

Natural Language Processing

CS181  
David Kauchak

## + Who are you and why are you here?

- Name/nickname
- Dept., college and year
- Why are you taking this course?
- What topics would you like to see covered?

## + Administrivia

- <http://www.cs.pomona.edu/classes/cs159/>
  - Office hours, schedule, assigned readings, assignments
  - Everything will be posted there
- Read the "administrivia" handout!
  - ~5 assignments (in a variety of languages)
  - 4 quizzes (dates are tentative)
  - in-class presentation
  - final project for the last month
    - teams of 2-3 people
    - research-like with write-up and presentation
  - class participation
  - readings
- Academic honesty and collaboration

## + Administrivia

- First assignment posted already
  - Shouldn't take too long
  - Due Monday at the beginning of class
- CS colloquium tomorrow
  - Text simplification
  - 4:15pm Rose Hill Theatre
- CS accounts

## + What to expect...

- This course will be challenging for many of you
  - assignments will be non-trivial
  - content can be challenging
- But it is a fun field!
- We'll cover
  - basic linguistics
  - probability
  - the common problems
  - many techniques and algorithms
  - common learning techniques
  - applications

## + Requirements and goals

- Requirements
  - Competent programmer
    - Mostly in Java, but I may allow/encourage other languages
  - Comfortable with mathematical thinking
    - We'll use a fair amount of probability, which I will review
    - Other basic concepts, like logs, summation, etc.
  - Data structures
    - trees, hashtables, etc.
- Goals
  - Learn the problems and techniques of NLP
  - Build real NLP tools
  - Understand what the current research problems are in the field

## + What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia

## + What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language...


- The book

+ **Key: Natural text**


**“A growing number of businesses are making Facebook an indispensable part of hanging out with their shingles. Small businesses are using ...”**

■ Natural text is written by people, generally for people

**Why do we even care about natural text in computer science?**



+ **Why do we need computers for dealing with natural text?**



**We knew the web was big...**


7/25/2008 10:12:00 AM

We've known it for a long time: the web is big. The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, we've seen a lot of big numbers about how much content is really out there. Recently, **even our search engineers stopped in awe about just how big the web is these days — when our systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once!**

How do we find all those pages? We start at a set of well-connected initial pages and follow each of their links to new pages. Then we follow the links on those new pages to even more pages and so on, until we have a huge list of links. In fact, we found even more than 1 trillion individual links, but not all of them lead to unique web pages. Many pages have multiple URLs with exactly the same content or URLs that are auto-generated copies of each other. **Even after removing those exact duplicates, we saw a trillion unique URLs, and the number of individual web pages out there is growing by several billion pages per day.**


+ **Web is just the start...**

e-mail



**247 billion e-mails a day**

corporate databases



twitter

**27 million tweets a day**

Blogs: **126 million** different blogs

http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/

+ **Why is NLP hard?**

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors

## + Why is NLP hard?

- **User:** Where is The Green Hornet playing in the Claremont Area?
- **System:** The Green Hornet is playing at the Ontario Mills theatre.
- **User:** When is **it** playing **there**?
- **System:** It's playing at 2pm, 5pm and 8pm
- **User:** I'd like 1 adult and 2 children for **the first show**. How much would **that** cost?

## + Why is NLP hard?

- Natural language:
  - is highly ambiguous at many different levels
  - is complex and contains subtle use of context to convey meaning
  - is probabilistic?
  - involves reasoning about the world
  - is highly social
  - is a key part in how people interact
- However, some NLP problems can be surprisingly easy

## + Different levels of NLP

pragmatics/discourse: how does the context affect the interpretation?

semantics: what does it mean?

syntax: phrases, how do words interact

words: morphology, classes of words

## + NLP problems and applications

What are some places where you have seen NLP used?

What are NLP problems?

## + NLP problems and applications

- Lots of problems of varying difficulty

- Easier

- Word segmentation: where are the words?

I would've like Dr. Kauchak to finish early. But he didn't.

北海 已 成为 中国 对 外 开 放 中 升 起 的 一 颗 明 星  
 Beihai already become China to outside open during rising (DE) one measure bright

- Speech segmentation
  - Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  - Language identification

Soy un profesor con queso.

