CS 784: Computational Linguistics
Lecture 1: Introduction

Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

January 7, 2025

Enrollment

- 700-level CS course: Lecture oriented graduate course.
- Non-CS and/or non-grad students will be on the waitlist.
  The CS grad office will issue permission numbers to clear the waitlist.
  - Undergrads will also need to sign a course override form.
  - Non-CS grad students are automatically on the waitlist with a confirmation email from me.
- First batch of waitlist will be submitted today.
- We have more interest than we can accommodate, so the waitlist order will be determined by a quiz in the first lecture.

People, Websites, and Books

Instructor: Freda Shi

Lectures: T/Th 4:00pm-5:20pm, DC 2568

Office hours: T/Th 5:20pm-5:50pm, DC 2568

Course website:
`https://cs.uwaterloo.ca/~fhs/teaching/wat-cs784-w25`

Discussion forum:
`https://piazza.com/uwaterloo.ca/winter2025/cs784`

People, Websites, and Books

Instructor: Freda Shi

Lectures: T/Th 4:00pm-5:20pm, DC 2568

Office hours: T/Th 5:20pm-5:50pm, DC 2568

Course website:
https://cs.uwaterloo.ca/~fhs/teaching/wat-cs784-w25

Discussion forum:
https://piazza.com/uwaterloo.ca/winter2025/cs784

There is no required textbook for this course. However, we will primarily refer to the following resources:

- Jurafsky and Martin, *Speech and Language Processing*, 3rd edition
  https://web.stanford.edu/~jurafsky/slp3/
- Clyde et al., *An Introduction to Bayesian Thinking*
  https://statswithr.github.io/book/

Other Recommended Readings

Here are some other readings that may be helpful to understand the course material:

- Jacob Eisenstein, *Introduction to Natural Language Processing* `https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf`
- Emily M. Bender, *Linguistic Fundamentals for Natural Language Processing*, I and II `https://link.springer.com/book/10.1007/978-3-031-02150-3` and `https://link.springer.com/book/10.1007/978-3-031-02172-5` (*Note: You may login with your WATID for free access to this book.*)

Grade Breakdown

- Assignments (30%: 15% each) – Due on February 12 and March 19
  Each assignment consists of a coding exercise and a paper review exercise.
- Course Project (40%) – Due on March 26
  Individual meta-analysis/survey of Machine Learning/AI literature focusing on an ambiguous keyword.
  Projects must be completed independently.
- Project Peer Review (10%) – Due on April 9
  Critical review of a peer's project.
- Final Examination (20%) – Date TBD (during final exam period)
  Format: Open-book and open-notes examination
  You will analyze provided research results using course methodologies, and design follow-up studies based on the presented findings.

Course Project

A survey/meta analysis of AI/ML literature, centered on an ambiguous keyword.

Candidate words include:

- alignment
- attention
- bias
- causal
- classifier
- grounding
- reality
- valence

Or any other keyword—please check with the instructor first.

Default corpus: `https://huggingface.co/datasets/Seed42Lab/AI-paper-crawl`

You may use any other relevant corpus—please check with the instructor first.

1-page midterm checkin by February 28: describe which word you've chosen, what you've done, and what you plan to do next.

Course Project Evaluation Criteria

- **Midterm Checkin** (5%): Clear description of the chosen keyword and the methodology.
- **Literature Review** (15%): Demonstrate thorough identification and accurate summarization of at least **two** senses of the chosen keyword, supported by appropriate scholarly references.
- **Meta Analysis** (15%): Execute a comprehensive meta-analysis with clear methodology and reproducible results.
- **Written Presentation** (5%): Exhibit clear, coherent, and professional academic writing and adhere to the formatting requirements throughout the report.

Formatting Requirements:

- 4-8 pages (with unlimited references) in the *ACL template: `https://github.com/acl-org/acl-style-files`

Additional Opportunities:

- Outstanding reports will be invited for co-authorship on a meta-analysis paper.
- Extra credit (up to 5%) available for expanding dataset coverage to other AI conferences/journals.

Lateness Policy

All assignments are due at 11:59 PM Eastern Time (Waterloo time) on the specified due date.

A universal 72-hour grace period applies to all assignments (including course project and peer review):

- Submissions within the grace period will be accepted without penalty.
- No additional extensions will be granted beyond the grace period.
- Assignment exemptions will only be considered in cases of medical emergency, provided that:
  - A valid Verification of Illness Form (VIF) is submitted.
  - The VIF covers the original due date and the entire grace period.

  If approved, the assignment weight will be redistributed to the final exam.

Collaboration and Citation Policy

- **Permitted Collaboration:** Discussion of assignments with fellow students is encouraged, but all submitted work (including code and written solutions) must be completed independently.
- **Citation Requirements:** All external sources must be properly cited, including:
  - Academic references and literature
  - Generative AI tools (e.g., ChatGPT) if used
  - Any other resources consulted during assignment completion
- **Student Responsibilities:**
  - Verify the accuracy of any AI-generated content
  - Ensure all submitted work reflects your own understanding
  - Provide complete and accurate citations for all sources

GenAI Policy

You may use GenAI tools however you can to assist your learning and growth, but remember these tools can't be fully trusted and should never replace your own understanding. It is essential to carefully verify the content they generate, especially when working with detailed numbers or using AI to refine your writeups.

