

+

**A NEW CAPTCHA APPROACH**

< <PREV RANDOM NEXT > >

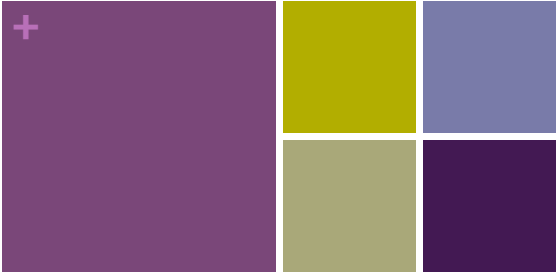
TO COMPLETE YOUR WEB REGISTRATION, PLEASE PROVE THAT YOU'RE HUMAN:

WHEN LITTLEFOOT'S MOTHER DIED IN THE ORIGINAL 'LAND BEFORE TIME,' DID YOU FEEL SAD?

YES  
 NO

(BOTS: NO LYING)

+



**Natural Language Processing**

CS457  
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?

+

**AI**

- How many of you were originally signed up for AI (or planned on signing up for AI)?
- Any particular topics you wanted to see covered?

## + Administrivia

- <http://www.cs.middlebury.edu/~dkauchak/classes/cs457/>
  - 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)
  - final project for the last 3-4 weeks
    - teams of 2-3 people
  - class participation
  - readings
- Honor code and collaboration

## + Administrivia

- First assignment posted already
  - Shouldn't take too long
  - Due Thursday at the beginning of class
- Lab access
- 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 machine learning techniques
  - AI/search
  - 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

**ALL NATURAL**

“A growing number of businesses are making Facebook an indispensable part of hanging out 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?

The Official Google Blog | Insights from Googlers into our products, technology, and the Google culture.

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



twitter

27 million tweets a day

corporate  
databases



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 Contagion playing in the Middlebury Area?
- **System:** Contagion is playing at the Marquis 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 liked Prof. Kauchak to finish early. But he didn't.*

## + NLP problems and applications

- Lots of problems of varying difficulty

- Easier

- Word segmentation: where are the words?

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

## + NLP problems and applications

- Lots of problems of varying difficulty
- Easier
  - Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  - I would've liked Prof. Kauchak to finish early. But he didn't.*
- Language identification
  - Soy un maestro con queso.*

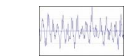
## + NLP problems and applications

- Easier continued
  - truecasing
    - i would've liked prof. 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  
smartest  
smart* → *smart*
  - speech recognition
  - text classification



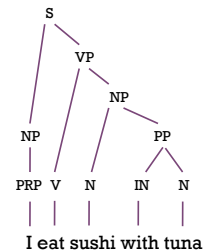
SPAM



sentiment analysis

## + 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 likes/VB likes/VBD eating/VBG bananas/NN /.

Parse
(SROOT
 (S
  (NP (PPRS My) (IN dog))
  (VP (VB likes))
  (VP (VBD eating))
  (P (DOT))
  (ADJP (NN bananas))))))
  
```

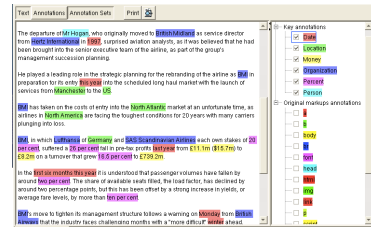


## + Where are we now?

- 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 extraction
  - Structured documents (very good!)
    - [www.dealtime.com](http://www.dealtime.com)
    - [www.froogle.com](http://www.froogle.com)
  - AKT technologies
- Lots of these
  - FlipDog
  - WhizBang! Labs
  - ...
  - work fairly well



## + Where are we now?

- CMU's NELL (Never Ending Language Learner)
- <http://rtw.ml.cmu.edu/rtw/>

Recently-Learned Facts [bunker](#)

instance

red\_harvester\_ant is an invertebrate

crookpot\_mushroom\_chicken is a type of meat

football is a hobby

colgate\_palmolive is a magazine

kirkwood is a city

ben\_takahashi plays the sport baseball

andy\_warhol is a visual artist in the field of printmaking

john\_hayward held the position of vice\_admiral

tom\_osborne works for nebraska

lg is a company headquartered in the city nashville

## + Where are we now?

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

who was the fifteenth president of the united states  Search

About 628,000 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-DAHQ-151-4629-33.C...

americanhistory.about.com/od/.../Jeff\_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

WolframAlpha<sup>™</sup> 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

WolframAlpha<sup>™</sup> computational knowledge engine

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

Using closest Wolfram|Alpha interpretation: united states

Input Interpretation: United States *Mathematica form*

### + Question answering

#### The Science Behind an Answer

Watson performs so fast that it can rival the greatest human contestants in understanding a Jeopardy! clue and arriving at a single, precise answer. The significance of this accomplishment can be difficult to comprehend.

Watch the video to see how the computing system designed to play Jeopardy! works.

The first person mentioned by name in "The Man in the Iron Mask" is this hero of a previous book by the same author.

Possible Answers

- balance
- ban
- bang
- bank
- bathe
- battle
- be
- beam
- beer
- beat
- become
- beg

### + Where are we now?

- Question answering
  - Many others...
    - TREC question answering competition
    - language computer corp
    - answerbus
    - ...

## + Where are we now?

### ■ Summarization

- NewsBlaster (Columbia)
  - <http://newsblaster.cs.columbia.edu/>

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. [\(article 5\)](#) 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 Attwell Baker [\(article 1\)](#)

## + Where are we now?

### ■ Voice recognition

- pretty good, particularly with speaker training
  - Apple OS has one built in:
    - "What time is it?"
    - "Switch to finder"
    - "Hide this application"
  - IBM ViaVoice
  - Dragon Naturally Speaking

### ■ Speech generation

- The systems can generate the words, but getting the subtle nuances right is still tricky
  - Apple OS
  - [translate.google.com](http://translate.google.com)

## + Other problems

- Many problems untackled/undiscovered
- "That's What She Said: Double Entendre Identification"
  - ACL 2011
  - <http://www.cs.washington.edu/homes/brun/pubs/pubs/Kiddon11.pdf>