…HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



Now

?

Many useful tools, but none that come even close to HAL's ability to communicate in natural language.

1950

Alan Turing
1912 - 1954

Turing Test

"Computing Machinery and Intelligence"
Mind, Vol. 59, No. 236, pp. 433-460, 1950

I propose to consider the question
"Can machines think?"…
We can only see a short distance ahead, but we can see plenty there that needs to be done.

Layers of Natural Language Processing

1. Phonetics & Phonology
2. Morphology
3. Syntax
4. Semantics
5. Pragmatics
6. Discourse

1. Phonetics & Phonology

The study of: language sounds, how they are physically formed; systems of discrete sounds, e.g. languages' syllable structure.

`dis-k&-'nekt`

`disconnect`

"It is easy to recognize speech."
"It is easy to wreck a nice beach."

JeetJet?

2. Morphology

The study of the sub-word units of meaning.

disconnect

"not" "to attach"

Even more necessary in some other languages, e.g. Turkish:

uygarlastiramadiklarimizdanmissinizcasina

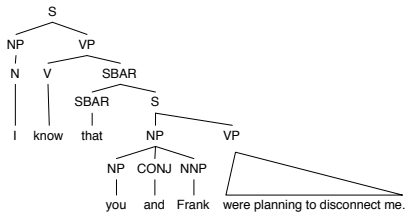
uygar las tir ama dik lar imiz dan mis siniz casina

(behaving) as if you are among those whom we could not civilize

3. Syntax

The study of the structural relationships between words.

I know that you and Frank were planning to disconnect me.



Not same structure:

You know me--Frank and I were planning to disconnect that.

4. Semantics

The study of the literal meaning.

I know that you and Frank were planning to disconnect me.

ACTION = disconnect
ACTOR = you and Frank
OBJECT = me

5. Pragmatics

The study of how language is used to accomplish goals.

What should you conclude from the fact I said something? How should you react?

I'm sorry Dave, I'm afraid I can't do that.

Includes notions of polite and indirect styles.

6. Discourse

The study of linguistic units larger than a single utterance.

The structure of conversations:
turn taking, thread of meaning.

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
…HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

Linguistic Rules

E.g. Morphology

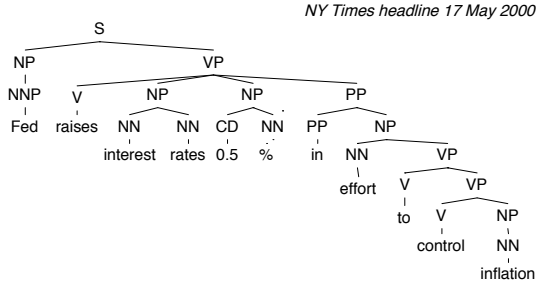
To make a word plural, add "s"

- dog → dogs
- baby → babies
- dish → dishes
- goose → geese
- child → children
- fish → fish (!)

Inherent Ambiguity in Syntax

```
Fed raises interest rates 0.5%
in effort to control inflation
```

NY Times headline 17 May 2000



Where are the ambiguities?

Part-of-speech ambiguities

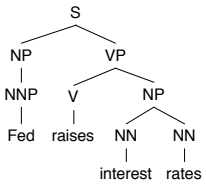
Syntactic attachment ambiguities

		VB				
	VBZ	VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

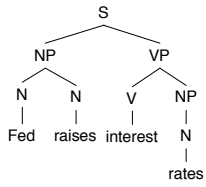
Word sense ambiguities: Fed →"federal agent"
interest →a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level.

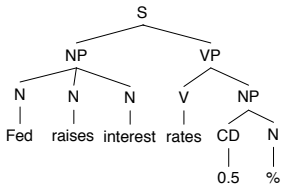
Effects of V/N Ambiguity (1)



Effects of V/N Ambiguity (2)



Effects of V/N Ambiguity (3)



Ambiguous Headlines

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Ban on Nude Dancing on Governor's Desk

What is grammatical and what isn't?

- John I believe Sally said Bill believed Sue saw.
- What did Sally whisper that she had secretly read?
- John wants very much for himself to win.
- Who did Jo think said John saw him?
- The boys read Mary's stories about each other.
- Mary, while John had had had had had had had had had was the correct answer.

What is grammatical and what isn't?

- John I believe Sally said Bill believed Sue saw.
- What did Sally whisper that she had secretly read?
- John wants very much for himself to win.
- Who did Jo think said John saw him?
- The boys read Mary's stories about each other.
- Mary, while John had had "had" had had "had had;" "had had" was the correct answer.