Ultimately, you are fully responsible for the accuracy and integrity of all work you submit in this course.

Academic Integrity and Intellectual Property

Property of UW:

- Lecture content, spoken and written (and any audio/video recording thereof).
- Lecture handouts, presentations, and other materials prepared for the course (e.g., PowerPoint slides).
- Questions or solution sets from various types of assessments (e.g., assignments, quizzes, tests, final exams).
- Work protected by copyright (e.g., any work authored by the instructor or TA or used by the instructor or TA with permission of the copyright owner).

Sharing intellectual property without the intellectual property owner's permission is a violation of intellectual property rights.

Prerequisites

There is no formal prerequisite for this course; however, the following background knowledge is strongly recommended:

- Basic knowledge of calculus, linear algebra, and probability.
- Programming proficiency (Python is preferred but not required).
- Fundamentals of algorithms, e.g., complexity analysis for simple algorithms.
- Understanding of basic data structures, e.g., lists, stacks, queues, trees, and graphs.
- Some basic-level machine learning knowledge, e.g., linear regression.

Prerequisites

There is no formal prerequisite for this course; however, the following background knowledge is strongly recommended:

- Basic knowledge of calculus, linear algebra, and probability.
- Programming proficiency (Python is preferred but not required).
- Fundamentals of algorithms, e.g., complexity analysis for simple algorithms.
- Understanding of basic data structures, e.g., lists, stacks, queues, trees, and graphs.
- Some basic-level machine learning knowledge, e.g., linear regression.

Scoring 75% or higher in the quiz later is a good indicator that you have necessary background knowledge.

Quiz in the First Lecture

We have more interest than we can accommodate in this course.

Students not enrolled in CS grad programs will be added to the waitlist—once cleared, the grad office will issue a permission number.

There will be a 30-minute quiz in the first lecture to determine everyone's rank on the waitlist.

Results of this quiz will not affect your grade in the course.

What is Computational Linguistics?

Computational linguistics is the scientific study of **language** from a **computational** perspective.

It is an interdisciplinary field that draws on linguistics, computer science, psychology, philosophy, cognitive science, neuroscience, and speech and hearing science.

What is Computational Linguistics?

Computational linguistics is the scientific study of **language** from a **computational** perspective.

It is an interdisciplinary field that draws on linguistics, computer science, psychology, philosophy, cognitive science, neuroscience, and speech and hearing science.

- **Computational linguistics** is the scientific study of language from a computational perspective.
- **Natural language processing** (NLP) is the application of computational techniques to the analysis and synthesis of natural language and speech.

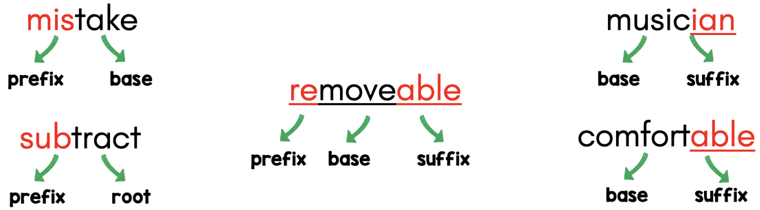
The two terms are sometimes used interchangeably.

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.



mistake
prefix base

subtract
prefix root

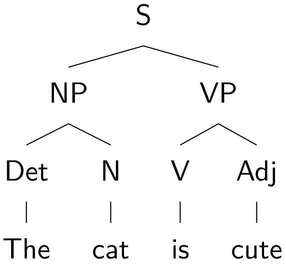
removeable
prefix base suffix

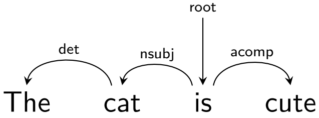
musician
base suffix

comfortable
base suffix

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.





Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.
- **Semantics**: The study of the meaning of words and sentences.

The cat is cute.

$$\lambda x.\mathrm{cat}(x) \wedge \mathrm{cute}(x)$$

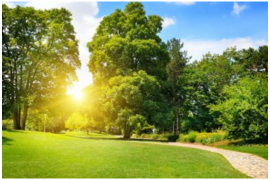
Each *object* refers to a variable (e.g., x), and each property refers to a predicate (e.g., $\mathrm{cat}(\cdot)$).

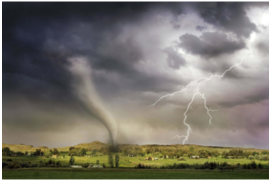
Recently, there has also been a trend of (and also debates around) using vector representations for words and sentences.

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.
- **Semantics**: The study of the meaning of words and sentences.
- **Pragmatics**: The study of how context affects meaning.

What an excellent weather!

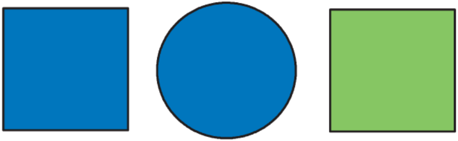




Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.
- **Semantics**: The study of the meaning of words and sentences.
- **Pragmatics**: The study of how context affects meaning.

