Course Details

- **Books:**
  - Jurafsky and Martin. Speech and Language Processing, 2 Ed
  - Manning and Schütze. Foundations of Statistical NLP

- **Prerequisites:**
  - CS 188 or CS 281 (grade of A or see me)
  - Strong in Java or equivalent
  - Deep interest in language
  - There will be a lot of statistics and programming

- **Work and Grading:**
  - Four coding assignments
  - Solo, turn in write-ups only
  - Final group project
  - Participation
  - Units

Announcements

- **Computing Resources:**
  - You will want more compute power than the instructional labs
  - Recommendation: start assignments early to find out whether what you have works

- **Communication:**
  - Announcements: webpage
  - Public discussion: newsgroup
  - My email: klein@cs.berkeley.edu

- **Enrollment:**
  - Undergrads stay after and see me

- **Questions?**

The Dream

- **It’d be great if machines could**
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us

- **But they can’t:**
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge

- **So:**

What is NLP?

- **Fundamental goal:** deep understand of broad language
  - Not just string processing or keyword matching!

- **End systems that we want to build:**
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering...
  - Modest: spelling correction, text categorization...
### Speech Systems
- **Automatic Speech Recognition (ASR)**
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV

### Information Extraction
- **Unstructured text to database entries**

<table>
<thead>
<tr>
<th>Person</th>
<th>Company</th>
<th>Post</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times</td>
<td>president and general manager</td>
<td>NY</td>
</tr>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times</td>
<td>executive vice president</td>
<td>NY</td>
</tr>
<tr>
<td>Lance R. Primis</td>
<td>New York Times</td>
<td>president and CEO</td>
<td>NY</td>
</tr>
</tbody>
</table>

- SOTA: perhaps 70% accuracy for multi-sentence templates, 90%+ for single easy fields

### Question Answering
- **Question Answering:**
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: “What is the capital of Wyoming?”
  - Can be harder: “How many US states’ capitals are also their largest cities?”
  - Can be open ended: “What are the main points of the report?”
  - Can vary greatly in difficulty and complexity of question and answering database
  - SOTA: Can do factoids, even when text isn’t a perfect match

### Machine Translation
- **Atlanta, prese il killer del palazzo di Giustizia**
  - ATLANTA: La grande paura di per 26 ore ha attirato attentare a un gruppo di persone nel palazzo di Giustizia e ilLA.

- **Atlanta, taken the killer of the palace of Justice**
  - ATLANTA - The great fear that for 26 hours has gripped students of school: man who killed three persons in palace of justice and that a customs agent has then killed, is disclosed to the police, after he has tried to escape in the lobby of one woman in a complex of apartment to the veranda of the city. For all the day, the center of the city, center of the City Square and end of Cochci 29th, heart of one populous metropolitan area, was remained paralyzed.

### Summarization
- **Condensing documents**
  - Single or multiple
  - Extractive or synthetic
  - Aggregative or representative
  - Even just shortening sentences
  - Very context-dependent
  - An example of analysis with generation

### Etc: Historical Change
- *Change in form over time, reconstruct ancient forms, phylogenies*
- *... just an example of the many other kinds of models we can build*
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?**
  - 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

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
  - **Engineering Methods**
    - Issues of scale
    - Sometimes, very ugly (but important) hacks
  - 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 cs281)
    - First speech systems
    - Machine translation (MT) hugely funded by military
    - Decent 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 the real tools used in NLP (language models, taggers, parsers, translation systems)
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

Some Early NLP History

- **1950s:**
  - 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 (FAHQT) 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+: Richer Statistical Methods**
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up

Some BIG Disclaimers

- The purpose of this class is to train NLP researchers
  - Some people will put in a LOT of time
  - There will be a LOT of reading, some required, some not – doing it all would be time-consuming
  - There will be a LOT of coding and running systems on substantial amounts of real data
  - There will be a LOT of statistical modeling (though we do use a few basic techniques very heavily)
  - There will be discussion and questions in class that will push past what I’ve presented in lecture, and I’ll answer them
  - Not everything will be spelled out for you in the projects

- Don’t say I didn’t warn you!

Problem: Ambiguities

- **Headlines:**
  - Iraqi Head Seeks Arms
  - Ban on Nude Dancing on Governor’s Desk
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
  - Hospitals Are Sued by 7 Foot Doctors

- Why are these funny?
Syntactic Analysis

- SOTA: 80-90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

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.

Semantic Ambiguity

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don’t fully nail down the meaning

Every morning someone’s alarm clock wakes me up

John’s boss said he was doing better

In general, every level of linguistic structure comes with its own ambiguities…

Dark Ambiguities

- Dark ambiguities: most analyses are shockingly bad (meaning, they don’t have an interpretation you can get your mind around)

This analysis corresponds to the correct parse of

“This will panic buyers!”

- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

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

Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora

- Examples
  - Newsavire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

Problem: Sparsity

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

Problem: Sparsity

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

- Words
  - N-gram models and smoothing
  - Classification and clustering
- Sequences
  - Part-of-speech tagging
  - Information extraction
  - Speech recognition / synthesis
- Trees
  - Syntax and semantics
  - Machine translation
  - Question answering
- Discourse
  - Reference resolution
  - Dialog systems

Next Week

- We’re going to do an experiment in competitive parsing
- Polls:
  - How many have an EECS research account?
  - How many have a laptop they can bring next week?
  - How many know what a prepositional phrase is?
- Also: Assignment 1 will be out very soon, and assignments will move fast