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

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

“...A growing number of businesses are making Facebook an indispensible 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?
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

Why is NLP hard?

- Natural language:
  - highly ambiguous at many different levels
  - complex and contains subtle use of context to convey meaning
  - probabilistic?
  - involves reasoning about the world
  - 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?

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
  - 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
  - Spell checking
    Identifying mispellings is challenging especially in the dessert.
  - OCR

NLP problems and applications

- Moderately difficult
  - Morphological analysis/stemming
    smarter smarter smart smart smartest
  - Speech recognition
  - Text classification
    SVM

NLP problems and applications

- Moderately difficult continued
  - Text segmentation: break up the text by topics
  - Part of speech tagging (and inducing word classes)
  - Parsing

I eat sushi with tuna
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.

We are good at grammar.

speech synthesis

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.

IBM hired Fred Smith as president.

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 are good at grammar.
  - speech synthesis

NLP problems and applications

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

<table>
<thead>
<tr>
<th>person</th>
<th>company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Smith</td>
<td>IBM</td>
<td>president</td>
</tr>
</tbody>
</table>
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?"

Text simplification

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

Where are we now?

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

Parsing

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

Alfonso Perez is a former Spanish football player.
Where are we now?

- Machine translation
  - Getting better every year
  - Enough to get the jist of most content, but still nowhere 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
    - www.froogle.com
    - AIT 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/

Where are we now?

- Information retrieval/query answering
  - Search engines:
    - Pretty good for some things
    - Does mostly pattern matching and ranking
    - No deep understanding
    - Still requires user to “find” the answer
Where are we now?

- Question answering
  - Wolfram Alpha

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the 44th president of the United States?</td>
<td>Barack Obama</td>
</tr>
<tr>
<td>What is the most popular car sales in the United States?</td>
<td>United States</td>
</tr>
</tbody>
</table>

Many others...

- TREC question answering competition
- Language computer corp
- AnswerBus

...
Where are we now?

- **Summarization**
  - NewsBlaster (Columbia)

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

Other problems

- Many problems untackled/undiscovered
- “That’s What She Said: Double Entendre Identification”
  - ACL 2011