

Course Overview

Lecture #1

Introduction to Natural Language Processing
CMPSCI 585, Spring 2004

University of Massachusetts Amherst



Andrew McCallum

Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein

1967

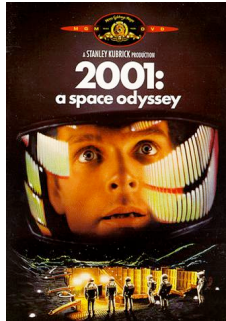


Stanley Kubrick,
filmmaker
1928 - 1999



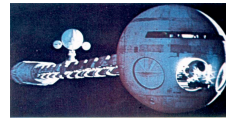
Arthur C. Clarke,
author, futurist,
1917 -

Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein



Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein

HAL



Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein

HAL's Capabilities

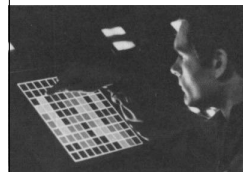
- Display graphics
- Play chess
- *Natural language production and understanding*

- Vision
- Planning
- Learning
- ...

Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein

Graphics

HAL



Now



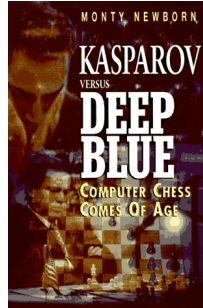
Andrew McCallum, UMass Amherst.
Including material from Chris Manning and Jacob Eisenstein

Chess

HAL



Now



Andrew McCallum, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

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, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

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.

Andrew McCallum, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

Layers of Natural Language Processing

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

Andrew McCallum, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

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, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

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

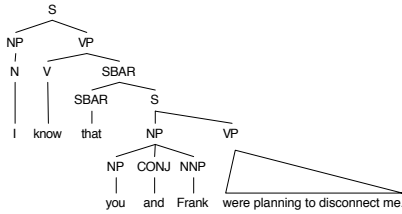
(behaving) as if you are among those whom we could not civilize

Andrew McCallum, CMU, Alex Ambrose, including material from Chris Manning and Jacob Devlin

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.

Andrew McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

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 McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

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 McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

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 McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

Linguistic Rules

E.g. Morphology

To make a word plural, add "s"

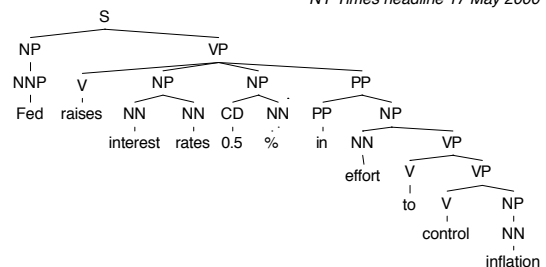
- dog → dogs
- baby → babies
- dish → dishes
- goose → geese
- child → children
- fish → fish (!)

Andrew McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

Inherent Ambiguity in Syntax

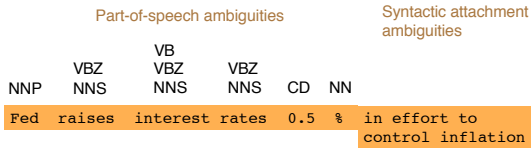
Fed raises interest rates 0.5%
in effort to control inflation

NY Times headline 17 May 2000



Andrew McCallen, UMass Amherst, including material from Chris Manning and Jacob Eisenstein

Where are the ambiguities?

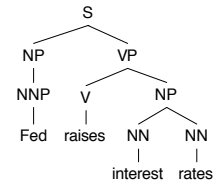


Word sense ambiguities: Fed → "federal agent"
 interest → a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level.

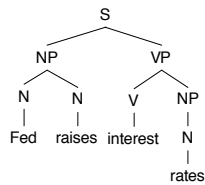
Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

Effects of V/N Ambiguity (1)



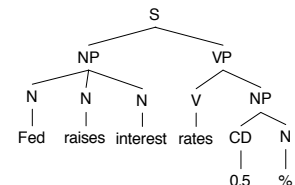
Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

Effects of V/N Ambiguity (2)



Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

Effects of V/N Ambiguity (3)



Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

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

Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

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 had was the correct answer.

Andrew McCallum, CMU, Stanford, including material from Chris Manning and Jacob Eisenstein

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.

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

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]

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

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

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

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

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

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.

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

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

Andreas McCallum, CMU, Andrew Senior, including material from Chris Manning and Jacob Eisenstein

Example Applications of NLP

A screenshot of a Google search for "natural language processing". The search results page shows several relevant links, including "Natural Language Processing" from Microsoft Research, "US's Natural Language Group" from USC, and "Foundations of Statistical Natural Language Processing" from MIT Press. The page also displays a sidebar with "Sponsored Links" and "NLP News".

Example Applications of NLP: MSWord spelling correction, grammar checking

If you use Microsoft Word you have no doubt noticed red any misspelled words (or, to be exact, all words that did you know that you can correct these errors simply Microsoft Word will give you a list of the words that it word you want appears in the list you simply pick it f

A screenshot of a Microsoft Word spelling correction dropdown menu. The word "appears" is highlighted in red. The dropdown menu shows a list of suggestions: "appears", "appease", "apparels", "appeals", and "appear". Below the list are buttons for "Ignore All", "Add", "AutoCorrect", and "Spelling...".

