USC Viterbi School of Engineering

CSCI662 Advanced Natural Language Processing

Units: 4
Fall 2017, TuTh 10am-11:50pm

Location: WPH B28

- Course materials and discussions through **Piazza**.
- Homeworks submitted through **Blackboard**.

Instructor: Prof. Kevin Knight
Office: SAL 242
Office Hours: Tuesdays 9am-10am (SAL 242)
Contact Info: knight@isi.edu

Teaching Assistant: Nada Aldarrab
Office: TBD
Office Hours: TBD
Contact Info: naldarra@usc.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.

Learning Objectives

This graduate course is intended for PhD students who want to deepen their understanding of natural language processing, machine learning, and automata theory. The course covers both foundations and current research.

Prerequisite(s): None.
Co-Requisite (s): None.
Concurrent Enrollment: None.
Recommended Preparation: Students are expected to be proficient in programming, algorithms and data structures, discrete math, and basic probability theory.

Course Notes
This course is letter graded. Lecture notes, supplementary material, and announcements are posted on Piazza, where student discussions are also held. Students submit homeworks via Blackboard.

Technological Proficiency and Hardware/Software Required
Students must have access to computational resources and are expected to be proficient in programming.

Required Readings and Supplementary Material
The course revolves around a set of lecture notes, distributed on the first day of class. A recommended but optional textbook is Jurafsky and Martin, *Speech and Language Processing* (2nd ed.). The course uses the following software:

- Carmel (http://www.isi.edu/licensed-sw/carmel). C++-based string automata toolkit.
- Tiburon (http://www.isi.edu/licensed-sw/tiburon). Java-based tree automata toolkit.
- TensorFlow (https://www.tensorflow.org/install). Recurrent neural network toolkit.

Description and Assessment of Assignments

- The course has 5 programming assignments, 10 quizzes, and a final research project. These teach how to construct natural language processing systems and stress empirical experimentation.
- Programming assignments are graded according to the correctness and clarity of the solutions. A small part of the grade may depend on the performance of a system relative to the rest of the class.
- Quizzes are graded according to correctness and clarity.
- The research project is graded according to the project's substantiality, correctness, and relevance to the course, as well as the clarity and depth of the project report.

Grading Breakdown

Assignment	Points	% of Grade
Programming assignment #1	50	6.25%
Programming assignment #2	100	12.5%
Programming assignment #3	150	18.75%
Programming assignment #4	100	12.5%
Programming assignment #5	100	12.5%
Quiz #1	10	1.25%
Quiz #2	10	1.25%
Quiz #3	10	1.25%
Quiz #4	10	1.25%
Quiz #5	10	1.25%
Quiz #6	10	1.25%
Quiz #7	10	1.25%

Quiz #8	10	1.25%
Quiz #9	10	1.25%
Quiz #10	10	1.25%
Final Project	200	25%
Total	**800**	**100%**

Assignment Submission Policy
Assignments must be submitted on Blackboard by midnight on the due date.

Additional Policies

- Students are expected to submit only their own work for homework assignments. They may discuss the assignments with one another but may not collaborate on answers.
- *Late assignments will be accepted with a 30% penalty, up to one week late.*

Course Project

The purpose of the course project is to give students the opportunity to work on a research project in natural language processing. Students may work singly or in groups of at most two. An individual project is roughly double the size of an average homework assignment, and a group project is roughly double the size of an individual project. Students may not submit the same project to CSCI622 and another class, but students may choose a project that is part of a larger research program.

Project Timeline:

- Initial project proposal due at beginning of class (on Blackboard).
- Final project proposal due at beginning of class (on Blackboard). This should include a report of your data preparation and baseline, which should be completed by this time.
- Interim project presentation (in class).
- Final project write-ups due (on Blackboard).

The project topic must concern statistical learning of the structure of natural language. It must treat a text as more than a bag of words, and it should learn a model (rather than, for example, a similarity function). The initial proposal must include: (1) a clear statement of the goal, and what would constitute success, (2) concrete description of the data to be used, and what processing is needed to make it usable, (3) the evaluation method, (4) the baseline solution method, (5) the proposed method, and (6) a sample results chart, with blank cells.

Sample projects. The instructors will provide a list of sample topics. However, students are strongly encouraged to devise projects of personal interest outside the provided list. Instructors can provide additional topics and will work with students to refine all project proposals. Sample projects include: (1) Unsupervised or discriminative context-free parsing using the Penn Treebank, (2) HMM word

alignment using bilingual Canadian Hansards, UN, or EU proceedings, (3) Automatically deciphering the Copiale manuscript, and (4) Translating passages from Dante's Divine Comedy from Italian into English, maintaining verse.

Grading breakdown of the course project:

- Proposal: 10%
- Interim in-class presentation: 10%
- Final report: 80%

Course Schedule

Date		Topic	Lecturer	Assigned
Aug	22	Introduction to NLP, applications. Basic stats. Text processing.	Knight	HW0 out (no credit)
	24	Finite state acceptors & transducers (FSA, FST). Weighted machines (WFSA/T), noisy channel model.	Knight	
	29	Language modeling.	Knight	HW1 out (WFSA)
	31	"	Knight	Quiz 1
Sep	5	String transformations and applications.	Knight	HW1 due
	7	"	Guest	Quiz 2
	12	**WORKSHOP**	Guest	HW2 out (WFST)
	14	**WORKSHOP**	Guest	Quiz 3
	19	Unsupervised learning, expectation maximization (EM).	Knight	HW2 due
	21	"	Knight	Quiz 4
	26	Efficient EM (forward-backward algorithm).	Knight	HW3 out (Unsup. ML)
	28	Decipherment.	Knight	Quiz 5
Oct	3	Bayesian inference.	Knight	
	5	Discriminative training (CRF, Perceptron).	Knight	Quiz 6
	10	Discriminative training (MERT).	Knight	HW3 due
	12	Context-free grammar (CFG), treebanks, parsing.	Knight	Quiz 7
	17	"	Knight	HW4 out (Parsing)
	19	Regular tree grammars (RTG), tree transducers.	Knight	Quiz 8
	24	Tree transformations and applications.	Knight	HW4 due
	26	Inside-outside algorithm.	Knight	Quiz 9
	31	Neural network models, distr. representations.	Knight	HW5 out (seq2seq)
Nov	7	Neural language models.	Knight	Quiz 10
	9	Neural sequence-to-sequence models.	Knight	HW5 due
	14	Neural network interpretation.	Knight	Project proposal due
	16	Dependency parsing, synchronous grammars (SCFG, STSG) & adjoining (TAG).	Knight	
	21	Meaning representation, semantic parsing, language generation, graph automata.	Knight	Project proposal due (revised to final)
	23	NO CLASS – THANKSGIVING	n/a	
	28	Interim Project Reports.	Students	
	30	Interim Project Reports.	Students	
Dec	11			Course project due

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* *http://emergency.usc.edu* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.