CSCI 562
Statistical Natural Language Processing

University of Southern California
Fall 2012

Time:
Tu/Th, 11:00am–12:20pm
Location:
KAP 148
Instructors:
Prof. David Chiang (chiang@isi.edu)
Prof. Kevin Knight (knight@isi.edu)
Teaching Assistant:
Hui Zhang (hzhang@isi.edu)

Course Description

Computers process massive quantities of information every day in the form of human language, yet machine understanding of human language remains one of the great challenges of computer science. How can advances in computing technology enable more intelligent processing of all this language data? Will computers ever be able to use this data to learn language like humans do? This course provides a systematic introduction to *statistical models of human language*, with particular attention to the *structures of human language* that inform them and the *structured learning and inference algorithms* that drive them. This is a lecture course, not a seminar course, but aims to cover both fundamental and cutting-edge research issues.

Audience This graduate course is intended for:

- students who want to understand current natural language processing (NLP) research,
- students interested in tools for building NLP applications,
- machine learning students looking for large-scale application domains, and
- students seeking experience with probabilistic methods that can be applied to a range of AI problems.

Prerequisites Students are expected to be proficient in programming, basic algorithms and data structures, discrete math, and basic probability theory. Undergraduate students are welcome as well. If you are not sure if you have enough background for this course, you are welcome to sit in on the first few lectures. There will be a Quiz 0 to test if you qualify for this course.

Related Courses This course is part of USC's curriculum in natural language processing. There is a sister course, Natural Language Processing (CSCI 544), offered in the Spring semester. You can take these two in either order.

Textbooks

- Class notes (to be distributed in class).
- Recommended but optional: Jurafsky and Martin, *Speech and Language Processing* (2nd ed.), Prentice Hall, 2008.

Requirements and policies

Coursework Students will experiment with existing NLP software toolkits and write their own programs. Grades will be based on:

- Quizzes (4 × 5% of grade): at the end of each unit.
- Programming assignments (5 × 12%): graded according to the correctness and clarity of the solutions. Sometimes a small part of the grade will depend on the performance of a system relative to the rest of the class.
- Research project (20%): graded according to the project's substantiality, correctness, and relevance to the course, as well as the clarity and depth of the project report.

All assignments and the project will be due at the beginning of class (11am) on the due date. Late assignments will be accepted with a 30% penalty up to one week late. No exceptions can be made except for a grave reason.

Students are expected to submit only their own work for homework assignments. They may discuss the assignments with one another but may not collaborate with or copy from one another.

Statement on Academic Integrity USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. *Scampus*, the Student Guidebook, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS.

Syllabus

Date	Topic	Instructor	Slides/Readings	Assignments
Tue, Aug 28	What is natural language processing? What is the empirical approach to NLP?	Knight		HW0: nothing to turn in.
Thu, Aug 30	Introduction to language.	Chiang		Quiz 0
	Unit One: Sequence Models (10 lectures)			
Tue, Sep 4	Basic automata theory. Finite-state acceptors and transducers.			
Thu, Sep 6				HW1 out
Tue, Sep 11				
Thu, Sep 13 *Add/drop deadline*	Probability theory and estimation. Weighted FSTs. The noisy-channel model.	Knight		
Tue, Sep 18				HW1 due
Thu, Sep 20	Language models and their applications. Learning models from data.			HW2 out
Tue, Sep 25		TBD		
Thu, Sep 27				

Tue, Oct 2 Thu, Oct 4	String transformations and their applications.	Knight	HW2 due Quiz 1
	Unit Two: Learning Language (8 lectures)		
Tue, Oct 9 Thu, Oct 11 Tue, Oct 16 Thu, Oct 18	Unsupervised training. Expectation-Maximization. The Forward-Backward algorithm.	Knight	HW3 out
Tue, Oct 23 Thu, Oct 25	Discriminative training. The perceptron algorithm. Conditional random fields. Max-margin training.	Chiang	HW3 due HW4 out
Tue, Oct 30			Quiz 2
Thu, Nov 1	Bayesian inference.	Knight	
	Unit Three: Tree Models (7 lectures)		
Tue, Nov 6 Thu, Nov 8	Context-free grammars (CFGs). Parsing. Inside-Outside algorithm.	Chiang	HW4 due
Tue, Nov 13		TBD	Initial project proposal due
Thu, Nov 15 *Withdraw deadline*	Statistical parsing.	Chiang	HW 5 out
Tue, Nov 20			Final project proposal due
Thu, Nov 22	Thanksgiving – no class		
Tue, Nov 27 Thu, Nov 29	Synchronous CFGs and machine translation.	Chiang	HW5 due Quiz 3
Tue, Dec 4 Thu, Dec 6	Project interim presentations	You	
Wed, Dec 19			Final projects due

Suggested Readings

[LY90] The estimation of stochastic context-free grammars using the inside-outside algorithm.
K. Lari and S. J. Young.
Computer Speech and Language, 1990: 4:35-56.

[SSP95] Principles and Implementation of Deductive Parsing.
Stuart Shieber, Yves Schabes, and Fernando Pereira.
Journal of Logic Programming, 1995: 24 (1-2): 3-36.

[H08] Tutorial on Semirings and Advanced Dynamic Programming in NLP (slides).
Liang Huang
Tutorial given at COLING 2008 and NAACL 2009.

[KS09] Tutorial on Writing Systems, Transliteration, and Decipherment (slides).
Kevin Knight and Richard Sproat
Tutorial given at NAACL 2009.

[CG98] An Empirical Study of Smoothing Techniques for Language Modeling
Stanley F. Chen and Joshua Goodman
Technical Report TR 10-98, Computer Science Group, Harvard University.

[DLR77] Maximum Likelihood from Incomplete Data via the EM Algorithm
A.P. Dempster, N.M. Laird, and D.B. Rubin
Journal of the Royal Statistical Society, Vol. 39, No. 1, 1977

[C97] The EM Algorithm
Michael Collins

[D02] The Expectation Maximization Algorithm
Frank Dellaert
Technical Report, 2002

[B06] The Expectation Maximization Algorithm -- A short Tutorial.

Sean Borman

Some useful Python modules

- bigfloat: floating point numbers with very large range