Language Evolves

- Morphology
  - We learn new words all the time: bioterrorism, cyberstalker, infotainment, thumb candy, energy bar
- Part-of-speech
  - Historically: "kind" and "sort" were always *nouns*: "I knowe that sorte of men ryght well." [1560]
  - Now also used as *degree modifiers:* "I'm sort of hungry." [Present] "It sort o' stirs one up to hear about old times." [1833]

Natural Language Computing is hard because

- Natural language is:
  - highly ambiguous at all levels
  - complex and subtle
  - fuzzy, probabilistic
  - involves reasoning about the world
  - embedded a social system of people interacting
    - persuading, insulting and amusing them
    - changing over time

Probabilistic Models of Language

To handle this ambiguity and to integrate evidence from multiple levels we turn to:

- Bayesian Classifiers (not rules)
- Hidden Markov Models (not DFAs)
- Probabilistic Context Free Grammars
- Maximum Entropy models
- …other tools of Machine Learning, AI, Statistics

Natural Language Processing

- Natural Language Processing (NLP) is the study of the computational treatment of natural languages:
  - 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 & Where is NLP

- Goals can be very far-reaching
  - True text understanding
  - Reasoning and decision-making from text
  - Real-time spoken dialog
- Or very down-to-earth
  - Searching the Web
  - Context-sensitive spelling correction
  - Analyzing reading-level or authorship statistically
  - Extracting company names and locations from news articles.
- These days, the later predominate (as NLP becomes increasingly practical, focused on performing measurably useful tasks *now*.
- Although language is complex, and ambiguity is pervasive, NLP can also be surprisingly easy sometimes:
  - rough text features often do half the job

Some brief history: 1950s

- Early NLP 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.

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 psycholtherapist
  - It's true, I am unhappy.
  - *Do you think coming here will make you not to be unhappy?*
  - I need some help; that much is certain.
  - *What would it mean to you if you got some help?*
  - Perhaps I could earn to get along with my mother.
  - *Tell me more about your family.*
- Early corpora: Brown Corpus (Kudera 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 cuube with supports a pyramid.
  - *OK.*

Some brief history: 1980s

- Procedural --> Declarative (including logic programming)
- Separation of processing (parser) from description of linguistic knowledge.
- 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)

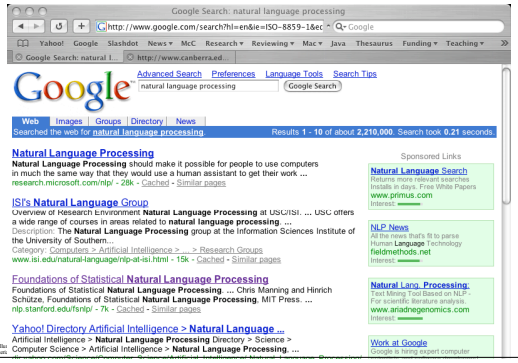
Some brief history: 1990s

- 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 probabilisitic/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.

