Human Language Technologies  
CSC 791T - Special Topics  

Fall 2006

Course Description
Human Language Technologies (HLT) is the field of artificial intelligence that focuses on the automated understanding and generation of human language. With a broad array of applications in conversational interfaces, web search, document analysis, and text mining, HLT has become the subject of growing interest in both academia and industry. HLT tasks range from information retrieval, information extraction, question answering, and text mining to text-based and spoken dialog understanding. Variously known as Natural Language Processing and Computational Linguistics, HLT has made significant strides in recent years with the introduction of probabilistic, statistical, corpus-based, and machine learning techniques.

This course presents the foundations of HLT and explores frontier HLT tasks. It introduces the principles and methods underlying the design of scalable computational models of human language with an emphasis on statistical frameworks, and it explores the issues bearing on the design of accurate and efficient solutions. It begins by introducing the fundamentals of language modeling, part-of-speech analysis, grammars, parsing, and semantic and pragmatic analyses, and then proceeds to information retrieval problems including text classification and text clustering. As time permits, it will also cover advanced topics such as open domain question answering, information extraction, text mining, natural language generation, and summarization. Cognitive and linguistic phenomena will be considered throughout the course.

Course Objectives

- Students will be able to comfortably read and critically assess the HLT literature.
- Students will be able apply HLT models to a given problem.
- Students will be able to apply empirical evaluation methodologies to assess HLT techniques.
- Students will be able to design new HLT models and implement them to create language analysis tools that solve a given problem with a specified level of accuracy and efficiency.
Instructor

James Lester
lester@csc.ncsu.edu
919.515.7534
EB2 2264

Prerequisites

- Graduate standing in Computer Science or Bioinformatics
- CSC520 or the equivalent (special arrangements will be made with Bioinformatics students)

Texts

The course will use two textbooks, one presenting a broad overview of the field and the other offering a specialized treatment of statistical approaches:


Syllabus

- **Introduction**
  - Introduction to Natural Language Processing
  - History and State of the Art in HLT
  - Linguistic Essentials

- **Introduction to Statistical HLT**
  - Mathematical Foundations
  - Information Theory (Entropy, The Noisy Channel Model)
  - Language Modeling (*n*-gram Models)

- **Sub-Sentential Processing**
  - Tokenization and Morphological Analysis
  - Large Corpora Methodologies and Techniques
  - Part-of-Speech Tagging (Transformation-Based Learning, HMMs)
  - Spelling Correction
• Parsing
  o Shallow Parsing and Chunking
  o CFGs
  o Lexicalized Parsing
  o Features and Unification
  o Probabilistic CFGs and Probabilistic Parsing

• Semantics and Pragmatics
  o Reference and Ambiguity Resolution (Anaphora, Ellipsis)
  o Attachments
  o Word Sense Disambiguation
  o Discourse
  o Dialogue Management (Text-Based and Speech-Based Conversational Agents)

• Information Retrieval
  o Introduction to Search and IR
  o Text Representation and Indexing
  o Search Engines
  o k-Nearest Neighbor Text Classification
  o Text Clustering

• Advanced Topics (Topic selection guided by student interest)
  o Information Extraction
  o Open Domain Question Answering
  o Natural Language Generation
  o Text Summarization
  o Text Mining

Course Scope
HLT covers a broad subject matter and some areas (e.g., those above) must necessarily be emphasized over others (e.g., lexical semantics, machine translation). Statistical approaches will be emphasized over symbolic approaches, and issues in representing and processing text will be emphasized over those in speech (e.g., computational phonology, speech recognition).

Format and Grading
The course will primarily follow a lecture format, augmented with student presentations. Grading will be based on problem sets, in-class presentations, class participation, and a course project. Selection and scoping of projects will be arranged with the instructor. Each student will give a presentation toward the end of the semester on his or her project.