Computational Linguistics

Course information

Instructor:	Roger Yu-Hsiang Lo (`roger.lo@xxx.yyy`)
Credits:	3
Time:	Tues/Thur 9:00 AM–10:30 PM
Location:	TBD
Online discussion forum:	TBD
Office hours:	Thurs 1:00 PM–2:00 PM and by appointment

Course overview

Computational linguistics is a scientific discipline focused on understanding written and spoken language through computational methods. In the first half of this course, we will explore foundational concepts and methods in computational linguistics. In the second half, we will examine recent developments and common applications of this field.

We begin with an overview of the key issues before delving into regular expressions and finite state machines, which are not only useful for matching words but also for describing morphological patterns across languages. Next, we will study language modelling, a cornerstone of computational linguistics. We will then explore classification methods and basic model evaluation techniques in machine learning, as many language-related tasks can be framed as classification problems. Subsequently, the course will cover part-of-speech tagging, computational semantics—vector semantics in particular—which have broad applications in natural language processing (NLP). Later, we will focus on neural network-based techniques, the driving force behind most modern language technologies. The course concludes with an exploration of real-world applications of computational linguistics.

Learning objectives

Upon completion of this course, you will be able to:

- Explain key concepts underlying current computational linguistic research;
- Implement algorithms or use existing libraries for common NLP tasks;
- Empirically evaluate the performance of computational models, including their errors.

Prerequisites

LING 250 or SDA 250: *Computational Text Analysis*. You are expected to be comfortable programming in Python, particularly with string manipulation and basic data structures. If you do not meet the prerequisite but still wish to enrol, please speak with me after the first class. Familiarity with probability and calculus will be beneficial but not required.

Course materials

Required:

- Jurafsky, Daniel, and James H. Martin. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models*. 3rd edition. [Book draft]
- Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55:77–84. [Full text]
- Lewis, Patrick et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474. [Full text]
- Xi, Zhiheng et al. 2023. The rise and potential of large language model based agents: A survey. [Full text]

Optional:

- Goldberg, Yoav. 2017. *Neural network methods for natural language processing*. Morgan & Claypool. [Full text]
- Eisenstein, Jacob. 2018. *Natural language processing*. [Full text]

Course format

Most class meetings will be lecture-based, with some lectures incorporating live-coding demonstrations. While you will be able to follow the lectures without completing the assigned readings and videos, I strongly encourage you to review them in advance to deepen your understanding of the topics. Lecture notes will be posted prior to class.

Assessment

- **Class attendance and participation** (10%): Attendance and participation will be assessed through in-class quizzes during lectures.
- **Homework assignments** (60%; 10% per assignment): There will be six mandatory homework assignments designed to help you understand course concepts and techniques and apply them to practical problems. Assignments are **due on Friday at midnight** (see the schedule below for specific dates) and should be submitted electronically. Except for HW1, you may work in groups of up to **three** people for each assignment. Late submissions will incur a 10% deduction for every 24 hours past the deadline. However, you have a five-day "grace period" for **ONE** assignment, allowing for a late submission without any point deduction (Life happens sometimes!). If you choose to use this option, please clearly indicate it on your assignment. If you have an emergency, please reach out to the instructor as soon as possible.
- **Final project** (30%): The final project offers an opportunity for you to explore a topic in computational linguistics in depth. You may choose from the following project types:

- **Independent research**: Extend techniques from a prior study or apply them to your own research;
- **Error analysis**: Conduct a detailed linguistic analysis of errors produced by an NLP application, focusing on what they reveal about the underlying model;
- **Literature review**: Survey and synthesize research on a specific theme in the field.

If you are uncertain about the scope of your project, feel free to consult me at any time. Midway through the course, I will check in to discuss your plans.

Projects can be completed individually or in groups of up to **three** people. The effort for group projects is expected to scale linearly with the number of group members. All group submissions must include a paragraph detailing each member's contributions.

