Course Overview

Lecture #1

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

`http://www.cs.umass.edu/~mccallum/courses/cl2006`

Where to find syllabus, announcements, slides, homeworks

Today's Main Points

- Why is natural language interesting and difficult, complex and ambiguous.
  - Why? How to we resolve this ambiguity?
- Six "layers" of natural language
- Computational Linguistics history, an overview, current successes
- Get to know each other, and our motivations for being here
- Course mechanics; what you can expect

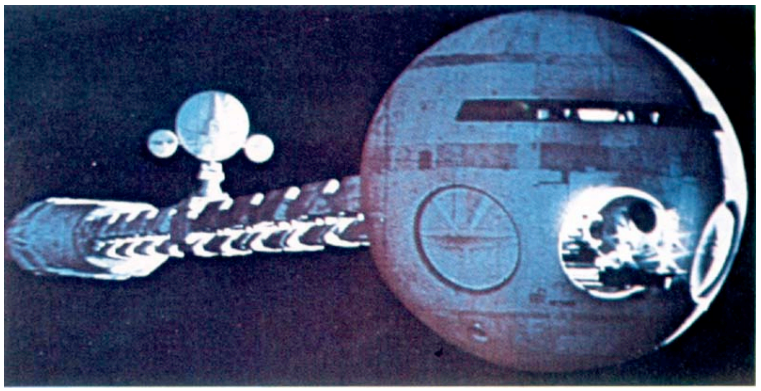
1967

Stanley Kubrick, filmmaker 1928 - 1999

Arthur C. Clarke, author, futurist, 1917 -

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

HAL



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

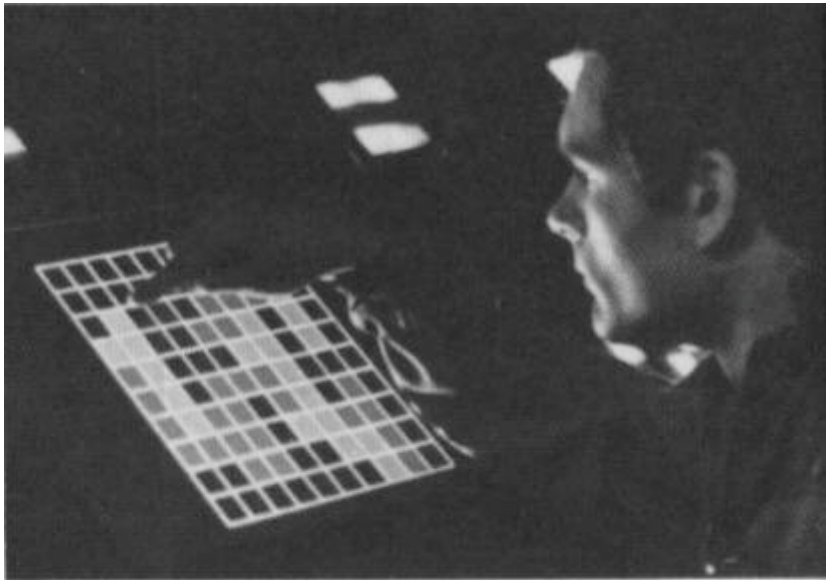
HAL's Capabilities

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

- Vision
- Planning
- Learning
- …

Graphics

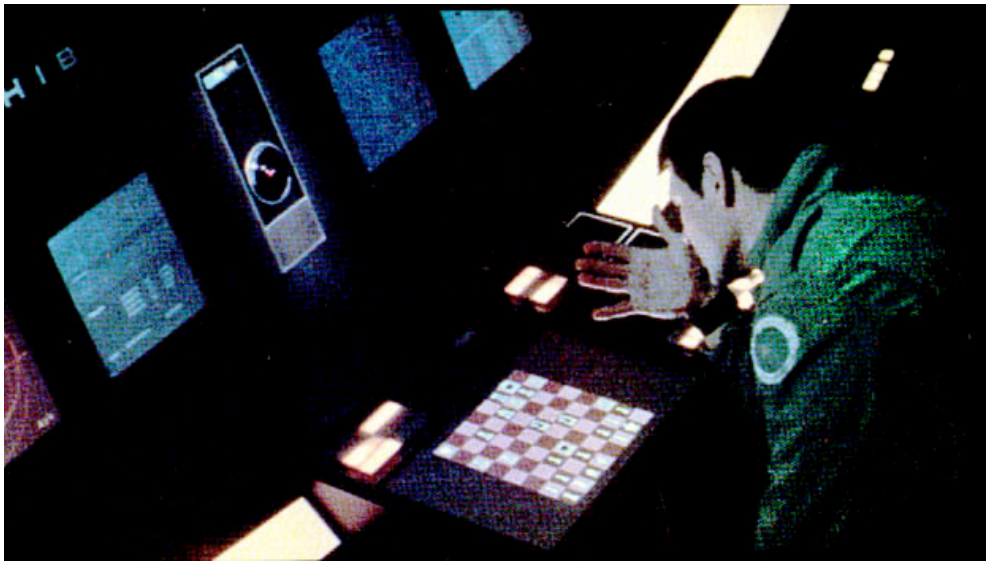
HAL



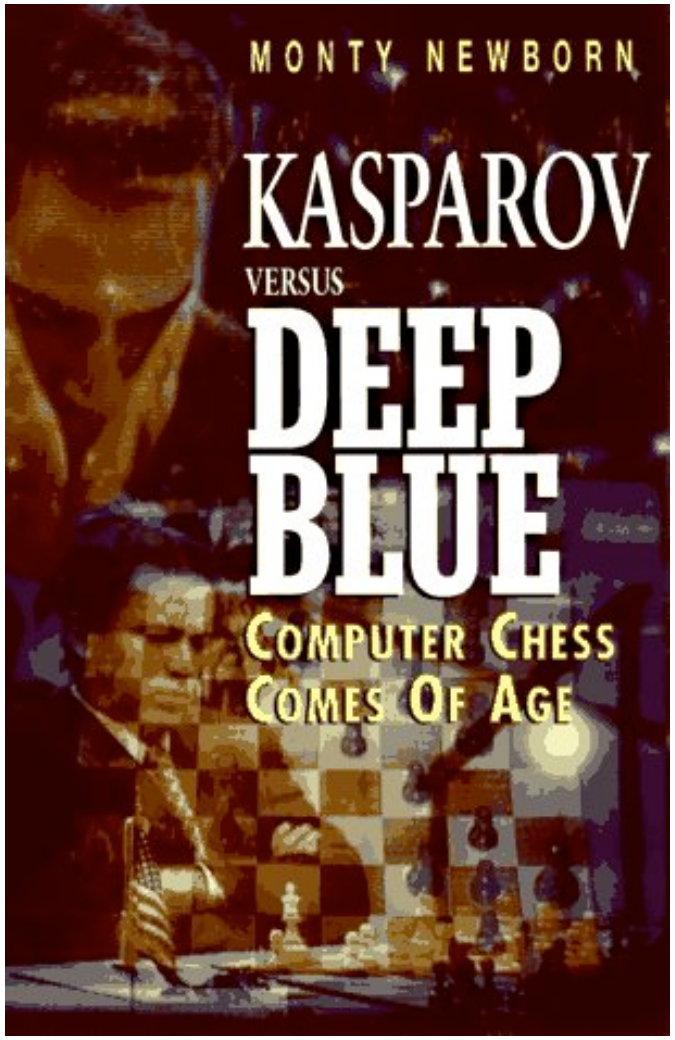
