CSCI 5832
Natural Language Processing

Lecture 1
Jim Martin

Today 1/15

- An exercise
- Overview of the field of NLP
- Administrivia
- Course topics
- Commercial relevance

What's this story about?

17 the
13 and
10 of
10 a
8 to
7 s
6 in
6 Romney
6 Mr
5 that
5 state
5 for
4 industry
4 automotive
4 Michigan
3 on
3 his
3 have
3 are
2 would
2 with
2 up
2 think
2 technology
2 speech
2 primary
2 neck
2 is
2 further
2 fuel
2 from
2 former
2 energy
2 campaigning
2 billion
2 bill
2 at
2 They
2 Senator
2 Republican
2 Monday
2 McCain
2 He
2 Gov
1 wrong
1 who
1 upon
1 unions
1 unfunded
1 ultimately
1 trade
1 top
1 took
1 together
1 throughout
1 they
1 there
1 task
1 t
1 support
1 successive
1 standards
1 some
1 signed
1 shake
1 set
1 science
1 said
1 rise
1 research
1 requires
1 representatives
1 remarkably
1 recent
1 rebuild
1 raising
1 pushed
1 presidential
1 polls
1 policy
1 plight
1 pledged
1 plan
1 people
1 or
1 off
1 measure
1 materials
1 mandates
1 losses
1 litany
1 leading
1 leadership
1 lawmakers
1 killer
1 jobs
1 job
1 its
1 issues
1 indicated
1 independent
1 increase
1 including
1 imposing
1 him
1 heavily
1 has
1 greenhouse
1 gone
1 gas
1 future
1 forever
1 focused
1 flurry
1 fluid
1 first
1 final
1 field
1 federal
1 essentially
1 emphasizing
1 emissions
1 efficiency
1 economic
1 don
1 domestic
1 do
1 disinterested
1 die
1 development
1 delivered
1 days
1 criticized
1 could
1 costs
1 contest
1 come
1 childhood
1 cause
1 cap
1 candidates
1 by
1 bring
1 between
1 being
1 been
1 be
1 back
1 automobile
1 automakers
1 asserted
1 aiding
1 ahead
1 agenda
1 again
1 after
1 advisers
1 acknowledged
1 With
1 Washington
1 There
1 Recent
1 President
1 New
1 Mitt
1 Mike
1 Massachusetts
1 Lieberman
1 Joseph
1 John
1 Iowa
1 In
1 I
1 Huckabee
1 Hampshire
1 Economic
1 Detroit
1 Connecticut
1 Congress
1 Club
1 Bush
1 Arkansas
1 Arizona
1 America

The story

Romney Battles McCain for Michigan Lead
By MICHAEL LUO

DETROIT — With economic issues at the top of the agenda, the leading Republican presidential candidates set off Monday on a final flurry of campaigning in Michigan ahead of the state's primary that could again shake up a remarkably fluid Republican field.

Recent polls have indicated the contest is neck-and-neck between former Gov. Mitt Romney of Massachusetts and Senator John McCain of Arizona, with former Gov. Mike Huckabee of Arkansas further back.

Mr. Romney's advisers have acknowledged that the state's primary is essentially do-or-die for him after successive losses in Iowa and New Hampshire. He has been campaigning heavily throughout the state, emphasizing his childhood in Michigan and delivered a policy speech on Monday focused on aiding the automotive industry.

In his speech at the Detroit Economic Club, Mr. Romney took Washington lawmakers to task for being a "disinterested" in Michigan's plight and imposing upon the state's automakers a litany of "unfunded mandates," including a recent measure signed by President Bush that requires the raising of fuel efficiency standards.

He criticized Mr. McCain and Senator Joseph I. Lieberman, independent of Connecticut, for a bill that they have pushed to cap and trade greenhouse gas emissions. Mr. Romney asserted that the bill would cause energy costs to rise and would ultimately be a "job killer."

Mr. Romney further pledged to bring together in his first 100 days representatives from the automotive industry, unions, Congress and the state of Michigan to come up with a plan to "rebuild America's automotive leadership" and to increase to $20 billion, from $4 billion, the federal support for research and development in energy, fuel technology, materials science and automotive technology.

