Course Description

This course is designed to introduce students to the fundamental concepts and ideas in natural language processing (NLP), and to get them up to speed with current research in the area. It develops an in-depth understanding of both the algorithms available for the processing of linguistic information and the underlying computational properties of natural languages. Word-level, syntactic, and semantic processing from both a linguistic and an algorithmic perspective are considered. The focus is on modern quantitative techniques in NLP: using large corpora, statistical models for acquisition, disambiguation, and parsing. Also, it examines and constructs representative systems.

Prerequisites

- Adequate experience with programming and formal structures (e.g., CS106B/X and CS103B/X).
- Programming projects will be written in Java 1.5, so knowledge of Java (or a willingness to learn on your own) is required.
- Knowledge of standard concepts in artificial intelligence and/or computational linguistics (e.g., CS121/221 or Ling 180).
- Basic familiarity with logic, vector spaces, and probability.

Intended Audience

Graduate students and advanced undergraduates specializing in computer science, linguistics, or symbolic systems.

Textbook and Readings

This year, the required text will be:

- Daniel Jurafsky and James H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall. The book won't be in time for the class. (June 2008 update: it's now available for purchase!) We will use a reader containing parts of the second edition. The reader is available for ordering at University Readers. You order it online and they ship it to you. The cost is $40.58. [Detailed purchasing instructions.] Once you've ordered it, you can have access to the first couple of chapters that we'll use online for free. If you have any difficulties, please e-mail orders@universityreaders.com or call 800.200.3908, and email the class email list. It's referred to as J&M in the syllabus. [Book website]

Of course, I'm also fond of:

- Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press. Buy it at the Stanford Bookstore (recommended class text) or Amazon ($64 new). Please see http://nlp.stanford.edu/fsnlp/ for supplementary information about the text, including errata, and pointers to online resources.

Other useful reference texts for NLP are:

- Gerald Gazdar and Chris Mellish. 1989. Natural Language Processing in X. Addison-Wesley. (Where X = Prolog, Lisp, or, I think, Snobol.)

Other papers with relevant material will occasionally be posted or distributed for appropriate class lectures.

Copies of in-class hand-outs, such as readings and programming assignments, will be posted on the syllabus, and hard copies will also be available outside Gates 158 (in front of Prof. Manning’s office) while supplies last.

**Assignments and Grading**

There will be three substantial programming assignments, each exploring a core NLP task. They are a chance to see real, close to state-of-the-art tools and techniques in action, and where students learn a lot of the material of the class.

There will be a final programming project on a topic of your own choosing.

Finally, there will be simple weekly online quizzes, which will aim to check that you are thinking about what you hear/read.

Course grades will be based 60% on programming assignments (20% each), 8% on the quizzes, and 32% on the final project.

Be sure to read the policies on late days and collaboration.
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topic</th>
<th>Assigned Reading</th>
<th>Other Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1</td>
<td>4/2/08</td>
<td>Introduction Overview of NLP. Statistical machine translation. Language models and their role in speech processing. Course introduction and administration.</td>
<td>M&amp;S 1.0-1.3, 4.1-4.2, Collaboration Policy. Good background reading: M&amp;S 1.0-1.3, 4.1-4.2, Collaboration Policy. Optional reading on Unix text manipulation (useful skill!): Ken Church's tutorial Unix for Poets. (If your knowledge of probability theory is limited, also read M&amp;S 2.0-2.1.7. If that's too condensed, read the probability chapter of an intro statistics textbook, e.g. Rice, Mathematical Statistics and Data Analysis, ch. 1.)</td>
<td>Distributed today: Programming Assignment 1.</td>
</tr>
<tr>
<td>Lecture 3</td>
<td>4/9/08</td>
<td>Statistical Machine Translation (MT), Alignment Models</td>
<td>Assigned reading: J&amp;M ch. 25, sections 25.0-25.5, 25.11</td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>4/11/08</td>
<td>Smoothing</td>
<td></td>
<td>Smoothings: absolute discounting, proving you have a proper probability distribution, Good-Turing implementation. Information theory examples and intuitions. Java implementation issues.</td>
</tr>
<tr>
<td>Section 2</td>
<td>Fri</td>
<td>The EM algorithm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lecture 6  
Mon 4/21/08  
**Information Extraction (IE) and Named Entity Recognition (NER).**  
**Assigned reading:**  
J&M secs 22.0-22.1 (intro to IE and NER).  
J&M secs. 5.5 and 5.7 (introduce HMMs, Viterbi algorithm, and experimental technique). If you’re not familiar with supervised classification and Naïve Bayes, read J&M sec 20.2 before the parts of ch. 5.  
**Alternative reading:** M&S 8.1 (evaluation), 7.1 (experimental methodology), 7.2.1 (Naïve Bayes), 10.2-10.3 (HMMs and Viterbi)  
**Background older IE reading:**  
Douglas E. Appelt. 1999. *Introduction to Information Extraction Technology*  