Now

Chess

HAL



Now



Natural Language Understanding

HAL

David Bowman:

Open the pod bay doors, Hal.

HAL:

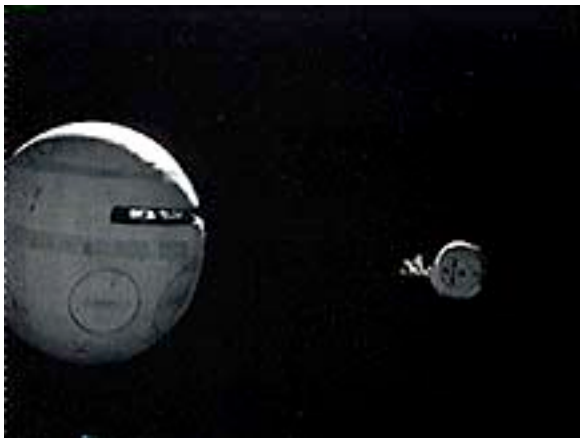
I'm sorry, Dave, I'm afraid I can't do that.

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.

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

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 Computational Linguistics

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?

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

2. Morphology

The study of the sub-word units of meaning.

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

uygarlastiramadiklarimizdanmissinizcasina

uygar las tir ama dik lar imiz dan mis siniz casina

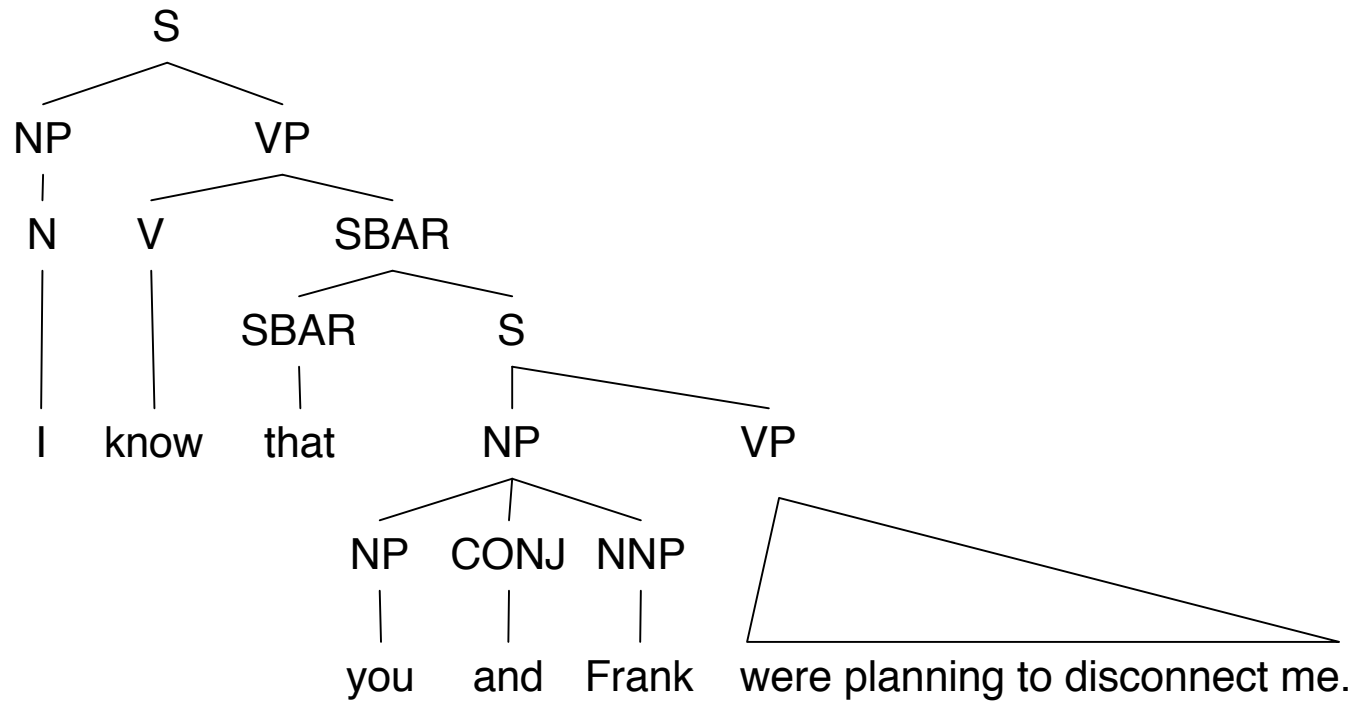
(behaving) as if you are among those whom we could not civilize

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

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

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

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.