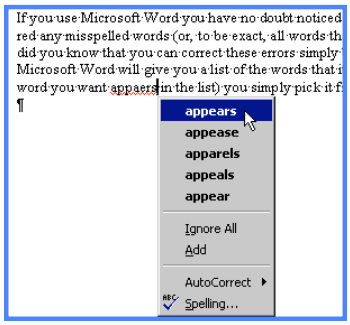
Some brief history: 2000s

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

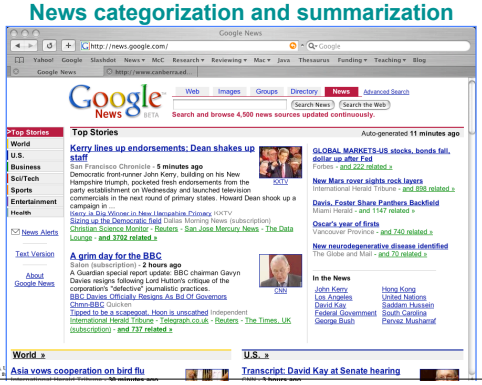
Example Applications of NLP



Example Applications of NLP:
MSWord spelling correction, grammar checking

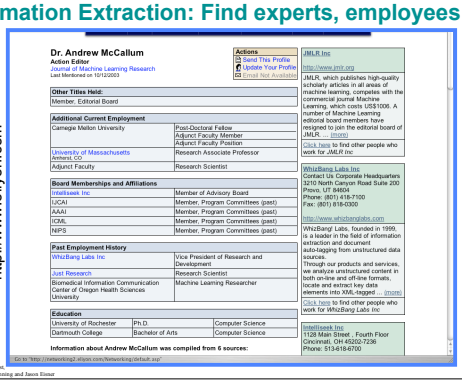


Example Applications of NLP:
News categorization and summarization

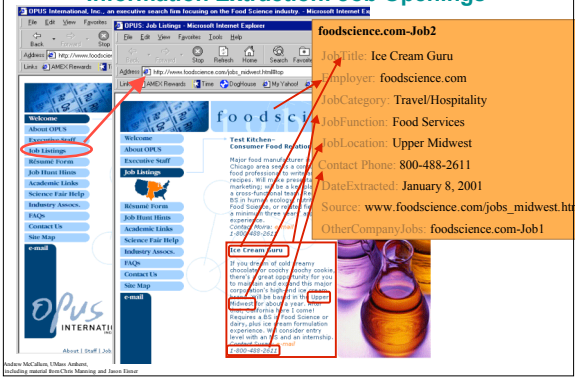


Example Applications of NLP
Information Extraction: Find experts, employees

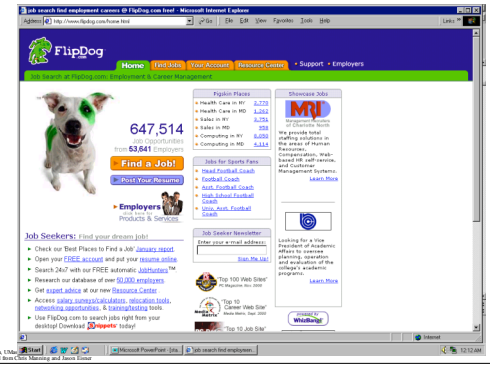
http://www.eliyon.com



Example Applications of NLP:
Information Extraction: Job Openings



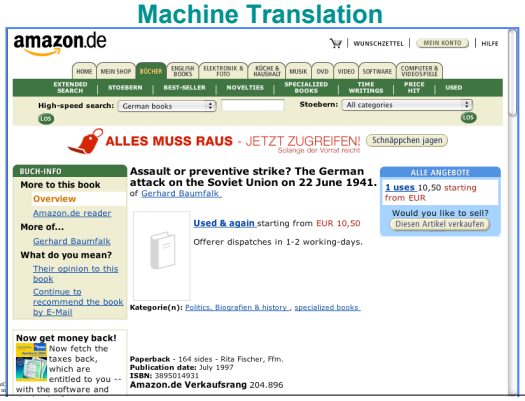
Example Applications of NLP: Information Extraction: Job Openings



Example Applications of NLP: Automatically Solving Crossword Puzzles

Example Applications of NLP: Question Answering

Example Applications of NLP: Machine Translation



Example Applications of NLP: Automatically generate Harlequin Romance novels?

Goals of the Course

- Introduce you to NLP problems and solutions.
- Relation to linguistics & statistics.
- Give you some hands-on practice with data and a handful of methods.

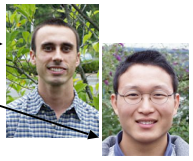
- At the end you should
  - Agree that language is subtle and interesting.
  - Feel some ownership over the formal & statistical models.
  - Be able to build some useful NLP system of your choosing.

This Class

- Assumes you come with some skills…
  - Some basic statistics, decent programming skills (in a language of your choice--although solutions will be in Java)
  - Some ability to learn missing knowledge
- Teaches key theory and methods for language modeling, tagging, parsing, etc.
- But it's something like an "AI Systems" class:
  - Hands on with data
  - Often practical issues dominate over theoretical niceties

Course Logistics

- Professor: Andrew McCallum
- TA: Aron Culotta
  Gary Huang



- Time: Tue/Thu 2:30-3:45pm
- Mailing list: `cs585@cs.umass.edu`
- More information on Web site:
  http://www.cs.umass.edu/~mccallum/courses/inlp2004

Grading

- 5 short written homeworks
  - should take less than 30 minutes each
  - some hands-on experience
  - help you set expectations for the mid-term and final
- 3 programming assignments
  - no way to really internalize without doing it
  - should be fun!
- Final project: with a partner
  - chance to explore a special interest at end of term
- Midterm & Final, and classroom participation

Syllabus Outline

- Grammars and parsing
- Foundations (probability & info theory)
- Language models, Spam filtering.
- Collocations, word clustering, disambiguation.
- Finite state machines, Markov models, Part-of-speech tagging.
- Modern parsing techniques.
- Information extraction, Semantics, Question answering, Dialog systems.

Recommended Reading

- Manning & Schutze
  - Chapter 11, section 1
    Context Free Grammars, topic of next class
- Manning & Schutze
  - Chapter 3, for background on linguistics.

Thank you!