University of Illinois at Chicago  
Fall 2015  

CS 521 — Statistical Natural Language Processing  
Course Syllabus  

Room:  B10 BH  
Time:  MWF 12-12:50  
Class Web Site:  via Blackboard  

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

Course Objectives  

The aim of this course is to cover empirical methods that are widespread in Natural Language Processing (NLP), and to provide an overview of current, cutting-edge research in the field. 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, language in social media, human-robot interaction.

At this point, a more appropriate name for this course should be Applied NLP, since 99% of NLP these days uses empirical methods. However, historically the field of 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; hence when this course was conceived, its title was meant in opposition to more linguistic oriented approaches to NLP.

Important Note: Frequency  

This class will not be offered next year (16-17).

Reading Materials  


2. **Additional materials** from: Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. The MIT Press, 1999

3. **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.

IMPORTANT. Students who don’t have NLP background, or whose NLP background is rusty, must study the following in the textbook on their own: sec 4.1-4.2; sec. 5.1-5.3, and 5.5 up to p. 144; ch 12; ch 13 up to p. 448; sec. 21.4-21.6; sec. 24.1

Important Note: Laptop Usage in class

I don’t mind if you use your laptop / tablet 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 (8/24)</td>
<td>Introduction, n-grams, smoothing</td>
<td>Ch. 4</td>
</tr>
<tr>
<td>Week 2 (8/31)</td>
<td>Tagsets, Intercoder Agreement</td>
<td></td>
</tr>
<tr>
<td>Week 3 (9/9-11)</td>
<td>HMMs</td>
<td>Ch. 5, 6</td>
</tr>
<tr>
<td>Week 4 (9/14)</td>
<td>Information Theory in NLP</td>
<td>Sec. 4.10, 4.11; Sec 2.2 Manning</td>
</tr>
<tr>
<td>Week 5 (9/21)</td>
<td>Maximum Entropy</td>
<td>Sec. 6.7, 6.8</td>
</tr>
<tr>
<td>Week 6 (9/28)</td>
<td>SVMs, LDA, RTM ...</td>
<td></td>
</tr>
<tr>
<td>Week 7-8 (10/5, 10/12)</td>
<td>Semantic Inferencing, Question Answering</td>
<td></td>
</tr>
<tr>
<td>Week 9-10 (10/19, 10/26)</td>
<td>Sentiment Analysis, Language in Social Media</td>
<td></td>
</tr>
<tr>
<td>Week 11-12 (11/2, 11/09)</td>
<td>Dialogue Processing, Multimodal Interaction</td>
<td></td>
</tr>
<tr>
<td>Week 13-14 (11/16, 11/23)</td>
<td>BioNLP, NLP for health sciences</td>
<td></td>
</tr>
<tr>
<td>Week 15 (12/1)</td>
<td>Catch up, and / or Project Presentations</td>
<td></td>
</tr>
</tbody>
</table>

From week 7 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 2 or 3 other 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 10/5 (??Wed 10/7)</td>
<td>Midterm (Open Book and/or Notes)</td>
</tr>
<tr>
<td>Week of 10/12</td>
<td>Project Proposal</td>
</tr>
<tr>
<td>Week of 11/9</td>
<td>Project “in progress” report</td>
</tr>
<tr>
<td>Week of 12/7 (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% each.
- **2-3 Paper Critiques**: 5% each.
- **Class Participation**: 2%
- **Project**: The remaining points, divided as follows:
  - project originality, work execution, thoroughness, etc: half of project points;
  - proposal + intermediate report $\frac{1}{3}$; final presentation $\frac{1}{3}$; final report $\frac{1}{4}$.

The project can be done in pairs.

## Possible readings

Here’s a list of potential papers we may discuss, or conferences where papers on the topic are presented. 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** [Carenini and Cheung, 2008](#), [Somasundaran et al., 2009](#), [Saggion and Funk, 2010](#), [Tata and Di Eugenio, 2012](#), [Recasens et al., 2013](#), [Yang and Cardie, 2013](#)

Dialogue Processing  Proceedings of SIGDIAL (ACL-ISCA Special Interest group on Discourse and Dialogue), e.g. http://aclweb.org/anthology/sigdial.html#2014_0.
[Di Eugenio et al., 2010] [Mairesse and Walker, 2010] [Chen and Di Eugenio, 2013] [Janarthanam et al., 2013] [Fang et al., 2014]

BioNLP, NLP for health sciences  Proceedings of workshops on biomedical NLP, health text mining http://aclweb.org/anthology/sigbiomed.html#2012_0
http://aclweb.org/anthology/W/W14/#1100
https://aclweb.org/anthology/W/W15/#3800
[Jha and Elhadad, 2010] [Angelova and Boytcheva, 2011] [Raghavan et al., 2012] [Mayfield et al., 2013] [Hassan et al., 2015]

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

Innovative Applications  NL and video retrieval: papers from e.g. ACM SIGMM conferences http://www.sigmm.org/view_conference
NL interfaces for educational applications: papers from workshops on NLP for Educational Applications https://aclweb.org/anthology/W/W15/#4400; Educational Data Mining Conferences, http://educationaldatamining.org/

References