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

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.

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Linguistic Rules

E.g. Morphology

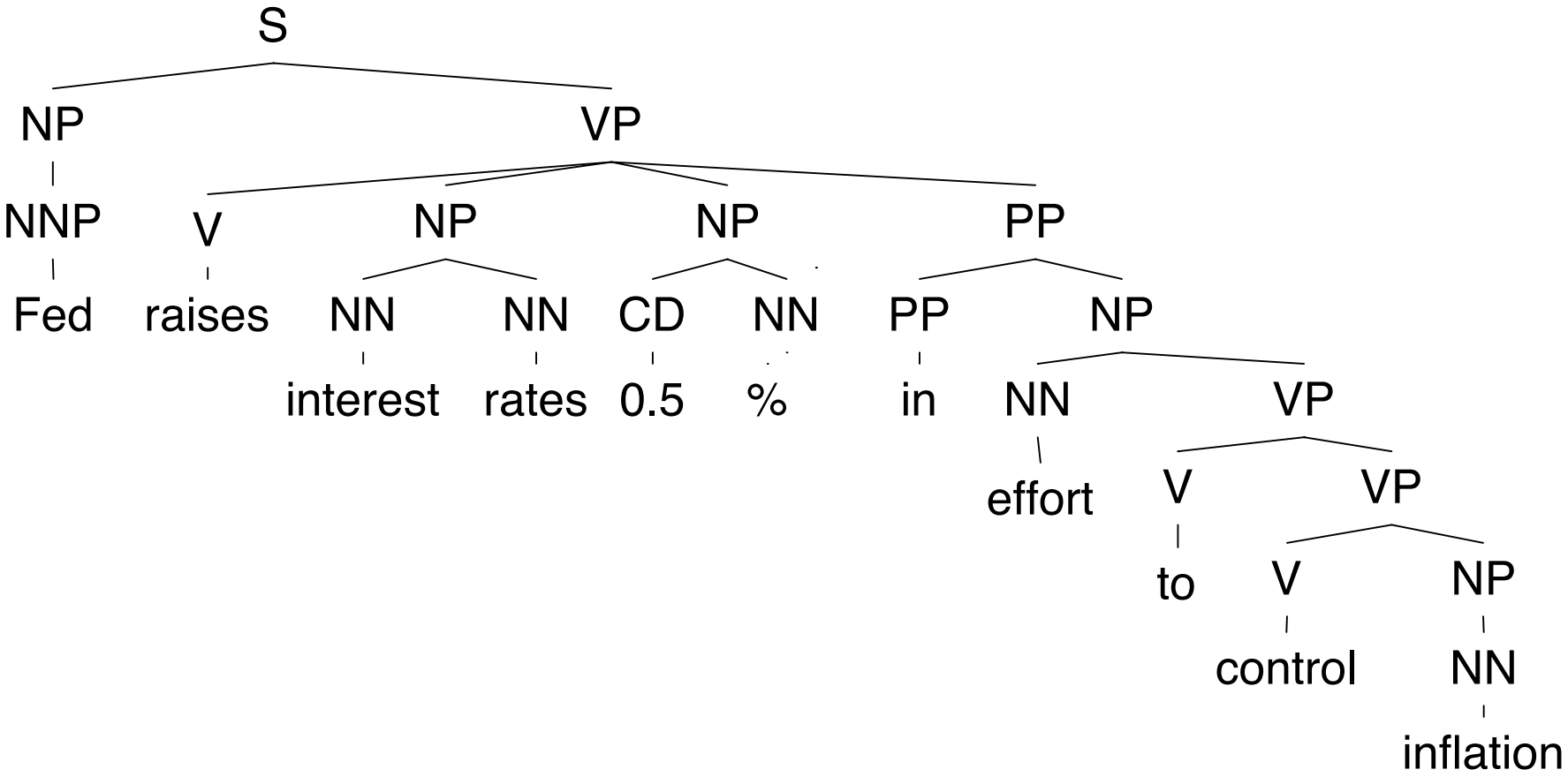
To make a word plural, add “s”

