CS 674 Spring 2000 - Project Guidelines

Project grades are based on the following four parts:

- Proposal: due March 15 in class, but by noon March 17 (hand in to Cindy Robinson, Upson 4146) is fine.
- Preliminary literature survey: due April 10
- Presentations: May 4-9 (tentative; I will be out of town May 1-3, hence the strange dates.)
- Final report: due May 16

Projects should be focused and include careful evaluation. For example:

- **Bad:** I implemented a neural net to assign part-of-speech tags.
- **Good:** I studied various factors impacting the performance of neural nets as part-of-speech taggers, including the number of nodes, the training algorithm used, and the amount of training data available. I discovered that the amount of training data was the key factor, corroborating the claims of X.
- **Bad:** I read six papers on Martian discourse. A summary of each paper follows.
- **Good:** Current theories of Martian discourse provide different accounts of the Double Anaphoric Missing Negative (DAMN). I considered six different mechanisms proposed for dealing with DAMN, and devised a uniform approach incorporating the good points of the aforementioned six methods.

Your *proposal* should be 2-3 pages, and must include a preliminary list of related references. You should describe what problem you intend to consider, what has been previously proposed (this can be sketchy, since you probably haven't done much background reading yet), what evaluation you intend to do, and what schedule you will follow for accomplishing your goals. Programming project proposals should list the resources you will use; survey proposals should include the questions you will "ask" of each paper you read.

For programming projects, the following resources are available: Brill's part-of-speech tagger, the Brown corpus (English, mixed subjects, with POS tags), MUC4 text (English, same subject, with POS tags), and WordNet. You are responsible for making sure that you can acquire and/or build what other resources you need – this should be reflected in your proposal schedule.

Your *preliminary literature survey* should be approximately five pages long plus bibliography, and should represent some of the major papers on the problem you are examining.

The *final report* should be about ten pages plus appendices for programming projects and about fifteen pages plus bibliography for surveys. For other types of projects, such as the presentation of a new theoretical result, fewer pages may be appropriate. In all cases, a thorough discussion of related work is expected.

Sources of ideas include the recommended texts and readings, the list of last year's projects (see below) and the numerous conference proceedings that are on reserve at the Engineering Library: a useful activity is to pick a few and scan through the abstracts. Once you have a preliminary idea, feel free to see me at office hours or make an appointment to talk about it.

(over)

Last year's projects:

- Stylometric Clustering: A comparrative analysis of data-driven and syntactic features (experiment)
- Learning Programming Language Parsing Rules (experiment)
- CFGs and Matrices and TAGs, oh my! (theoretical project)
- An Analysis of Automatically Constructed Thesauri (survey)
- Anaphora Resolution Algorithms on Noun Phrases (experiment)
- Word Sense Disambiguation in Malay (experiment)
- The Importance of Different Parts of Speech to Word Sense Disambiguation (project)
- Part of Speech Tagging: An Experiment in Extending the Unsupervised Learner of Disambiguation Rules
- A Linguistics Bias Case-Based Learning Approach to Anaphora Resolution (experiment)
- Word Sense Disambiguation with Semantic Networks: Evaluating Resnik's Semantic Similarity Measure in the Disambiguation of Free Text (experiment)
- An Unsupervised Grammar Inducer and Parser Based on part of Speech n-Grams (experiment)
- Tree Adjoining and Combinatory Categorial Grammars as Natural Language Formalisms, or How I Learned to Stop Worrying and Love the Lambda Calculus (theoretical project)
- Introducing Explicit Syntactic/Semantic Representation into a Statistical Machine Translation System: An Investigation of the Use of Linguistic Information to Assist the Candide MT System (theoretical project)