Vector Representations

- The first slide was a basic vector representation for the meaning of a text
  - Also known as a "bag of words" representation
- Discourse segments, sentence boundaries, syntax, word order are all ignored.
- Roughly, all that matters is the set of words that occur and how often they occur

Vector Representations

- These representations are the basis for many interesting and useful systems
- BUT there has to be something better.
- Much of NLP is directed at finding representations that do a better job at capturing the meaning and intent behind texts.

Natural Language Processing

- What is it?
  - We're going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
  - We will be secondarily concerned with the insights that such computational work gives us into human processing of language.

Why Should You Care?

Two trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication

Major Topics

1. Words
2. Syntax
3. Meaning
4. Discourse

} 5. Applications exploiting each

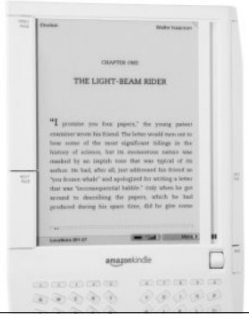
Applications

- First, what makes an application a *language processing application* (as opposed to any other piece of software)?
  - An application that requires the use of knowledge about human languages
    - Example: Is Unix wc (word count) an example of a language processing application?

Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities

What's missing



Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the following interaction with HAL the computer from 2001: A Space Odyssey

HAL from 2001

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

What's needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
- How groups of words clump
  - What the clumps mean

What's needed?

- Dialog
  - It is polite to respond, even if you're planning to kill someone.
  - It is polite to pretend to want to be cooperative (I'm afraid, I can't…)

Real Example

What is the Fed's current position on interest rates?

- What or who is the "Fed"?
- What does it mean for it to to have a position?
- How does "current" modify that?

Caveat

NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with perfect solutions/algorithms
- We can't let either of those facts get in our way

Administrative Stuff

- Waitlist/SAVE
  - Course is open
- Web page
  - www.cs.colorado.edu/~martin/csci5832.html
- Reasonable preparation
- Requirements

CAETE

- This venue tends to encourage students to act like they are viewing the taping of a TV show.
- You're not, you're part of the show.
- You must participate.

Web Page

The course web page can be found at.
www.cs.colorado.edu/~martin/csci5832.html.

It will have the syllabus, lecture notes, assignments, announcements, etc.

You should check it periodically for new stuff.

Mailing List

- There is a automatically generated mailing list.
- Mail goes to your official CU email address.
  - I can't alter it so don't ask me to send your mail to gmail/yahoo/work or whatever
  - You can set up a forward yourself
  - But you can only send to the list from your CU account

Preparation

- Basic algorithm and data structure analysis
- Ability to program
- Some exposure to logic
- Exposure to basic concepts in probability
- Familiarity with linguistics, psychology, and philosophy
- Ability to write well in English

Requirements

- Readings:
  - Speech and Language Processing by Jurafsky and Martin, Prentice-Hall 2008
    - Draft version of the 2nd Ed.
  - Various conference and journal papers
- Around 4 or 5 assignments
- 3 quizzes
- Final comprehensive exam on Monday May 5 from 1:30 to 4:00.

Programming

- All the programming will be done in Python.
  - It's free and works on Windows, Macs, and Linux
  - It's easy to install
  - Easy to learn

Programming

- Go to www.python.org to get started.
- The default installation comes with an editor called IDLE. It's a serviceable development environment.
- Python mode in emacs is pretty good. It's what I use but I'm a dinosaur.
- If you like eclipse, there is a python plug-in for it.

Grading

- Assignments – 30%
- Quizzes – 30%
- Final Exam – 30%
- Participation – 10%

Course Material

- We'll be intermingling discussions of:
  - Linguistic topics
    - E.g. Morphology, syntax, discourse structure
  - Formal systems
    - E.g. Regular languages, context-free grammars
  - Applications
    - E.g. Machine translation, information extraction

Linguistics Topics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse processing

My biases…

- I'm not terribly into phonology or speech
- I care about meaning in general, and word meanings in particular

Topics: Techniques

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Unification
  - Lambda calculus
- First order logic
- Probability models
- Supervised machine learning methods

Topics: Applications

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation
- Stand-alone
- Enabling applications
- Funding/Business plans

Next Time

- Read Chapter 1
- Download, install and learn Python. The first assignment will be given out next time.