- 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

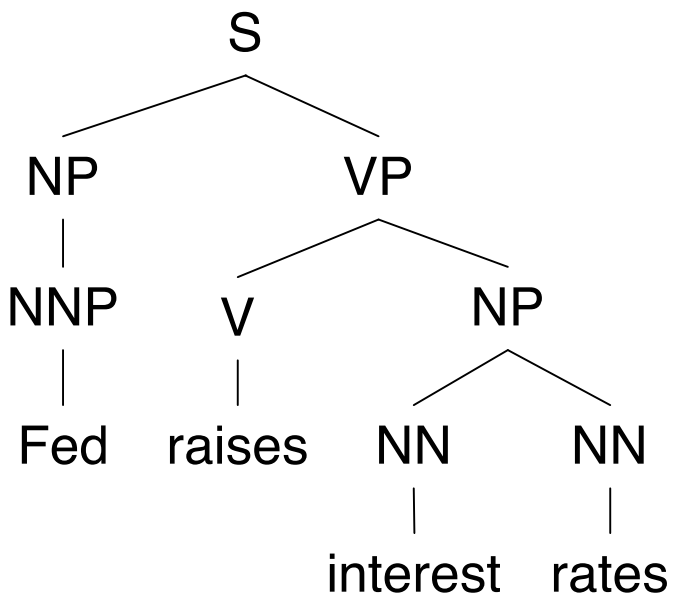
Syntactic attachment ambiguities

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

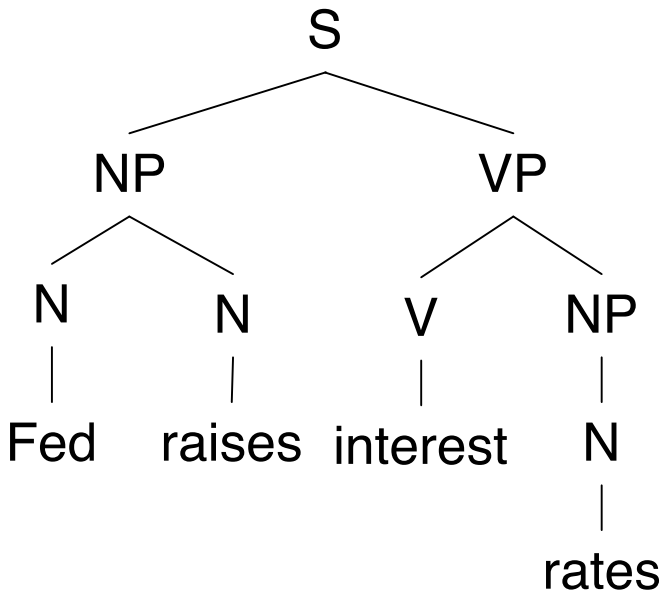
Word sense ambiguities: Fed →"federal agent"
interest →a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level.

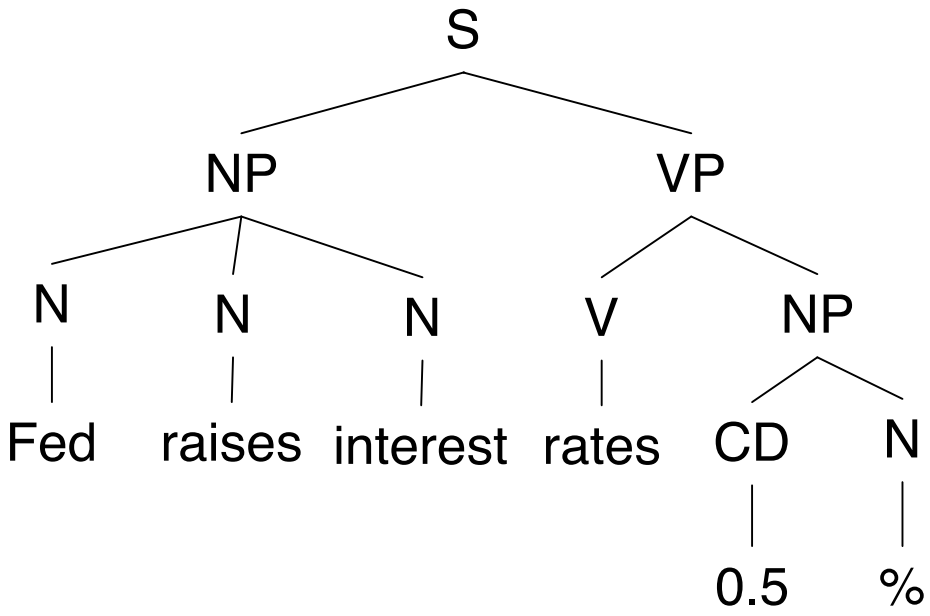
Effects of V/N Ambiguity (1)



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
- Local HS Dropouts Cut in Half
- 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
  - interpretation involves ***combining evidence***
  - involves reasoning about the world
  - embedded a social system of people interacting
    - persuading, insulting and amusing them
    - changing over time

Probabilistic Models of Language

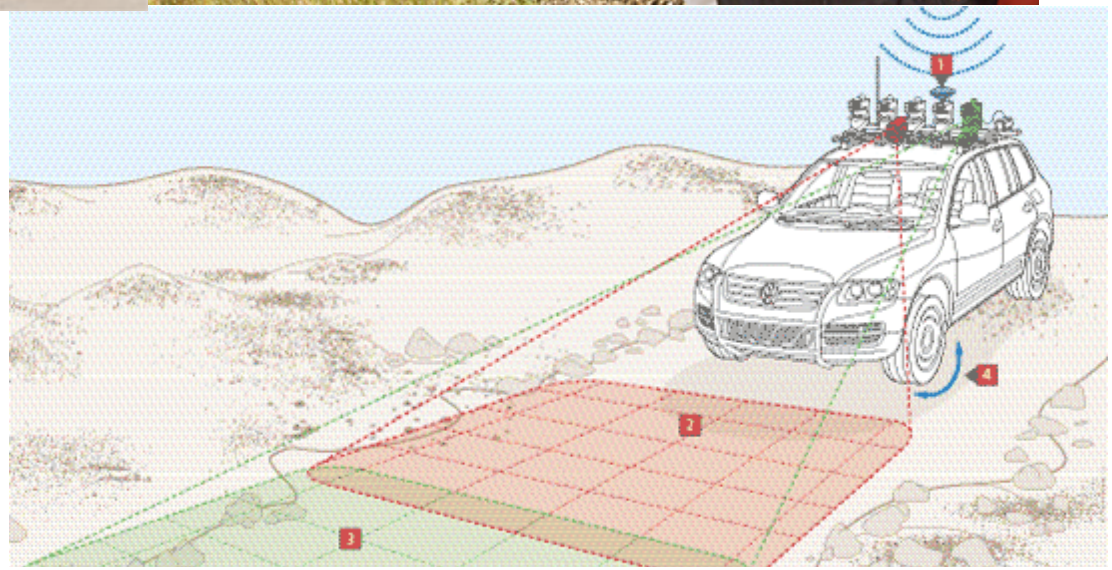
To handle this **ambiguity** and to **integrate evidence** from multiple levels we turn to:

The tools of probability:

- Bayesian Classifiers (not rules)
- Hidden Markov Models (not DFAs)
- *Probabilistic* Context Free Grammars

- …other tools of Machine Learning, AI, Statistics

Another Area where Probabilistic Combination of Evidence Won



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Natural Language Processing

- Natural Language Processing (NLP) studies how to get computers to do useful things with natural languages:
  - Most commonly Natural Language Understanding
  - The complementary task is Natural Language Generation

Engineering Goal

- NLP draws on research in Linguistics, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, 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

Linguistics

- Linguistics is the study of natural languages:
  - Understanding this naturally-occurring phenomenon.
  - Structure, meaning, how acquired, differences and commonalities across languages.

