Course Overview

Lecture #1

Computational Linguistics

CMPSCI 585, Fall 2007

University of Massachusetts Amherst

Andrew McCallum

http://www.cs.umass.edu/~mccallum/courses/inlp2007

Where to find syllabus, announcements, slides, homeworks

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

• Natural Language Processing 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 -
HAL
HAL’s Capabilities

• Display graphics
• Play chess
• Natural language production and understanding

• Vision
• Planning
• Learning
• …
Graphics

HAL

Now
Chess

HAL

Now

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

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”

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?
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 (!)
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

<table>
<thead>
<tr>
<th>Fed</th>
<th>raises</th>
<th>interest</th>
<th>rates</th>
<th>0.5 %</th>
<th>in effort to</th>
<th>control inflation</th>
</tr>
</thead>
</table>

Syntactic attachment ambiguities

<table>
<thead>
<tr>
<th>Fed</th>
<th>raises</th>
<th>interest</th>
<th>rates</th>
<th>0.5 %</th>
<th>in effort to</th>
<th>control inflation</th>
</tr>
</thead>
</table>

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)

```
S
  /\   
 NP VP
  /\   /
 NNP V NP
 |   | /
 Fed raises NN NN
    /  
   interest rates
```
Effects of V/N Ambiguity (2)

```
S
 /   |
NP   VP
 /   / |
N   N  V  N
Fed raises interest rates
```
Effects of V/N Ambiguity (3)

S

NP

N
Fed

N
raises

N
interest

N
rates

VP

V
CD

NP

N
0.5
%

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

• 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.

• Linguistics draws on research in Natural Language Processing, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.
Example Applications of NLP

Natural Language Processing
Natural Language Processing should make it possible for people to use computers in much the same way that they would use a human assistant to get their work done.

ISI's Natural Language Group
Overview of Research Environment Natural Language Processing at USC/ISI. ... USC offers a wide range of courses in areas related to natural language processing.

Foundations of Statistical Natural Language Processing
Foundations of Statistical Natural Language Processing, ... Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press.

Yahoo! Directory Artificial Intelligence > Natural Language Processing
Artificial Intelligence > Natural Language Processing Directory > Science > Computer Science > Artificial Intelligence > Natural Language Processing, ...
Example Applications of NLP:
MSWord spelling correction, grammar checking

If you use Microsoft Word, you have no doubt noticed it will correct any misspelled words (or, to be exact, all words that did you know that you can correct these errors simply by using use Microsoft Word. It will give you a list of the words that it word you want appears in the list). You simply pick it from the list.
Example Applications of NLP:

News categorization and summarization
Example Applications of NLP
Information Extraction: Find experts, employees

http://www.zoominfo.com

Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner
Example Applications of NLP: Information Extraction: Job Openings

foodscience.com-Job2

JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.html

OtherCompanyJobs: foodscience.com-Job1

If you dream of cold creamy chocolate or ooey gooey cookie dough, there's a great opportunity for you to realize your culinary dreams with this major corporation's ice cream line. This position will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or Dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.

Contact: Email 1-800-488-2611
Example Applications of NLP:

Information Extraction: Job Openings
Example Applications of NLP: Automatically Solving Crossword Puzzles
Example Applications of NLP: Question Answering

Question:

who is married to bill gates

Possible answers:  XML  TXT

- Bill was married to Melinda French Gates in 1994 in Hawaii.
- Mary Gates, Bill's mother, biggest fan, and strongest prodder, finally laid down an ultimatum in 1993. She was dying of cancer, and wanted to see her only son married.
- Bill Gates married Melinda French in Hawaii on January 1, 1994, and his mother died a few months later.
- 1994 Bill Gates and Melinda French married in Hawaii on New Years Day.
Example Applications of NLP: 
Machine Translation
Example Applications of NLP:
Automatically generate Harlequin Romance novels?
Goals of the Course

• Introduce you to Natural Language Processing problems and solutions.

• Ultimate focus on handling ambiguity by probabilistic integration of evidence.

• Give you some hands-on practice with data and a handful of methods.
This Class

• Assumes you come with some skills…
  – Some basic math/probability, decent programming skills
    (We will use Python; tutorial coming next week.)
  – 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: David Mimno
  Karl Schultz
- Assistants: Hanna Wallach
  Khash Rohanimanesh

- Time: Tue/Thu 2:30-3:45pm
- Mailing list: 585-staff@cs.umass.edu
- More information on Web site:
  http://www.cs.umass.edu/~mccallum/courses/inlp2007
Grading

• 7 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 backgrounds
  – chance to explore a special interest at end of term

• Midterm & Final, and classroom participation
For Linguistics Students: 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 parts:**
  - 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.
To Do This Week

• Visit course Web site, browse around.

• Read Chapters 1 and 2 in Jurafsky & Martin textbook
  – Available on line! See course web site.

• Install Python on your computer
  – Get extensive help from the TAs if you like!
Thank you!