CSE6390 3.0 Special Topics in AI & Interactive Systems II
Introduction to Computational Linguistics
Insructor: Nick Cercone – 3050 CSEB – nick@cse.yorku.ca
Tuesdays,Thursdays 10:00-11:30 – South Ross 104
Fall Semester, 2010

CSE6390 Course Calendar (10 September 2010 update)

#	Date	Title	Asgn's
	Part I: Computational Linguistics, Language, Natural Language Processing, Theory and Applications		
1	14 Sept 10	**Course Introduction** Course information: overview of course; logistics and administrivia, textbook and other main references, evaluation scheme, academic honesty policy, tentative course schedule; resources Introduction to computational linguistics and natural language processing (NLP); what is a natural language and other kinds of languages; cchallenges for language processing; what is Computational Linguistics; short history of CL/NLP; example applications Handouts: course description; active reading; paper writing; resources; adjectives and adverbs; knowledge representation; assignment initial; assignment big; assignment big grading; course calendar Files: Lecture 1 notes (ppt).	**A0 out** **A2 out**
2	16 Sept 10	**Introduction to CL & NLP** Some reasons why NLP is hard; ambiguities at all levels of NLP, examples of ambiguities; domain knowledge is useful: to interpret questions, to answer questions, to model the user. Some philosophy of language – representation, interpretation of adjectives and adverbs. Handouts: whatisCL?; semantic nets; updated course calendar Files: Lecture 2 notes (ppt).	**A0 due**
3	21 Sept 10	**Introduction to CL & NLP** Some example NLP applications: NL interfaces to databases (SystemX), NL interfaces to internet search engines (NLAISE & EMATISE), machine translation (GRMT). Handouts: updated course calendar; word sense disambiguation; sketch of word sense disambiguation; Yarowsky algorithm; machine learning introduction; machine learning book Files: Lecture 3 notes (ppt).	
	Part II: Linguistic Background - Unification-based approach to NLP		
4	23 Sept 10	**Words and Morphology** Words, words, words; morphemes, stems, affixes, stemming, morphological processes: inflection, derivation, compounding, clitics; Parts-of-speech (POS), POS tagging, open and closed categories, corpus linguistics	

		Handouts: Yawowsky paper; Synder & Palmer paper; Files: Lecture 4 notes (ppt).	
5	28 Sept 10	**Lexical Categories, Logic, Syntax, Grammar** Explain handouts; Lexical categories; POS tagging examples; Logic and resolution principle theorem proving and its role in NLU; Syntax: phrase structure, phrases, clauses, sentences; parsing, parse tree examples; Context-Free Grammars (CFG); Are NLs context-free? review: examples, parse trees, Handouts: assignment small; assignment small grading; project suggestions; project grading; heuristics; lexical category; early syntax theory; logic and resolution; chapter 3 (parsing) and chapter 18 (annotated bibliography) of Grune & Jacob's book on Parsing Techniques (second edition); papers – Earley CFG parser, Kaplan's lexical function grammar, Shieber's non CFGness of NL, Files: Lecture 5 notes (ppt).	**A1 out**
6	30 Sept 10	**NL Grammar Hierarchies** Class exercise; More notes on regular expressions, finite state automata, Markov algorithms, CFG, Typical phrase structure rules in English: Sentence (S), Noun Phrase (NP), Verb Phrase (VP), Prepositional Phrase (PP), Adjective Phrase (ADJP), Adverbial Phrase (ADVP); Handouts: updated course calendar; In-class exercise; Regular expressions, finite state machines and the pumping lemma; Markov algorithms; Carlo Strapparava's handout on FSA and regular expressions; Tutorial book – Picking up Perl; Practical Earley Parsing Files: Lecture 6 notes (ppt).	
7	5 Oct 10	**Parsing and Context Free Grammars** CFG; derivations, language generated by a CFG, left-most and right-most derivations, ambiguous sentences, bracketed representation of parse trees; Natural Language Phenomena: agreement, movement, subcategorization; Handouts: Lisp materials – Quickie Lisp, Good Lisp Style, Cooper's Book, Getting starterd in GNU Common Lisp. McCarthy's paper, Graham's book; Left corner paring; BNF grammars; Files: Lecture 7 notes (ppt).	
8	7 Oct 10	**Semantics and Pragmatics** Heads and dependency; head-feature principle, dependency trees, arguments and adjuncts; Elements of semantics: semantic analysis, lexical semantics: word senses Handouts: Subcategorization; Sample projects – ugproject1, ugproject2, bronislova; Presentation; Communications; Student projects documentation; student projects revised; updated course calendar; Regular expressions and finite state automata; Ratnaparkhi's statistical parser Files: Lecture 8 notes (ppt).	
9	12 Oct10	**Unification-based approach to NLP** Some final parsing and semantics examples; Unification-based approach to NLP; bits of history, First-order predicate logic: constants, variables, functions, terms, predicates, formulae, sentences, axioms, theorems, inference rules; examples, Resolution-based inference system by Robinson; Unification Handouts: Prolog material – Learn Prolog now, Logic programming and Prolog, Prolog book; Lisp code for regular expression parser; Representational typology	**A1 Due**