We will also offer a "default project" in lieu of a final project if you are unable to identify a topic of interest. While this alternative exists, I encourage you to take advantage of this opportunity to pursue a self-directed project. Computational linguistics is a booming and exciting field, and you are likely to find a topic that piques your interest!

Grading scale

Percentage grades will be assigned for all assessments and converted to final letter grades based on the following scale:

Letter grade	% grade	Definition
A+	90–100	Excellent performance
A	85–89	
A−	80–84	
B+	76–79	Good performance
B	72–75	
B−	68–71	
C+	64–67	Satisfactory performance
C	60–63	
C−	55–59	
D	50–54	
F	0–49	Unsatisfactory performance (fail)

Communication

For course-related questions, please follow these steps for the quickest response:

1. Consult this syllabus.
2. Post your question on the online discussion forum or ask classmates.
3. Meet with me during office hours.

For personal questions, feel free to email me directly. I aim to respond within 48 hours.

Academic integrity statement

You are expected to abide by SFU's academic integrity. Scientific research is a collaborative effort and relies on proper attribution of each party's contributions. Copying others' work without proper citation is strictly prohibited and violates the ethical standards of the university.

For this course, you are allowed to use artificial intelligence tools, including generative AI, to gather information, review concepts, or to help produce assignments. However, you should be aware that the code/text generated by AI may by inaccurate, biased, or incomplete. You are ultimately accountable for the work you submit, and any content generated or supported by an artificial intelligence tool must be cited appropriately.

Accessibility

- **Accommodation for students with disabilities**: Students requiring academic accommodations due to a disability or medical condition should reach out to Centre for Accessible Learning (CAL). More information is available on this page.
- **Well-being**: Being a student at any level can be challenging. You should always prioritize your well-being if you experience physical or psychological difficulties. Please refer to this page for resources provided by the university.

Schedule & topical outline

Week #	Date	Topics	Readings	Assignment due
Week 1	09/05	**Introduction** - Course overview - Software set-up	- Install `Jupyter Notebook`	
Week 2	09/10 09/12	**Computational morphology** - Regular expressions - Finite state machine - Edit distance	- SLP ch. 2	HW1 due on 09/13
Week 3	09/17 09/19	**Language modelling** - N-grams - Perplexity - Smoothing	- SLP ch. 3	
Week 4	09/24 09/26	**Classification** - Naive Bayes - Logistic regression - Evaluation metrics - Data split	- SLP ch. 4, 5 - Video: Bayes theorem	HW2 due on 09/27
Week 5	10/01 10/03	**POS tagging** - Hidden Markov Model - CRFs	- SLP ch. 17	

Week #	Date	Topics	Readings	Assignment due
Week 6	10/08 10/10	**Computational semantics** - Vector semantics - Embedding - Similarity metrics	- SLP ch. 6 - Video: vectors	HW3 due on 10/11
Week 7	10/17	**Topic modelling**	- Blei (2012)	
Week 8	10/22 10/24	**Neural network** - Feed-forward networks - Backpropagation - Neural language models	- SLP ch. 7 - Video: NN part 1, 2, 3, 4	HW4 due on 10/25
Week 9	10/29 10/31	**RNNs & Transformers** - Vanishing gradients - LSTM - Attention - Layer normalization	- SLP ch. 8, 9 -Video: transformers, attention	
Week 10	11/05 11/07	**Large language models** - Pre-training & fine-tuning - RLHF - Prompt engineering - Jailbreaking	- SLP ch. 10 - Video: LLMs	HW5 due on 11/08
Week 11	11/12 11/14	**Sentiment analysis** **Information extraction**	- SLP ch. 20, 22	
Week 12	11/19 11/21	**RAG** **LLM agent**	- Lewis et al. (2020) - Xi et al. (2023)	HW6 due on 11/22
Week 13	11/26 11/28	**Automatic speech recognition** - CTC - Text-to-speech	- SLP ch. 16	
Week 14	12/03	**What's next**		Final project due on 12/20