Lecture 7  
Wed 4/23/08  
**Maximum Entropy Classifiers**  
**Assigned Reading:**  
J&M secs 6.6-7 (maximum entropy models)  
**Additional references:**  
M&S section 16.2  

Section 3  
Fri 4/25/08  
**Corpora and other resources**  

Lecture 8  
Mon 4/28/08  
**Maximum Entropy Sequence Classifiers**  
**Assigned Reading:**  
J&M secs. 6.0-6.4 and 6.8-6.9 (HMMs in detail and then MEMMs).  
**Other references:**  
Adam Berger, *A Brief Maxent Tutorial*  
Distributed today: Final project guide  

Lecture 9  
Wed 4/30/08  
**IE and text mining**  
**Assigned reading:** J&M secs. 22.2, 22.4.  
**Maxent NER reading:** Jenny Finkel et al., 2005. *Exploring the Boundaries: Gene and Protein Identification in Biomedical Text*  
Due today: Programming Assignment 2  
Distributed today: Programming Assignment 3  

Section 4  
Fri 5/2/08  
**Maximum entropy sequence models**  

Lecture 10  
Mon 5/5/08  
**Syntax and Parsing for Context-Free Grammars (CFGs)** Parsing, treebanks, attachment ambiguities. Context-free grammars. Top-down and bottom-up parsing, empty constituents, left recursion, and repeated work. Probabilistic CFGs.  
**Assigned reading:** J&M ch. 13, secs. 13.0-13.3.  
**Background reading:** J&M ch. 9 (or M&S ch. 3). This is especially if you haven’t done any linguistics courses, but even if you have, there’s useful information on treebanks and part-of-speech tag sets used in NLP.  

Lecture **Dynamic Programming for Parsing** Dynamic programming for parsing. The CKY algorithm. Accurate
unlexicalized PCFG parsing.

**Assigned reading:** J&M sec. 13.4


**Due today:** final project proposals

### Section 5

**Fri**

5/9/08

**Lecture**

[Moved forward from 5/19/08] Semantic Role Labeling

**Assigned reading:** J&M secs. 19.4, 20.9

**Further reading:**


### Lecture

Lexicalized Probabilistic Context-Free Grammars (LPCFGs)

Lexicalization and lexicalized parsing. The Charniak, Collins/Bikel, and Petrov & Klein parsers.

**Assigned reading:** J&M ch. 14 (you can stop at the end of sec. 14.7, if you'd like!)

**Alternative reading:** M&S Ch. 11

**Optional readings:**


### Lecture

Modern Statistical Parsers

Search methods in parsing: Agenda-based chart, A*, and "best-first" parsing. Dependency parsing. Discriminative parsing. **Assigned reading:** J&M ch. 14 (you can stop at the end of sec. 14.7, if you’d like!)

**Alternative, less up-to-date reading:** M&S 8.3, 12

**Optional readings:**


Lecture 15  Computational Semantics
Wed 5/21/08 An Informal but Respectable Approach to Computational Semantics
J&M ch. 18 (you can skip secs. 18.4 and 18.6-end, if you wish).

Mon 5/26/08 Memorial Day
no class

Lecture 16  Compositional Semantics II
Wed 5/28/08 An Informal but Respectable Approach to Computational Semantics
J&M ch. 18 (you can skip secs. 18.4 and 18.6-end, if you wish).

Lecture 17  Lexical Semantics
Mon 6/2/08 Reading: (Okay, I'm not so naive as to think that you'll actually read this in week 9 of the quarter....) J&M secs. 19.0-9.3. Further reading: J&M secs 20.0-20.8 (not included in reader, I'm afraid). "-[slides: see last time]
I. Androutsopoulos et al., Language Interfaces to Databases

Lecture 18  Question Answering (QA) TREC-style robust QA, textual inference

Monday 6/9/08 Final Project Presentations
Students will give short (~5 min) presentations on their final projects during the time slot allocated for a final exam.