Introduction and Logistics

Rob Minneker

Fire Safety and Evacuation Plan

- https://healthsafety.cs.washington.edu/wp-content/uploads/2025/06/Gates-Center-FSEP-2025.06.05.pdf
  - Please review and be aware of your exits
  - No elevators
  - Meet at assembly point (Rainier Vista is safe bet)

COURSE OVERVIEW AND STRUCTURE

What is NLP? NLU? LM? LLM? LMM? what???

- NLP: Natural language processing
- NLU: Natural language understanding
- LM: Language model
- LLM: Large language model
- LMM: Large multimodal model
- NLP is a broad term that encompasses all things computers + *human* language (both understanding and generating): e.g. speech recognition, language translation, sentiment analysis, summarization, etc.

Course Objectives (1/4)

Natural Language Processing (NLP) investigates computational approaches to analyzing, generating, and understanding human language.

- Objective: Develop a rigorous understanding of core NLP concepts and algorithms
  - Formal models: probabilistic, neural, and symbolic paradigms
  - Foundational tasks: parsing, tagging, language modeling

Course Objectives (2/4)

- Emphasis on practical, hands-on experience with real-world NLP applications
  - Implementation of core methods (e.g., text classification, natural language generation)
  - Use of standard datasets and evaluation metrics/approaches

Course Objectives (3/4)

- Critical engagement with current NLP research literature
  - Analyzing recent advances and their limitations
  - Understanding open problems and future research directions

Applications:

- Machine translation, question answering, information extraction
- Language generation, summarization, sentiment analysis

Course Objectives (4/4)

- Learning outcomes:
  - Formulate and analyze NLP models mathematically
  - Design, implement, and critically evaluate NLP systems
  - Synthesize and critique cutting-edge NLP research
  - Leverage these tools/systems to do good things

Course Structure and Logistics (1/4)

- The course is organized around lectures, assignments, a final project and interactive participation, with no final exam or in-class quizzes and no quiz section
- Emphasis is placed on continual engagement and critical thinking rather than high-stakes testing
  - This structure encourages regular review and deeper understanding of NLP topics

Course Structure and Logistics (2/4)

- Assignments are the primary means of assessment
  - Each student has 5 penalty free late days
    - You may use a maximum of 3 days (72 hours) on any one assignment
  - Anything when late days are exhausted will receive no credit

Course Structure and Logistics (3/4)

- Grading breakdown:
  - Major components: assignments, final project, and participation
  - Contributions:
    - Assignments (50%)
      - A1 (15%), A2 (15%), A3 (20%)
    - Project (50%)
      - Various checkpoints (mini assignments) throughout the quarter
    - Participation (up to 4%)
      - Course surveys, completion-based activities

Course Structure and Logistics (4/4)

- Assignments
  - 3 longer assignments with coding/analysis, notebook based
    - Largely auto graded (hidden tests)
    - 517 has extra problem or so, 447 can do for some extra points
    - 2 weeks from release to submission
- Project
  - More details on Thursday
    - tldr; Group project, 517 has more ambiguity

Teaching Staff

- Instructor: Rob
- TAs:
  - Victoria (Lead)
  - Anagha
  - Ben
  - Boe
  - Jize
  - Khushi
  - Min
  - Praveer

NATURAL LANGUAGE PROCESSING: SCOPE AND SIGNIFICANCE

Definition and Scope of NLP (1/4)

Natural Language Processing (NLP) is the computational study of human language, aiming to enable machines to interpret, generate, and interact using natural language.

Definition and Scope of NLP (2/4)

- NLP spans multiple levels of linguistic analysis:
  - Phonology: study of speech sounds and their patterns
  - Morphology: structure of words and morphemes
  - Syntax: grammatical structure of sentences
  - Semantics: meaning of words, phrases, and sentences
  - Pragmatics: context-dependent aspects of meaning and use

Definition and Scope of NLP (3/4)

- Core problems in NLP involve mapping between forms and meanings:
  - **Parsing:** mapping input text to syntactic structures
    - e.g., deriving a parse tree for “The cat sat on the mat.”
  - **Tagging:** assigning linguistic labels (e.g., POS tags) to tokens
    - e.g., “flies” as a verb or a noun depending on context
  - **Translation:** converting text from one language to another
    - e.g., English → French: “Hello” ↦ “Bonjour”
  - **Question Answering:** extracting or generating answers from text
    - e.g., Given “What is the capital of France?” answer “Paris is the capital of France”

Definition and Scope of NLP (4/4)

- Now is the time of LLMs (or GenAI more broadly)
  - ChatGPT
  - DeepSeek
  - Claude
  - GitHub Copilot
  - Llama Etc.
  - Not just text-to-text, also text-to-image, text-to-video, … (*multimodality*)

Historical Perspective (1/3)

- Early rule-based approaches:
  - Relied on hand-crafted linguistic rules, such as context-free grammars (CFGs).
  - Example: Parsing with a formal grammar where production rules define valid sentences.

$$S \to NP\ VP$$

- Limitations: Poor scalability and brittle handling of real-world ambiguity.

Historical Perspective (2/3)

- Rise of statistical NLP:
  - Introduced probabilistic models, leveraging large corpora and frequency statistics.
  - Example: Hidden Markov Models (HMMs) for part-of-speech tagging.

$$P(T,W) = \prod_{i=1}^{n} P\left(w_i \mid t_i\right) P(t_i \mid t_{i-1})$$

- Enabled robust disambiguation and data-driven inference.