		Files: Lecture 9 notes (ppt).	
10	14 Oct 10	**HPSGs** Unification review, HPSG Introduction, Principles, Rules, Examples, Modularity Handouts: Intro to HPSG; ALE manual; Elementary principles of HPSG; Encyclopaedia HPSG; HPSG Linguistic approach; Foundations of HPSG; Flickinger's thesis Files: Lecture 10 notes (ppt).	
11	19 Oct 10	**HPSGs** How its done, Examples, Examples, Examples Handouts: Hermes NL access; Hermes grammar and lexicon; Lexical rules; Prolog and NL analysis; review of Prolog and NL analysis; Charniak's edge based chart parser; Logic-Based Implementation of Head-Driven Phrase Structure Grammar; HPSG grammars in ALE; Prolog compared to Lisp; Python tutorial; Categorial grammar Files: Lecture 11 notes (ppt).	
12	21 Oct 10	**Final HPSG, Statistical Approach to NLP** Handouts: Chapter 1 of Manning's book; Chapter 1 of Jurafsky's book; common n-gram method; Shannon's 1948 paper; Shannon's 1951 paper; Statistical NLP paper; Probability for linguists; Using Python book; Files: Lecture 12 notes (ppt).	
	Part III: Statistical Approach to NLP - Statistical Methods in NL Processing and Data Analysis and Part V (1st part) Student Presentations		
13		**Information Retrieval and the Vector Space Model (Jun Miao)** Typical IR system architecture, steps in document and query processing in IR, vector space model, tfidf - term frequency inverse document frequency weights, term weighting formula, cosine similarity measure, term-by-document matrix, reducing the number of dimensions, Latent Semantic Analysis, IR evaluation Handouts: nlp09.pdf; VectorSpaceImplementation-6per.pdf; 07Models-VSM.pdf; E09-3009.pdf; ieee-sw-rank.pdf; ir4up.pdf; p613-salton.pdf; Polettini Information Retrieval.pdf; 2.doc; Vector space model.doc Files: Lecture 13 IR and VSM notes (ppt).	
14		**Text Classification (Zahidur Rahman – now Nick Cercone)** Text classification and text clustering, Types of text classification, evaluation measures in text classification, F-measure, Evaluation methods for classification: general issues - over fitting and under fitting, methods: 1. training error, 2. train and test, 3. n-fold cross-validation Handouts: nlp10.pdf; 10[1].1.1.4.4417.pdf; chap16.pdf; IG-Mercer-Kernel-	