Scientific Goal

- Linguistics draws on research in Natural Language Processing, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.

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.

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 cube 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)

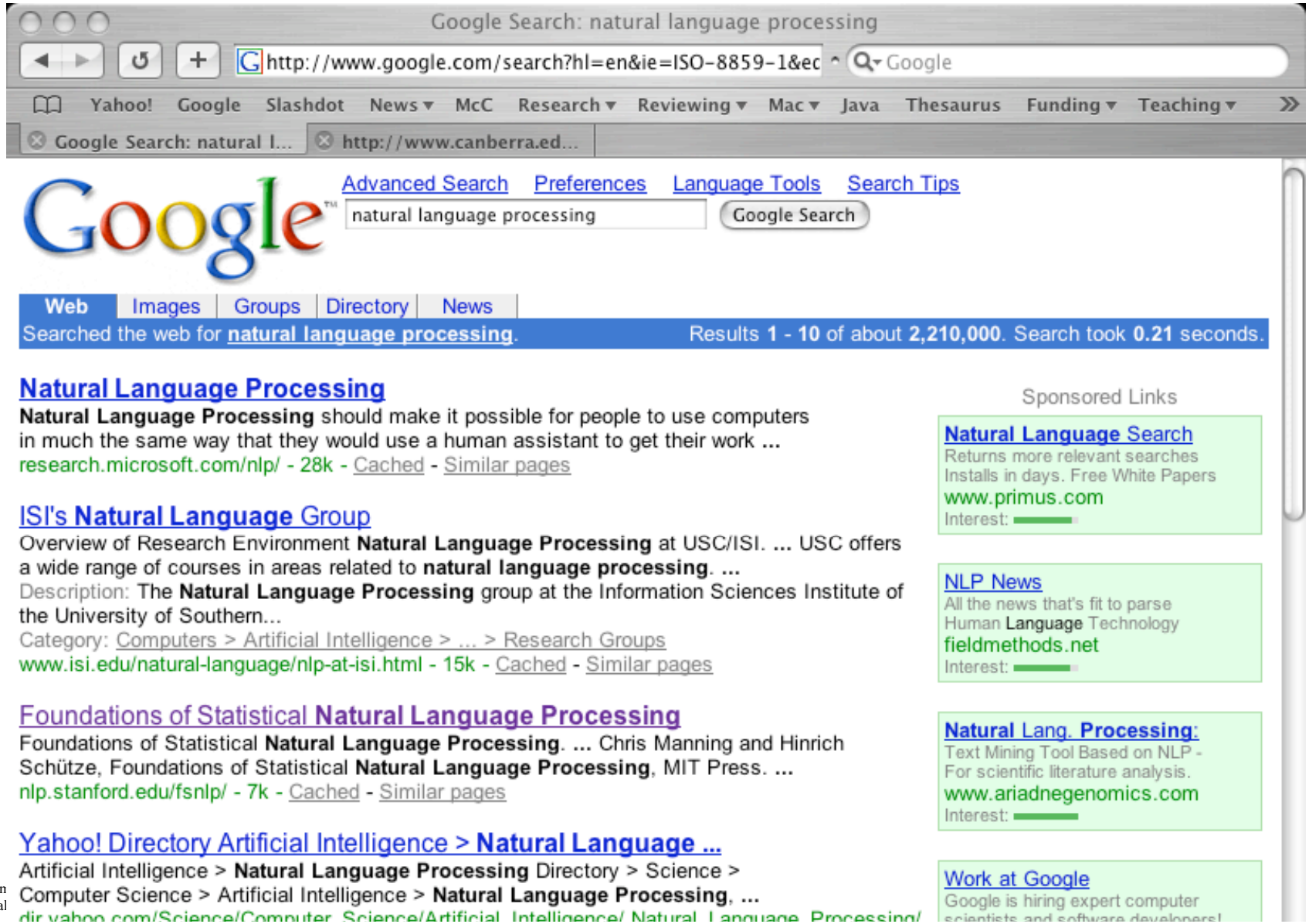
Some brief history: 1990s

- 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 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