Historical Perspective (3/3)

- Advent of neural and deep learning methods:
  - Shifted focus to distributed representations and end-to-end learning.
  - Example: Word embeddings (e.g., word2vec) model words as dense vectors, capturing semantic similarity.
  - Deep architectures (e.g., transformers) achieve state-of-the-art results in diverse NLP tasks.

Applications:

- Machine translation, question answering, sentiment analysis

NLP in AI and Society (1/3)

Natural Language Processing (NLP) enables computational systems to analyze, understand, and generate human language, serving as a core component in modern artificial intelligence.

- NLP bridges human communication and machine intelligence
  - Facilitates information exchange between users and AI-driven systems
  - Essential for tasks requiring understanding of unstructured, ambiguous text

NLP in AI and Society (2/3)

- Key applications in contemporary technology:
  - Search engines: interpret queries, rank and retrieve relevant documents
  - Virtual assistants: parse spoken/written language, execute user commands
  - Information extraction: identify entities, relations, and facts from text corpora
  - Summarization: condense large documents while preserving salient information

NLP in AI and Society (3/3)

Societal Impact and Ethical Considerations:

- Widespread deployment of NLP systems raises critical issues:
  - Algorithmic bias: NLP models may amplify social biases present in training data
  - Privacy: processing sensitive language data can compromise user confidentiality
  - Misinformation: automated generation and summarization may propagate errors or distortions

FUNDAMENTAL CONCEPTS AND THEORY

Probability and Statistics in NLP (1/4)

Probability and statistics provide the mathematical foundation for modeling uncertainty and learning from data in NLP.

- Discrete probability spaces underpin language modeling
  - The probability of a sequence $w_1, w_2, \dots, w_n$ is given by

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P\left(w_i \mid w_1, \dots, w_{i-1}\right)$$

Probability and Statistics in NLP (2/4)

- Markov assumptions (e.g., n-gram models) simplify modeling by limiting context
- Maximum Likelihood Estimation (MLE) is used to fit statistical language models
  - For unigram models:

$$\hat{P}(w) = \frac{\mathrm{count}(w)}{\sum_{w'} \mathrm{count}\,(w')}$$

- MLE provides unbiased estimators under i.i.d. assumptions

Probability and Statistics in NLP (3/4)

- Evaluation metrics for NLP tasks are grounded in statistical decision theory
  - Accuracy: proportion of correct predictions over all cases
  - Precision, recall, and F1 for imbalanced or structured outputs:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

Probability and Statistics in NLP (4/4)

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Applications:

- Language modeling, speech recognition, and information extraction rely on probabilistic models and statistical estimation
- Evaluation metrics inform model selection and error analysis in NLP research

Machine Learning Paradigms in NLP (1/4)

Machine Learning paradigms provide foundational frameworks for modeling linguistic phenomena in NLP, shaping how language data is represented, learned from, and generalized.

Machine Learning Paradigms in NLP (2/4)

- **Supervised vs. Unsupervised Learning**
  - Supervised learning uses labeled data (x, y) to learn a mapping $f_\theta: x \rightarrow y$ (e.g., text classification, NER).
  - Unsupervised learning infers structure from unlabeled data x, e.g., clustering or learning word embeddings.

Machine Learning Paradigms in NLP (3/4)

- **Feature Representation for Language**
  - Early NLP relied on hand-crafted features: n-grams, POS tags, parse trees.
  - Modern approaches use distributed representations: word embeddings (e.g., Word2Vec, GloVe), contextual vectors from transformers.
  - Formally, embeddings map discrete tokens to $\mathbb{R}^d$:

$$\text{Embed}: w \mapsto \mathbf{v}_w \in \mathbb{R}^d$$

Machine Learning Paradigms in NLP (4/4)

- **Transfer Learning and Pre-trained Models**
  - Large models are pre-trained on massive corpora via self-supervised objectives (e.g., masked language modeling).
  - Pre-trained representations are fine-tuned for downstream tasks, dramatically reducing labeled data requirements.
  - Example: BERT, GPT, and their variants set new state-of-the-art on many NLP benchmarks.

COURSE LOGISTICS AND RESOURCES

Course Website and Communication (1/3)

Course Website and Communication

- The course website serves as the authoritative source for schedules, policies, and resource links.
  - All lecture materials, assignment deadlines, and readings are posted here.
  - Updates to the syllabus or schedule are reflected promptly.
  - https://courses.cs.washington.edu/courses/cse447/26wi
  - cs.uw.edu/447 also works for any current course

Course Website and Communication (2/3)

- Use Ed!
  - Students are encouraged to post questions, share insights, and discuss assignments.
  - Please tag your posts appropriately to help others find them and dedup efforts

Course Website and Communication (3/3)

- Official announcements are communicated via Ed
  - Students must monitor institutional email accounts regularly.
  - Time-sensitive updates (e.g., schedule changes) are distributed through multiple channels.

Assignment and Project Logistics (1/2)

- Homework guidelines:
  - Regular submission deadlines; late policy strictly enforced
  - Submissions via Gradescope
  - Each assignment targets core concepts
    - Text Classification and N-gram language models
    - Neural Text Classification and Neural Language Modeling
    - Transformers and Natural Language Generation

Assignment and Project Logistics (2/2)

- Academic integrity and collaboration:
  - Discussion of concepts encouraged; all code and writeups must be individual
  - Plagiarism, including copying from peers or online sources, is a violation
  - Proper attribution required for external resources or libraries
  - GenAI use is permitted and encouraged (must cite this)
    - Particularly for studying or targeted coding help; less useful for conceptual questions on assignments

Next time

- Project overview/requirements
- Words and Tokens (~J&M Ch. 2)

Sources

Content derived from: course website, prior quarters