

University of Illinois at Chicago
Spring 2005

CS 594 — Statistical Natural Language Processing Course Syllabus

Room: 1325 SEO

Time: MW 12:00 – 1:15

URL: via Blackboard

Instructor: Barbara Di Eugenio

Office: 916 SEO

Phone: 6-7566

E-mail: bdieugen@cs.uic.edu

Office Hours: TBA

Course Objectives

In the last 10 years, the field of Natural Language Processing (NLP) has come to rely more and more on statistical techniques. These techniques exploit large corpora and are the foundations of applications that are much more robust than their symbolic counterparts, however at the expense of cognitive perspicuity. This seminar 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 applications that use those techniques, especially parsing, semantic inferencing, dialogue understanding and generation, summarization.

Reading Materials

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

Articles from the literature.

Reference book on NLP: Daniel Jurafsky and James H. Martin. Speech and Language Processing. Prentice Hall, 2000 (on reserve in the library for 421).

Prerequisite / Corequisite

CS 421

Tentative Schedule

Dates	Topic	Readings
Weeks 1-7	Foundations	
Week 1	Introduction	Ch. 1, 4
Week 2	Mathematical Foundations	Ch. 2
(1/17)	Martin Luther King Day, no class	
Week 3-4	N-gram models	Ch. 6
Week 5-6	Markov Models	Ch. 9
Week 7	Supervised / unsupervised classification	
Weeks 8-15	Applications	
Week 8	POS tagging, word sense disamb	[Bri95, Yar95]
Week 9-10	Parsing	[Col03, Hwa04]
(3/21-25)	Spring Break, no class	
Week 11	Semantic Inferencing	[GJ02]
Week 12	Discourse and Dialogue Processing	[SLKW02]
Week 13	Natural Language Generation, Summarization	[BR00, BL04]
Week 14	Catch up	
Week 15	Project Presentations	

Important Dates

Date	Event
2/2 (W)	Grammar quiz
3/2 (W)	Project Proposal
4/6 (W)	Project “in progress” report
4/25-27 (M-W)	Project Presentations

Grading Criteria

The class will be graded as follows:

- **Quiz on English Grammar:** 5%.
- **Paper Presentation(s):** 20%.
- **5 Paper Critiques:** 5% each, 25% total.
- **Project:** 50%.

References

- [BL04] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proc. of NAACL-HLT*, 2004.
- [BR00] Srinivas Bangalore and Owen Rambow. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th Conference on Computational Linguistics (COLING'2000)*, 2000.
- [Bri95] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–566, 1995.
- [Col03] Michael Collins. Head-driven statistical methods for natural language parsing. *Computational Linguistics*, 29, 2003.
- [GJ02] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3), 2002.
- [Hwa04] Rebecca Hwa. Sample selection for statistical parsing. *Computational Linguistics*, 30(3), 2004.
- [SLKW02] Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research (JAIR)*, 15:105–133, 2002.
- [Yar95] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.