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

In Class Presentation

Students will make two in-class presentations, a formal presentation on course material and an informal presentation based on the student's course project. The formal presentation/seminar will be 60 minutes followed by 10 minutes for questions. At the end of the semester two classes will be reserved for the shorter *informal* presentation by each student on his/her project in order to inform the class of the student's project. This *informal* presentation is intended to solicit feedback from students in the class and the instructor on any aspects of the project, and for critical commentary to take place between presenter and class. The shorted (project) presentation will count for 3% of your grade. The longer formal presentation will be a presentation of selected course material to the class and will count for 12% of your grade. Topics for the formal presentation include:

1. Information Retrieval and the Vector Space Model
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

2. Text Classification
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

3. Parser Evaluation, Text Clustering and CNG Classification
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

4. Probabilistic Modeling and Joint Distribution Model
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

5. Fully Independent Model and Naive Bayes Model
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

6. N-gram Model
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

7. Hidden Markov Model
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.

8. Bayesian Networks
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,

Each student should pick a topic as soon as possible and inform me by email of their choice. Topics will be assigned on a first-come first-serve basis. If two or more students pick the same topic the student who responded first will be assigned it. Students without topics by class on *October 28, 2010* will have one of the remaining topics assigned to them.

Topics will be scheduled for presentation as follows:

Topics	Date	Topics	Date
1	2 Nov 2010	**5**	16 Nov2010
2	4 Nov 2010	**6**	18 Nov 2010
3	9 Nov 2010	**7**	23 Nov 2010
4	11 Nov 2010	**8**	25 Nov 2010

The formal presentations will be graded according to the following criteria (12 points):

1. Organization 2 ____
2. Completeness 1 ____
3. Presentation 6 ____
   a. competence
   b. content
4. Understanding 2 ____
5. Materials submitted 1 ____

Notes:

Please keep within the times allotted for the presentations and make them crisp and to the point. Do not read your slides or make them too "busy". Try to get across the ideas and not get bogged down into minute details. Practice your presentations with friends first. Good luck.

The informal project presentations will be graded according to the following criteria (3 points):

1. Organization & Backgound 1 ____
2. Project degree of difficulty/effort expended 1 ____
3. Presentation & Understanding 1 ____