		Performance-2006.pdf; joachims_98a.pdf; lodhi02a.pdf; Text categorization.doc; Text classification and Naive Bayes.doc; http.doc Files: Lecture 14 Text Classification (ppt); Lecture 14 – cohen (ppt); lecture 14 - Rosen-Zvi (ppt)	
15		**Parser Evaluation, Text Clustering and CNG Classification (Ameeta Agrawal)** Parser evaluation: PARSEVAL measures, labeled and unlabeled precision and recall, F-measure; Text clustering: task definition, the simple k-means method, hierarchical clustering, divisive and agglomerative clustering; evaluation of clustering: inter-cluster similarity, cluster purity, use of entropy or information gain; CNG -- Common N-Grams classification method Handouts: nlp11.pdf; 10e-eval-2x3.pdf; 0712.3705.pdf; 774_paper.pdf; acl07parseval.pdf; D07-1066.pdf; getPDF.jsp.pdf; lre98.pdf; p9-clark.pdf; p37-lewis.pdf; p60-simov-ranlp03.pdf; pe08rimell_constructing.pdf; syntax.pdf; versley-tlt05.pdf; 13.doc Files: Lecture 15 6390E_Mee_Parser_Clustering_CNG(ppt).	
16		**Probabilistic Modeling and Joint Distribution Model (Haluk Madencioglu)** Elements of probability theory, Generative models, Bayesian inference, Probabilistic modeling: random variables, random configurations, computational tasks in probabilistic modeling, spam detection example, joint distribution model, drawbacks of joint distribution model Handouts: nlp12.pdf; nlp13.pdf; 08Models-Prob.pdf; 10[1].1.1.23.9849.pdf; ECIR2008TutorialHiemstra-new.pdf; Fuhr_92.pdf; IR-Probablistic-strategy.pdf; lecture20.pdf; Model_challenges1.doc; Please check the on.doc Files: Lecture 16 haluk-presentationn (pdf).	
17		**Fully Independent Model and Naive Bayes Model (Nikolay Yakovets)** Fully independent model, example, computational tasks, sum-product formula; Naive Bayes model: motivation, assumption, computational tasks, example, number of parameters, pros and cons; N-gram model, language modeling in speech recognition Handouts: nlp14.pdf; nlp15.pdf; 10[1].1.1.48.529.pdf; 10[1].1.1.65.9324.pdf; 10[1].1.1.73.5412.pdf; 10[1].1.1.112.8246.pdf; KDD96-061.pdf; AA28.txt Files: Lecture 17 FullyIndependentAndNaiveBayesModels-NY (pdf).	
18		**N-gram Model (Bahareh Sarrafzadeh)** N-gram model: n-gram model assumption, graphical representation, use of log probabilities; Markov chain: stochastic process, Markov process, Markov chain; Perplexity and evaluation of N-gram models, Text classification using language models Handouts: nlp16.pdf; 10[1].1.1.87.754.pdf; 01342667.pdf; aaac.pdf; D07-1045.pdf; DalTREC05spam.pdf; fulltext.pdf; IJCAI09-252.pdf; J92-4003.pdf; N03-1020.pdf; pacling05a.pdf; pst04.pdf; N-Grams.html; henke-ch6.ppt; Lecture4N-Grams.ppt Files: Lecture 18 Ngram Models (pps).	
19		**Hidden Markov Model (Rados Jovanovic)** Smoothing: Add-one (Laplace) smoothing, Bell-Witten smoothing; Hidden Markov Model, graphical representations, assumption, HMM POS example, Viterbi algorithm -- use of dynamic programming in HMMs.	

		Handouts: nlp17.pdf; C96-2141.pdf; hmm14.pdf; hmm tutorial; For a tutorial on HMM's see.doc Files: Lecture 19 (ppt).	
20		**Bayesian Networks (Nariman Farsad)** Bayesian Networks, definition, example, Evaluation tasks in Bayesian Networks: evaluation, sampling, inference in Bayesian Networks by brute force, general inference in Bayesian Networks is NP-hard, efficient inference in Bayesian Networks, Handouts: nlp18.pdf; bayesinf05.pdf; bayesnet09.pdf; bayesstruct05.pdf; BN.pdf; gaussbc12.pdf; naive02.pdf; tr-95-06.pdf; shortbayes03.pdf; prob18.pdf Files: Lecture 20 (ppt).	
21		**Probabilistic Parsing (Dmitri Shuralyov)** PCFG as a probabilistic model; Computational tasks for PCFG model: Evaluation, Learning, Simulation, proper PCFG,Probabilistic inference: marginalization, efficient inference, CYK algorithm Handouts: nlp21.pdf; nlp22.pdf; 1104.pdf; acl2003-chinese.pdf; C00-1017.pdf; DOPlecture.pdf; iicall06.pdf; lex-parser.pdf; P04-1069.pdf; p406-nederhof.pdf; unlexicalized-parsing.pdf; Sampson/ Probabilistic Parsing; The Stanford NLP (Natural Language Processing) Group; Probabilistic parsers on the web.doc; thadh-meissnem-1-PA3report.doc Files: Lecture 21 (ppt).	
	Part V (2nd part): Student Project Presentations		
22		**Student Project Presentations**	**A2 Due**
23		**Student Project Presentations**	
24		**Wrap-up and Course Review**	
		Projects Due	