Course Information

http://www.cs.berkeley.edu/~klein/cs288/fa14/

https://piazza.com/berkeley/fall2014/cs288/

Course Requirements

- **Prerequisites:**
  - CS 188 (CS 281a) and preferably CS170 (A-level mastery)
  - Strong skills in Java or equivalent
  - Deep interest in language
  - Successful completion of the first project
  - There will be a lot of math and programming

- **Work and Grading:**
  - Six assignments (individual, jars + write-ups)
  - This course is a major time-commitment!

- **Books:**
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
  - Also: Manning and Schatz, Foundations of Statistical NLP

Other Announcements

- **Course Contacts:**
  - Webpage: materials and announcements
  - Piazza: discussion forum

- **Enrollment:** We'll try to take everyone who meets the requirements

- **Computing Resources**
  - You will want more compute power than the instructional labs
  - Experiments can take up to hours, even with efficient code
  - Recommendation: start assignments early

- **Questions?**

Language Technologies

**Goal: Deep Understanding**
- Requires context, linguistic structure, meanings..

**Reality: Shallow Matching**
- Requires robustness and scale
- Amazing successes, but fundamental limitations

AI: Where Do We Stand?

Source: Slate Pioneers
Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV

Text to Speech (TTS)
- Text in, audio out
- SOTA: totally intelligible (if sometimes unnatural)

Example: Siri

- Siri contains
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

Text Data is Superficial

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

... But Language is Complex

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

Deeper Linguistic Analysis

Hurricane Emily howled toward Mexico’s Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters.

Learning Hidden Syntax

<table>
<thead>
<tr>
<th>Personal Pronouns (PRP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP-1</td>
</tr>
<tr>
<td>PRP-2</td>
</tr>
<tr>
<td>PRP-3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proper Nouns (NNP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP-12</td>
</tr>
<tr>
<td>NNP-2</td>
</tr>
<tr>
<td>NNP-1</td>
</tr>
<tr>
<td>NNP-15</td>
</tr>
<tr>
<td>NNP-3</td>
</tr>
</tbody>
</table>
Search, Facts, and Questions

Example: Watson

Language Comprehension?

Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation

Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)
Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.

That would be an interim solution which would make it possible to work towards a binding charter in the long term.

It would be a transitional solution which would lead to a binding charter.

Chairs are solid?

Deeper Understanding: Reference

Q: Who signed the Serve America Act?
A: Barack Obama

More Data: Machine Translation

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x DATA</td>
<td>[this [would] [be] [a] [transitional] [solution] [which] [would] [eventually] [lead] to] [a] [binding] [charter] [ ]</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[this [would] [be] [a] [transitional] [solution] [which] [would] [lead] to] [a] [binding] [charter] [ ]</td>
</tr>
</tbody>
</table>

100x DATA

1000x DATA

Data and Knowledge

- Classic knowledge representation worry: How will a machine ever know that...
  - Ice is frozen water?
  - Beige looks like this:
  - Chairs are solid?

- Answers:
  - 1980: write it all down
  - 2000: get by without it
  - 2020: learn it from data

Names vs. Entities

Deeper Understanding: Reference

Q: Who signed the Serve America Act?
A: Barack Obama

Names vs. Entities

Presidential photo (Barack Obama)
Example Errors

Input:
America Online announced on Monday that the company plans to update its instant messaging service.

Correct:
America Online is the company's instant messaging service.

Guess:
America Online the company is its instant messaging service.

Discovering Knowledge

Grounded Language

Grounding with Natural Data

... on the beige loveseat.

What is Nearby NLP?

- **Computational Linguistics**
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

- **Cognitive Science**
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!

- **Speech Processing**
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Example: NLP Meets CL

- Example: Language change, reconstructing ancient forms, phylogenies
  - ... just one example of the kinds of linguistic models we can build
What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
  - We’ll focus on what makes the problems hard, and what works in practice...

Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of cs281a)
    - Basic linguistics background (ling100)
    - Strong coding skills (Java), well beyond cs61b
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps

- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

  - This semester: new projects (speech, translation, analysis)

Some BIG Disclaimers

- The purpose of this class is to train NLP researchers
  - Some people will put in a LOT of time – this course is more work than most classes (grad or undergrad)
  - There will be a LOT of reading, some required, some not – you will have to be strategic about what reading enables your goals
  - There will be a LOT of coding and running systems on substantial amounts of real data
  - There will be a LOT of machine learning / math
  - There will be discussion and questions in class that will push past what I present in lecture, and I’ll answer them
  - Not everything will be spelled out for you in the projects
  - Especially this term: new projects will have hiccups

  - Don’t say I didn’t warn you!

Some Early NLP History

- 1950’s
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
  - Toy models: MT using basically word-substitution
  - Optimism!

- 1960’s and 1970’s: NLP Winter
  - Bar-Hillel (PAKCO) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - ... but toy domains / grammars (SHRDLU, LUNAR)

- 1980’s and 1990’s: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - Evaluate everything

- 2000’s: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - Begin to get both breadth and depth

Problem: Structure

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor’s Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Why are these funny?

Problem: Scale

- People did know that language was ambiguous!
  - ...but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - ...they didn’t realize how bad it would be
Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire

<table>
<thead>
<tr>
<th>Number of Words</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20000</td>
<td>0.1</td>
</tr>
<tr>
<td>40000</td>
<td>0.4</td>
</tr>
<tr>
<td>60000</td>
<td>0.8</td>
</tr>
<tr>
<td>80000</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Outline of Topics

- Words and Sequences
  - Speech recognition
  - N-gram models
  - Working with a lot of data

- Structured Classification

- Trees
  - Syntax and semantics
  - Syntactic MT
  - Question answering

- Machine Translation

- Other Topics
  - Reference resolution
  - Summarization
  - Diachronics
  - ... 

A Puzzle

- You have already seen N words of text, containing a bunch of different word types (some once, some twice...)
- What is the chance that the N+1\textsuperscript{th} word is a new one?