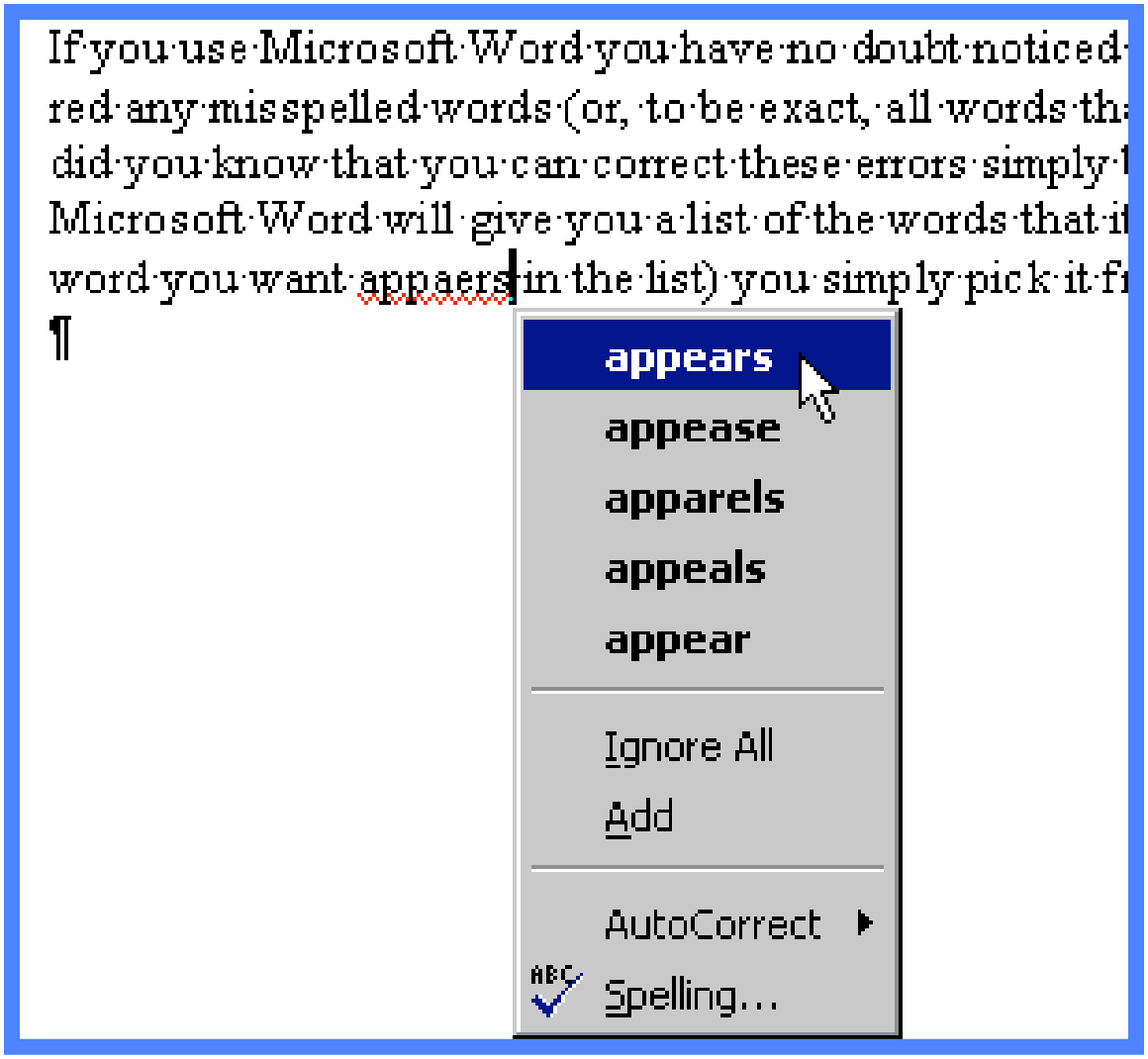
- A bit early to tell! But maybe:
  - Continued surge in probability, Bayesian methods of evidence combination, and joint inference.
  - Emphasis on meaning and knowledge representation.
  - Emphasis on discourse and dialog.
  - Strong integration of techniques, and levels: brining together statistical NLP and sophisticated linguistic representations.
  - Increased emphasis on unsupervised learning.

Example Applications of NLP



Andrew McCallun
including material

Example Applications of NLP:
MSWord spelling correction, grammar checking



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

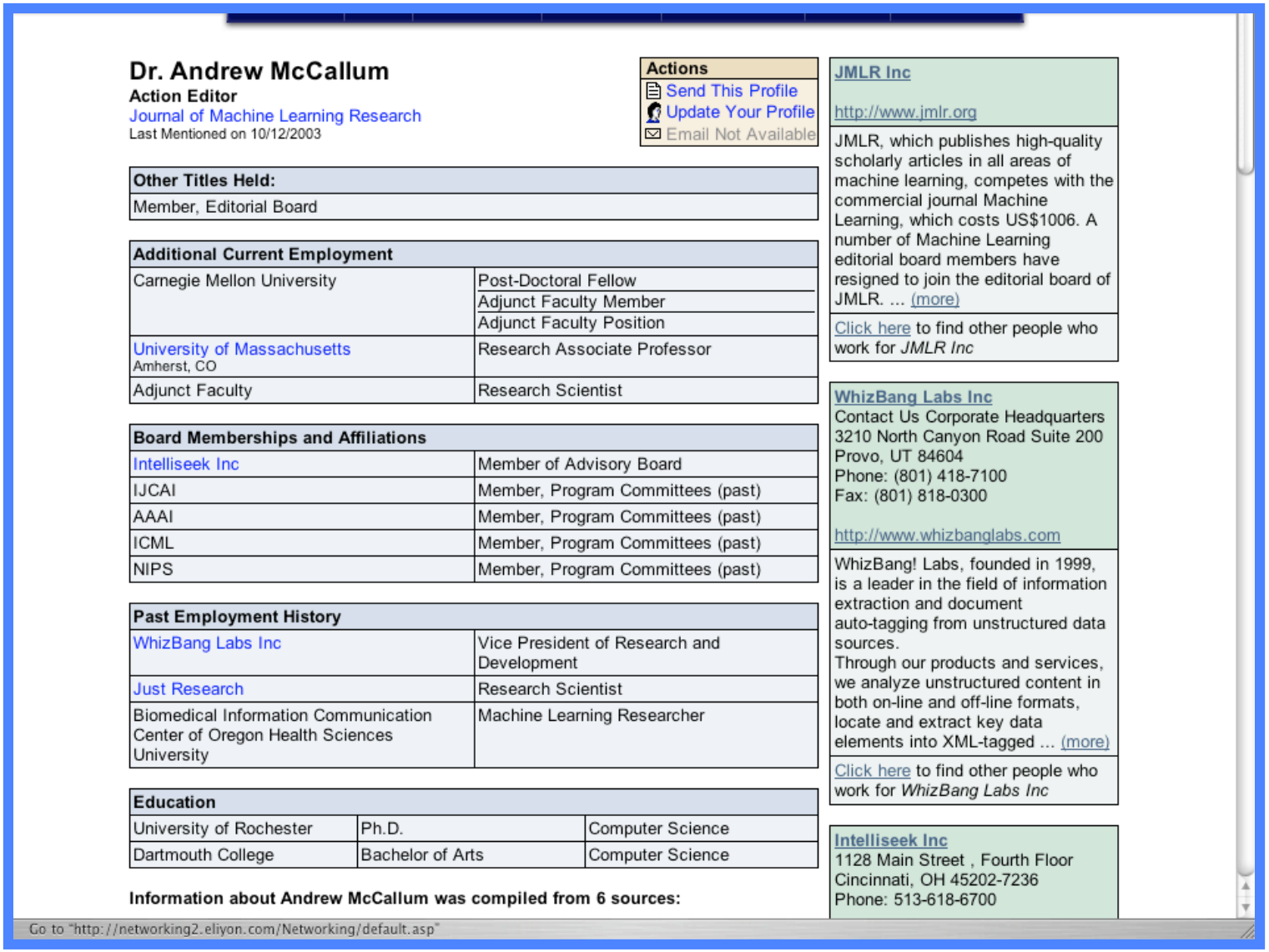
Example Applications of NLP:

Andrew McCallum
including material

Example Applications of NLP

Information Extraction: Find experts, employees

http://www.eliyon.com



Dr. Andrew McCallum
Action Editor
Journal of Machine Learning Research
Last Mentioned on 10/12/2003

Actions
Send This Profile
Update Your Profile
Email Not Available

Other Titles Held:
Member, Editorial Board

Additional Current Employment	
Carnegie Mellon University	Post-Doctoral Fellow
	Adjunct Faculty Member
	Adjunct Faculty Position
University of Massachusetts Amherst, CO	Research Associate Professor
Adjunct Faculty	Research Scientist

Board Memberships and Affiliations	
Intelliseek Inc	Member of Advisory Board
IJCAI	Member, Program Committees (past)
AAAI	Member, Program Committees (past)
ICML	Member, Program Committees (past)
NIPS	Member, Program Committees (past)

Past Employment History	
WhizBang Labs Inc	Vice President of Research and Development
Just Research	Research Scientist
Biomedical Information Communication Center of Oregon Health Sciences University	Machine Learning Researcher

Education		
University of Rochester	Ph.D.	Computer Science
Dartmouth College	Bachelor of Arts	Computer Science

Information about Andrew McCallum was compiled from 6 sources:

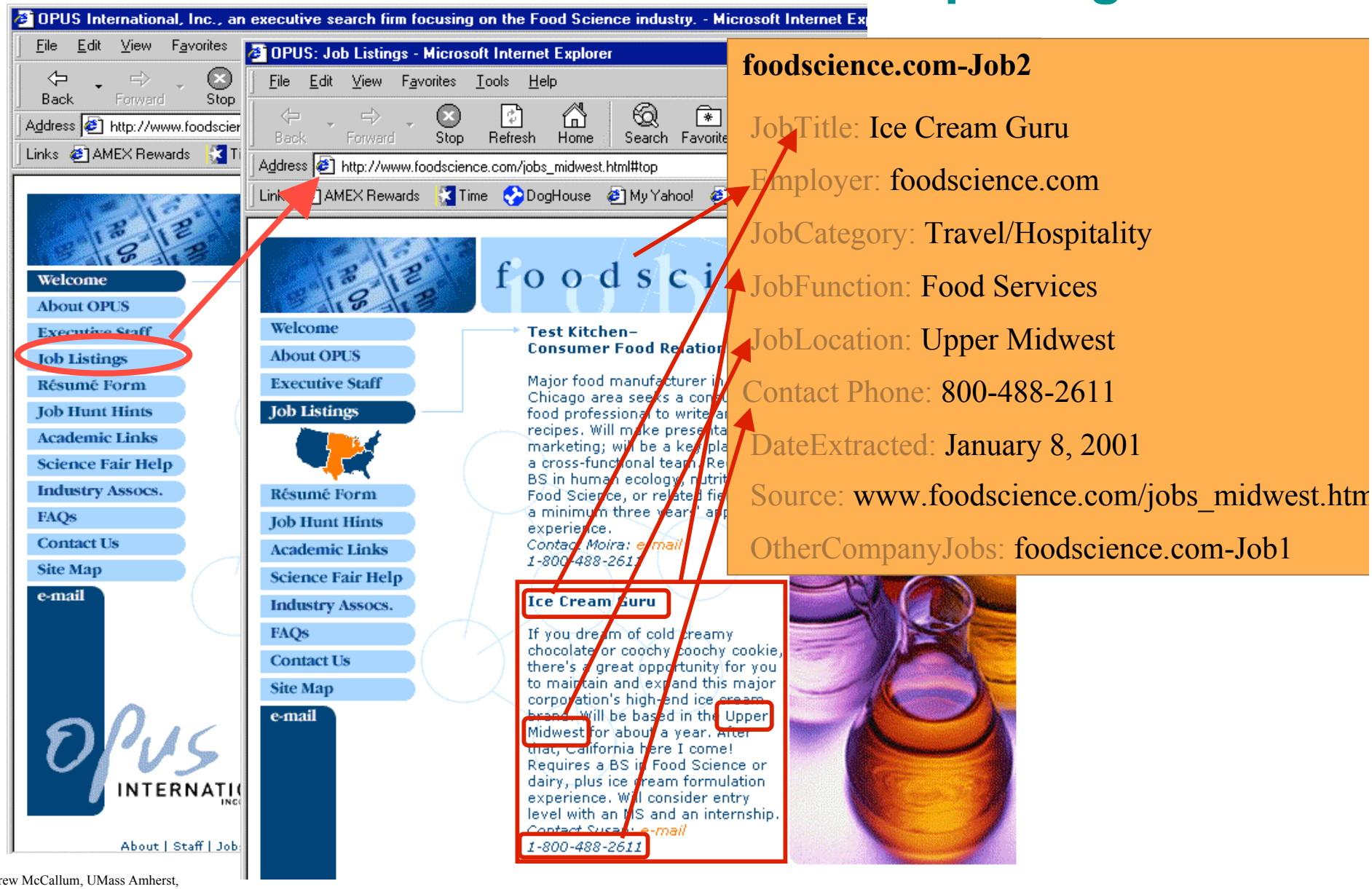
JMLR Inc
http://www.jmlr.org
JMLR, which publishes high-quality scholarly articles in all areas of machine learning, competes with the commercial journal Machine Learning, which costs US$1006. A number of Machine Learning editorial board members have resigned to join the editorial board of JMLR. ... (more)
Click here to find other people who work for *JMLR Inc*

