CS 784 Computational Linguistics – Winter 2025 – Professor Freda Shi

Introduction

Computational Linguistics is the branch of Computer Science that studies the computer processing, comprehension, and generation of natural (i.e., human) languages. This course will provide students with both theoretical and practical knowledge in the fundamentals of syntax, semantics, and pragmatics. It will provide the basic background for students planning to do a thesis in this area. It will also be a self-contained course for anyone interested in natural language processing (e.g., including computer scientists, humanists, cognitive scientists, linguists, and researchers in arts computing).

Textbook

Jurafsky and Martin (2024). Speech and Language Processing

Outline

- Words: definition, tokenization, morphology, word senses

- Lexical semantics: distributional semantics, word embeddings, word clustering

- Text Classification: classifiers, linear models, features, naive Bayes, training linear classifiers via loss function optimization, loss functions, stochastic gradient descent

- Neural Network Basics: MLP, CNN, RNN and Transformers, training NNs

- Language Modeling: n-gram models, smoothing, neural network--based language modeling

- Sequence Labeling: part-of-speech tagging, named entity recognition, hidden Markov models, conditional random fields, dynamic programming, Viterbi

- Syntax: weighted context-free grammar, dependency syntax, inference algorithms

- Semantics: compositionality, semantic role labeling, frame semantics, lambda calculus, semantic parsing, grounded semantics

- Pragmatics: phenomena, rational speech act model

- Cross-lingual NLP: translation, decoding, lexicon induction, unsupervised translation

- Large language models: challenges, prompting

Prerequisites

Basic knowledge of calculus, linear algebra, and probability; programming proficiency (no specific language required but Python is preferred); a machine learning/AI course (e.g., CS 480/680, CS 486/686) is recommended but not required.