## + NLP problems and applications

- Easier continued

- truecasing

i would've like dr. kauchak to finish early. but he didn't.

- spell checking

Identifying misspellings is challenging especially in the dessert.

- OCR

4

## + NLP problems and applications

- Moderately difficult

- morphological analysis/stemming

smarter  
 smarter  
 smartly → smart  
 smartest  
 smart

- speech recognition



- text classification



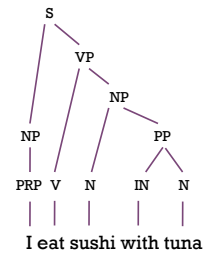
## + NLP problems and applications

- moderately difficult continued

- text segmentation: break up the text by topics

- part of speech tagging (and inducing word classes)

- parsing



## + NLP problems and applications

- moderately difficult continued
- word sense disambiguation

As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

- grammar correction

We am good at grammar.

- speech synthesis

## + NLP problems and applications

- Hard (many of these contain many smaller problems)
- Machine translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

## + NLP problems and applications

- Information extraction

IBM hired Fred Smith as president.

person	company	position
Fred Smith	IBM	president

## + NLP problems and applications

- Summarization

A company that acts as a middle man between content companies and Internet service providers is accusing Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior. (article 8) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 6) At issue is the cable operator's decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 9) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials overseeing the government's review of the proposed merger. (article 5) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the term of their merger agreement to March of next year. (article 4) Democrat Michael Copps fears the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 3) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Atwell Baker (article 1)

## + NLP problems and applications

- Natural language understanding
  - Text => semantic representation (e.g. logic, probabilistic relationships)
- Information retrieval and question answering
  - "How many programmers in the child care department make over \$50,000?"
  - "Who was the fourteenth president?"
  - "How did he die?"

## + NLP problems and applications

- Text simplification

**Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.**



**Alfonso Perez is a former Spanish football player.**

## + Where are we now?

- Many of the "easy" and "medium" problems have reasonable solutions
  - spell checkers
  - sentence splitters
  - word segmenters/tokenizers

## + Where are we now?

- Parsing
  - Stanford Parser (<http://nlp.stanford.edu:8080/parser/>)

```

Stanford Parser
Please enter a sentence to be parsed.
My dog also likes eating bananas.

Language: [English] Sample Sentence [Parse]

Your query:
My dog also likes eating bananas.

Tagging
My/P/PRP dog/NN also/CC like/VB eat/ing/VBG banan/NN /./

Parse
(S(00)
  (S (NP (PP (IN) in) (NP (NN) dog))
    (VP (VBZ) is) (NP (NN) also)
      (VP (VBZ) like)
        (VP (VBZ) eat)
          (NP (NN) banana))))))
  (P .)
)

```

## + Where are we now?

- Machine translation
  - Getting better every year
  - enough to get the jist of most content, but still no where near a human translation
  - better for some types of text
  
- translate.google.com
  
- Many commercial versions...
  - systran
  - language weaver

## + Where are we now?

- Information retrieval/query answering
  - search engines:
    - pretty good for some things

who was the fifteenth president of the united states

About 928,000 results (0.17 seconds) [Advanced search](#)

[James Buchanan - Fast Facts - Fifteenth President James Buchanan](#)

James Buchanan, Fifteenth President of the United States. Credit: Library of Congress, Prints and Photographs Division. LC-84-10211-65283-D.C. — americanhistory.about.com/od/.../alf\_j\_buchanan.htm - Cached - Similar

- does mostly pattern matching and ranking
  - no deep understanding
  - still requires user to "find" the answer

## + Where are we now?

- Question answering
  - wolfram alpha

computational knowledge engine

who is the fifteenth president of the united states

Input interpretation:  
United States President 15<sup>th</sup>

Result:  
James Buchanan

## + Where are we now?

- Question answering
  - wolfram alpha

computational knowledge engine

what is the most popular car color in the united states?

↳ Using closest Wolfram|Alpha interpretation: united states

Input interpretation: Mathematica form  
United States