Imagine you are talking to someone, and you want to refer to the middle object with one word. Which word would you use, **blue** or **circle**?



[Figure credit: Frank & Goodman (2012)]

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.
- **Semantics**: The study of the meaning of words and sentences.
- **Pragmatics**: The study of how context affects meaning.

In this course, we will focus on the last four areas from computational perspectives.

For historical reasons, the fields of computational linguistics and NLP have been more closely associated with text processing.

Phonetics and phonology are more closely associated with speech processing.

Computational Linguistics: Areas

- **Phonetics and phonology**: The study of the sounds of language.
- **Morphology**: The study of the structure of words.
- **Syntax**: The study of the structure of sentences.
- **Semantics**: The study of the meaning of words and sentences.
- **Pragmatics**: The study of how context affects meaning.

In this course, we will focus on the last four areas from computational perspectives.

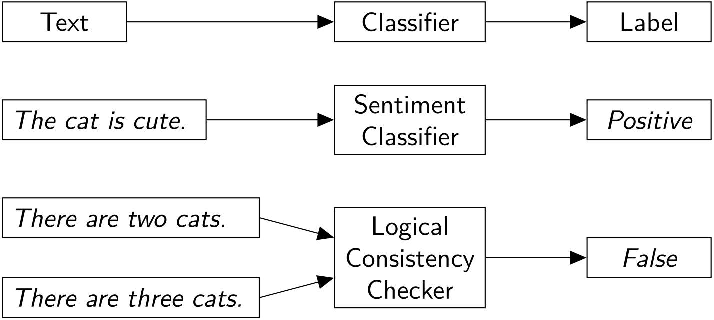
For historical reasons, the fields of computational linguistics and NLP have been more closely associated with text processing.

Phonetics and phonology are more closely associated with speech processing.

Recent years have seen a convergence of the two fields, with many techniques and models being applied to both text and speech.

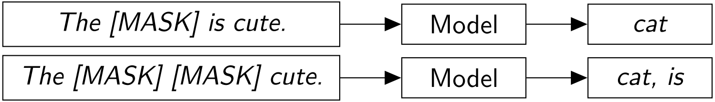
Computational Linguistics: Modeling Approaches

- **Classification**: Assigning a label to some text.



Computational Linguistics: Modeling Approaches

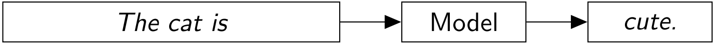
- **Classification**: Assigning a label to some text.
- **Language modeling**: Predicting a word or a few words given the context.
  - **Masked language modeling** (e.g., BERT): predicting a masked word/a (given) number of masked words given the context.



The [MASK] is cute. → Model → *cat*

The [MASK] [MASK] cute. → Model → *cat, is*

  - **Autoregressive language modeling** (e.g., GPT): predicting the next word given the context.



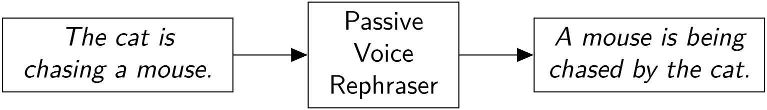
The cat is → Model → *cute.*

  - Autoregressive LM is a special case of masked LM; however, it fits better with generative requirements.
  - Both tasks can be viewed as classification.

[BERT: Devlin et al. (2019), GPT-2: Radford et al. (2019)]

Computational Linguistics: Modeling Approaches

- **Classification**: Assigning a label to some text.
- **Language modeling**: Predicting a word or a few words given the context.
- **Sequence-to-sequence modeling**: Predicting a sequence of words given another sequence of words.


The cat is chasing a mouse.
Passive Voice Rephraser
A mouse is being chased by the cat.

Taxonomy

Above are the taxonomies of CL/NLP tasks from the perspectives of **areas** and **modeling approaches**.

Whenever describing something scientifically, figure out:

- How many factors do we need to consider?
- What are the possible values of each factor?

Taxonomy as Coordination System

- **Classification-based sentiment analysis**: Semantics, classification.

Taxonomy as Coordination System

- **Classification-based sentiment analysis**: Semantics, classification.
- **BERT pretraining**:

Taxonomy as Coordination System

- **Classification-based sentiment analysis**: Semantics, classification.
- **BERT pretraining**: Some morphology, mostly somewhere between syntax and semantics, masked language modeling.

Taxonomy as Coordination System

- **Classification-based sentiment analysis**: Semantics, classification.
- **BERT pretraining**: Some morphology, mostly somewhere between syntax and semantics, masked language modeling.
- **Text summarization**:

Taxonomy as Coordination System

- **Classification-based sentiment analysis**: Semantics, classification.
- **BERT pretraining**: Some morphology, mostly somewhere between syntax and semantics, masked language modeling.
- **Text summarization**: Mostly semantics, sometimes pragmatics, sequence-to-sequence modeling.

Next

Basic methods: Regular expressions (SLP3 Chapter 2.1), probability, and basic information theory (optional: TTIC 31230 Lectures 1 and 8).