Example Applications of NLP: News categorization and summarization

A screenshot of Google News showing a grid of news stories categorized by region (World, U.S., Business, Sci/Tech, Sports, Entertainment, Health). Each story includes a headline, a brief summary, and a thumbnail image. For example, under "U.S.", there is a story about Kerry's endorsement of Dean.

Example Applications of NLP: Information Extraction: Find experts, employees

A screenshot of a professional profile page for Dr. Andrew McCallum. The page is organized into sections: "Other Titles Held", "Additional Current Employment", "Board Memberships and Affiliations", "Past Employment History", "Job Research", and "Education". Each section lists various roles, institutions, and dates. A sidebar on the right contains contact information and a bio.

Example Applications of NLP: Information Extraction: Job Openings

A screenshot of a job opening page for "Ice Cream Guru" at foodscience.com. The page is annotated with red boxes and arrows highlighting key information extracted by NLP, such as the job title, employer, location, and contact details. The annotations include: "Job Title: Ice Cream Guru", "Employer: foodscience.com", "Job Category: Travel/Hospitality", "Job Function: Food Services", "Job Location: Upper Midwest", "Contact Phone: 800-488-2611", "Date Extracted: January 8, 2001", and "Source: www.foodscience.com/jobs_midwest.htm".

Example Applications of NLP: Information Extraction: Job Openings

A screenshot of a job opening page for "FlipDog". The page features a large image of a dog and a prominent phone number "647.514.6384". The page is annotated with red boxes and arrows highlighting key information extracted by NLP, such as the phone number and the company name. The annotations include: "Phone Number: 647.514.6384" and "Company Name: FlipDog".

Example Applications of NLP: Automatically Solving Crossword Puzzles

The screenshot shows the OneAcross website interface. At the top, there's a navigation bar with links like Home, Crosswords, Cryptograms, Anagrams, Reference, Forum, and Languages. Below that, there's a search bar and a 'Crossword Clue Search' section. A text box contains the clue: 'Having trouble getting the last word in that puzzle? Having trouble getting the first? See if our search engine can help! Unlike pure pattern dictionary searches, we actually analyze the clue as well.' Below the text box are input fields for 'Clue:' and 'Pattern:'. To the right, there's a 'Get!' button. Further down, there's a 'How to Search:' section explaining the search process. At the bottom, there are examples of clues and patterns, such as 'Clue: Trout Basket' and 'Pattern: 5', and 'Clue: Out' and 'Pattern: 2121'.

Example Applications of NLP: Question Answering

The screenshot shows the AnswerBus website interface. At the top, there's a navigation bar with links like Home, Crosswords, Cryptograms, Anagrams, Reference, Forum, and Languages. Below that, there's a search bar and a 'Crossword Clue Search' section. A text box contains the clue: 'Having trouble getting the last word in that puzzle? Having trouble getting the first? See if our search engine can help! Unlike pure pattern dictionary searches, we actually analyze the clue as well.' Below the text box are input fields for 'Clue:' and 'Pattern:'. To the right, there's a 'Get!' button. Further down, there's a 'How to Search:' section explaining the search process. At the bottom, there are examples of clues and patterns, such as 'Clue: Trout Basket' and 'Pattern: 5', and 'Clue: Out' and 'Pattern: 2121'.

Example Applications of NLP: Machine Translation

The screenshot shows the Amazon.de website interface. At the top, there's a navigation bar with links like Home, Mein Konto, Suchen, and Hilfe. Below that, there's a search bar and a 'High-speed search:' section. A text box contains the clue: 'Assault or preventive strike? The German attack on the Soviet Union on 22 June 1941.' Below the text box are input fields for 'Clue:' and 'Pattern:'. To the right, there's a 'Get!' button. Further down, there's a 'How to Search:' section explaining the search process. At the bottom, there are examples of clues and patterns, such as 'Clue: Trout Basket' and 'Pattern: 5', and 'Clue: Out' and 'Pattern: 2121'.

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 
- Time: Tue/Thu 1-2:15pm
- Mailing list: cs585@cs.umass.edu
- More information on Web site:
<http://www.cs.umass.edu/~mccallum/courses/inlp2004>

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein

Take Home Points for Today

- Six layers of language
 - Phonetics, Morphology, Syntax, Semantics, Pragmatics, Discourse.
- Language is complex, ambiguous.
 - Why? How do humans resolve this ambiguity?
- NLP definition, goals, theoretical tools, current successes.

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein

Thank you!

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein

Syllabus Outline

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein

Inherent Amiguities

- " Example sentence: "I made her duck"•I cooked waterfowl for her. •I cooked waterfowl belonging to her. •I created the (plaster?) duck she owns. •I caused her to quickly lower her head or body. •I waved my magic wand and turned her into undifferentiated waterfowl. •(Explain all these ambiguities in linguistic terms. See Jur&Martin, p. 4.

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jacob Eisenstein