USC CSCI 662 (Advanced Natural Language Processing) Syllabus – Fall 2024

This is a survey course on Natural Language Processing aimed at PhD students who are well equipped to do research but may not have had formal exposure to topics. It is “advanced” in the sense that students are expected to independently read papers and follow up on topics they don’t understand. Prior experience in NLP is not required. The first part of the course covers fundamental tools such as linear and nonlinear classifiers, feed forward, recurrent, and transformer neural network architectures, data preparation and gathering. The second part will cover various tasks and advanced topics. Students will present short talks on papers from this year’s top conferences. Topics are subject to change without warning due to instructor or class whim. Active participation and debate are encouraged.

Prerequisites

There are no formal prerequisites, however you should have a good understanding of the following (though we will review):

- Linear algebra (vector and matrix math basics)
- Probabilities: random variables, discrete and (some) continuous distributions, Bayes’ Theorem, Chain Rules
- Calculus: mostly derivatives, or the ability to refresh this info
- Programming: python using a Linux environment, mostly *without* a Jupyter notebook

Instructors

- Jonathan May (Professor). Office hours 12-1M and 9-10 W in TBD or by appointment
- Alexander Spangher (TA). Office hours TBD

Contact Us

- Course website: https://jonmay.github.io/USC-CS662/
- slack, office hours, or in class. Do not email.

Grading

- 10% in-class participation
- 10% questions during assigned slots for paper presentation and project presentation (5% each)
- 10% in-class selected paper presentation
- 30% three homeworks (10% each)
- 40% project (done in small groups), over four components:
    - proposal (5%)
    - report v1 (5%)
    - in-class presentation (10%)
    - report v2 (20%)

Reading

- No purchased material is required! Everything should be available on line for free; contact an instructor if you are having trouble finding material.

- Required textbooks:
    - Jurafsky and Martin: "Speech and Language Processing" (2023 draft)
    - Eisenstein: "Natural Language Processing" (2018 github)
- Required papers – TBD, see schedule on website, but watch for changes!

Lectures (subject to change)

- Introduction
- Data Basics
    - Basic Processing
    - Basic Resources
- Linear Classifiers
    - Naive Bayes
    - Perceptron
    - Logistic Regression
- Nonlinear Classifiers
    - XOR problem
    - Feed-Forward NN
- Distributional Features
    - PPMI
    - Word2Vec
- Language Models
    - N-gram statistical
    - N-gram feed-forward
    - Recurrent (LSTM)
- Transformer
- Pretrained Language Models
- Prompting and Large Language Models
- Beyond Transformer (Mega, State space models)
- Reinforcement Learning with Human Feedback
    - PPO
    - DPO
- Ethics
- More on Data
    - Sources
    - Annotation
        - Agreement statistics
    - For Evaluation
        - statistical significance tests
- Syntax
    - Part-of Speech Tags
    - HMM decoding
    - Syntax Trees
    - Constituencies and CFG Parsing
    - Dependencies and Shift-Reduce Parsing

- Semantics
    - Lexical
    - Compositional
    - Semantic Structures (AMR, maybe CCG)
- Machine Translation
    - History of Translation
    - Evaluation approaches
        - BLEU
        - Other non-BLEU model-free types (e.g. Meteor)
        - COMET
        - BLEURT (maybe: YiSi)
    - Statistical MT outline
    - LSTM with Attention
    - Transformer Revisited
- Dialogue
    - Task-oriented
        - State Tracking
        - Slot Filling
        - Evaluation
    - Free chat
- Information Extraction
    - History of the topic
    - Entity Recognition
    - Event Recognition
    - Co-reference and grounding
- Question Answering and Information Retrieval
- Agents
- Discourse
- Special Topics (TBD; subset of potential areas listed)
    - Legal NLP
    - Multimodal NLP
    - Speech Recognition and Synthesis