WhizBang Labs Inc
Contact Us Corporate Headquarters
3210 North Canyon Road Suite 200
Provo, UT 84604
Phone: (801) 418-7100
Fax: (801) 818-0300
http://www.whizbanglabs.com
WhizBang! Labs, founded in 1999, is a leader in the field of information extraction and document auto-tagging from unstructured data sources.
Through our products and services, we analyze unstructured content in both on-line and off-line formats, locate and extract key data elements into XML-tagged ... (more)
Click here to find other people who work for *WhizBang Labs Inc*

Intelliseek Inc
1128 Main Street , Fourth Floor
Cincinnati, OH 45202-7236
Phone: 513-618-6700

Go to "http://networking2.eliyon.com/Networking/default.asp"

Example Applications of NLP: Information Extraction: Job Openings



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Example Applications of NLP: Information Extraction: Job Openings

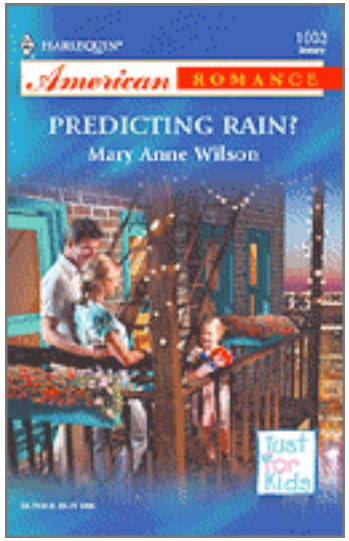
Example Applications of NLP: Automatically Solving Crossword Puzzles

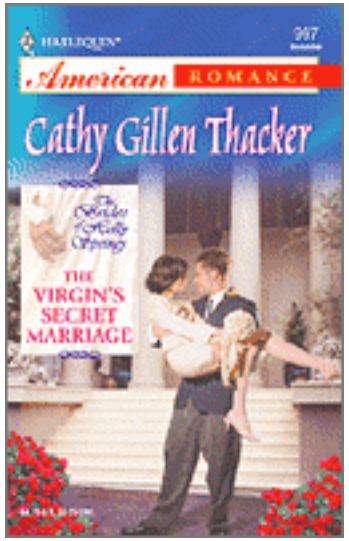
Example Applications of NLP: Question Answering

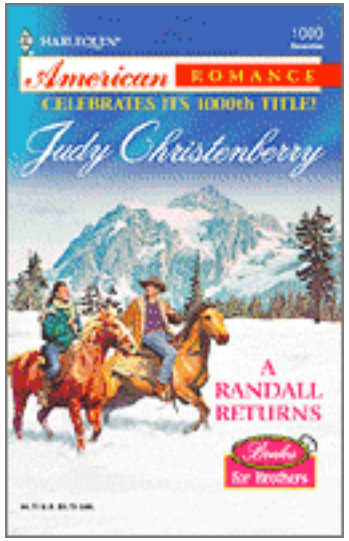
Andrew M
including

Example Applications of NLP: Machine Translation

Example Applications of NLP:
Automatically generate Harlequin Romance novels?







Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Goals of the Course

- Introduce you to Computational Linguistics problems and solutions.
- Ultimate focus on handling ambiguity by probabilistic integration of evidence.
- Using these tools to empower Linguistics' goals. (More on this next class.)
- Give you some hands-on practice with data and a handful of methods.

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
- TAs: Pallika Kanani
  Greg Druck
  Assistant: Gideon Mann

- Time: Tue/Thu 2:30-3:45pm

- Mailing list: **`compling-class@cs.umass.edu`**

- More information on Web site:
  http://www.cs.umass.edu/~mccallum/courses/cl2006

Grading

- 6 short written homework / programming assignments.
  - no way to really internalize without doing it
  - some hands-on experience
  - should be fun!
  - should take about 1-2 hours each.
- Random, informal in-class “collaborative quizzes”
  - help you set expectations for the mid-term and final
- Final project: with a small team, mixed CS & Ling
  - chance to explore a special interest at end of term
- Midterm & Final, and classroom participation

Programming? Yipes!

- Yes, but with *extensive* support for those w/out experience.

- Historically popular language for CL courses:
  - Prolog (clean, hard to learn, counter-intuitive)
  - Perl (quick, but obfuscated syntax, messy to read)
  - Interpreted, rapid prototyping
- Why **Python** is better-suited:
  - easy to learn, clean syntax, powerful features
  - becoming increasingly popular in CompLinguistics!
  - Extensive tutorials, CompLing support, toolkits, data, etc.
- Many CS students don't know it either: put you on more equal footing.

Syllabus Outline

- Two halves:
  - First: hands-on course, introductory, methods, HW
  - Second: more like a seminar + project
- First half:
  - Language, structures, and computation
  - Foundation of probability and information theory
  - Use those foundations to work with language

- Example topics:
  - Language models, language prediction, spam filtering.
  - Collocations, word clustering, word sense disambiguation.
  - Finite state machines, Markov models, Part-of-speech tagging.
  - Modern parsing techniques.
  - Information extraction, semantics, question answering, discourse.

This Week

- Visit course Web site, browse around.
- Read Chapters 1 and 2 in Jurafsky & Martin textbook
- Install Python on your computer
  - Get extensive help from the TAs if you like!

Thank you!