University of Illinois at Chicago
Fall 2012

CS 521 — Statistical Natural Language Processing
Course Syllabus

Room: LH 320
Time: MWF 12-12:50
URL: via Blackboard, cs521.fall.2012

Instructor: Barbara Di Eugenio
Office: 916 SEO
Phone: 6-7566
E-mail: bdieugen@cs.uic.edu
Office Hours: TBA

Course Objectives

The field of Natural Language Processing (NLP), also called Computational Linguistics or Human Language Technology, studies the processing of human languages, from both a linguistic and a computational / technological perspective. In the last twenty years, NLP has come to rely more and more on machine learning and statistics. These techniques exploit large corpora and support applications that are more robust that their symbolic counterparts, however at the expense of linguistic perspicuity. This course is meant to provide students with both the foundations necessary to understand cutting edge research papers that use these techniques in any area of NLP; and with an in-depth and critical look at some of the “hottest” applications that use those techniques, especially semantic inferencing, question answering, opinion mining and sentiment analysis, dialogue management, summarization.

Important Note: Frequency

This class is offered every TWO years only.

Reading Materials

Required Textbook: Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. The MIT Press, 1999
Articles from the literature.

Prerequisite

CS 421 or equivalent, or permission of the instructor. To be allowed in the class, students must have foundations in at least one of Artificial Intelligence, Information Retrieval, Machine Learning / Data Mining. Please discuss your background with the instructor.
Additionally, students who don’t have NLP background, or whose NLP background is rusty, must study Chapter 3 in the textbook on their own.

**Important Note: Laptop Usage in class**

I don’t mind if you use your laptop in class but its usage must be related to class – ie taking notes. Hearing constant typing is distracting to the instructor and classmates. I reserve the right to ask you to close it down if I find it disruptive.

**Tentative Schedule**

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td><strong>Weeks 1-7: Foundations</strong></td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>Introduction</td>
<td>Ch. 1, 4</td>
</tr>
<tr>
<td>Week 3</td>
<td>Basics: Probability, Information Theory</td>
<td>Ch. 2</td>
</tr>
<tr>
<td>Week 4-5</td>
<td>N-gram models</td>
<td>Ch. 6</td>
</tr>
<tr>
<td>Week 6</td>
<td>Markov Models</td>
<td>Ch. 9</td>
</tr>
<tr>
<td>Week 7</td>
<td>Probabilistic CFGs</td>
<td>Ch 11-12</td>
</tr>
<tr>
<td></td>
<td>Maximum Entropy, SVMs, CRFs, ...</td>
<td></td>
</tr>
<tr>
<td>Week 8 &amp; 9</td>
<td>Semantic Inferencing, Question Answering</td>
<td></td>
</tr>
<tr>
<td>Week 10</td>
<td>Opinion Mining and Sentiment Analysis</td>
<td></td>
</tr>
<tr>
<td>Week 11</td>
<td>Language in Social Media</td>
<td></td>
</tr>
<tr>
<td>Week 12</td>
<td>Dialogue Processing</td>
<td></td>
</tr>
<tr>
<td>Week 13-14</td>
<td>Summarization, Natural Language Generation,</td>
<td></td>
</tr>
<tr>
<td>Week 15</td>
<td>Catch up, or Project Presentations</td>
<td></td>
</tr>
</tbody>
</table>

From week 8 on, we will read articles from the literature. Each student will be asked to present 1 or 2 papers, and to be the discussant for 4 or 5 of those papers – this means writing a short written critique for the paper in question and be ready to participate in discussion. Exact workload will depend on class size.

**Important Dates**

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week of 10/8</td>
<td>Project Proposal</td>
</tr>
<tr>
<td>Week of 10/15 (??Wed 10/17)</td>
<td>Midterm (Open Book)</td>
</tr>
<tr>
<td>Week of 11/19</td>
<td>Project “in progress” report</td>
</tr>
<tr>
<td>Week of 12/10 (Finals Week)</td>
<td>Project Presentations</td>
</tr>
</tbody>
</table>
Grading Criteria

The class will be graded as follows:

- **Midterm**: 25%.
- **Paper Presentation(s)**: 10%.
- **4-5 Paper Critiques**: 4-5% each, 20% total.
- **Project**: 45%, of which:
  - project originality, work execution, thoroughness, etc: 25%;
  - proposal and intermediate report 5%; final presentation 5%; final project report 10%.

Possible readings

Here's a list of potential papers we may discuss. It’s not an exhaustive list by any means, it is meant to provide you with some ideas about the kind of papers we will be looking at.

**Semantic Inferencing, and Question Answering.** [Mooney, 2008; Moschitti et al., 2008; Toutanova et al., 2008; Chen et al., 2010; Branavan et al., 2012; Liang et al., 2013], papers from Proceedings of Textual Entailment Challenges (e.g., http://www.nist.gov/tac/2011/RTE/index.html)

**Sentiment Analysis, Opinion Mining and Recommender Systems** [Hu and Liu, 2004; Choi et al., 2006; Ding et al., 2008; Carenini and Cheung, 2008; Somasundaran et al., 2009; Saggion and Funk, 2010; Tata and Di Eugenio, 2012]

**Topic Modeling and Summarization** [Radev et al., 2003; Nastase, 2008; Branavan et al., 2009; Nenkova, 2013]

**Social Media Language Processing** Papers from workshops on Language in Social media, e.g. http://research.microsoft.com/en-us/events/lsm2012/

**Dialogue Processing** [Rieser and Lemon, 2008; Forbes-Riley and Litman, 2008; Di Eugenio et al., 2010; Mairesse and Walker, 2010]

**Innovative Applications** NL and video retrieval: papers from e.g. ACM SIGMM conferences http://www.acmmm12.org and TRECVID conferences http://www.itl.nist.gov/iaui/894.02/projects/trecvid/ NL interfaces for educational applications: [Woo et al., 2006; Fosatti et al., 2009], papers from Educational Data Mining Conference, http://educationaldatamining.org/ NL for medical applications:[Jha and Elhadad, 2010]
References


[Fossati et al., 2009] Davide Fossati, Barbara Di Eugenio, Stellan Ohlsson, Christopher Brown, Lin Chen, and David Cosejo. I learn from you, you learn from me: How to make iList learn from students. In *AIED09, the 14th International Conference on Artificial Intelligence in Education*, Brighton, Great Britain, 2009.


