Syllabus

CS 678

Advanced Natural Language Processing

Instructors

Ziyu Yao (ziyuyao [at] gmu [dot] edu)
Office Hours: Email for appointments.

TA

TBD (XXX [at] gmu [dot] edu)
Office Hours: TBD

Meets

Thursday, 4:30 to 7:10 PM, Horizon Hall 2016.
Safe Return to Campus: Students are expected to follow the university's Safe-Return-to-Campus Policy (including mask wearing, daily health check, etc.) for attending any classes. Please check out the policy before coming to the campus and the classroom. Note that students who choose not to abide by these expectations will be referred to the Office of Student Conduct for failure to comply.

Course Web Page

https://nlp.cs.gmu.edu/course/cs678-fall23/.
We will use **Blackboard** for course materials/assignments/grading, and **Piazza** for Q&A (sign up link: TBD).

Course Description

Massive amounts of information in our daily life are expressed in natural language. In this class, we will study building computing systems that can process, understand, and communicate in natural language. The class will start with an introduction to the foundations of natural language processing (NLP), and then focus on cutting-edge research problems in NLP. Each section will introduce a particular problem or phenomenon in natural language, describe why it is difficult to model, and demonstrate recent models that were designed to tackle this problem. In the process of doing so, the class will cover different techniques that are useful in creating and applying neural network models. The class will include assignments, short quizzes, and a final project.

Learning Outcomes

- Be familiar fundamental NLP problems and the methods powering the current state-of-the-art in language technologies, such as large pre-trained neural language models;
- Understand the limitations of current technologies and be able to make informed decisions about addressing them using advanced machine learning techniques;
- Design, implement, and evaluate a computing-based solution to address a language-related problem using state-of-the-art tools.

Prerequisites

CS 580 (Intro to AI) or CS 584 (Data Mining). You should be proficient in (a) Algorithms and Data Structures and (b) Probability and Statistics (STAT 344) or equivalent. Students should be experienced with writing substantial programs in Python. Please contact the instructor if you have questions about the necessary background.

Class Format

The class will be in-person. As the class aims to provide skills necessary to familiarize the students with, and to do cutting-edge NLP research, the classes and assignments will be at least partially implementation-focused. In general, each class will take the following format:

- *Reading:* Before some lectures, you will be pointed to some reading materials *(see "Reading Materials" in course schedule)* that you should read before coming to class that day.
- *Summary/Elaboration/Questions:* The instructor will summarize the important points of the reading material, elaborate on details that were not included in the reading while fielding any questions. Finally, new material on cutting-edge methods, or a deep look into one salient method will be covered.
- *Demo/Code Walk:* In some classes we will walk through some demonstration code that implements a simple version of the main concepts presented in the reading material.

Grading

There will be no midterm or final exam. Your final grade will be dependent on:
Three homework assignments (40%): In the first weeks of the semester, we will have two programming assignments (each worth 10% of your grade, to be completed independently) to ensure that everybody gets a minimal hands-on experience with building state-of-the-art neural networks for NLP. The third assignment (worth 20% of your grade) will be a longer one, where you will experience the full pipeline of building a dataset and a NLP solution for a problem. These assignments will be useful (if not necessary) for implementing your projects later in the class.

- HW0: Preliminaries and Introduction
- HW1: Implementing a small BERT model
- HW2: Text Classification from Scratch

Bonus questions and points will be available for all homeworks. All homeworks are to be done individually. You should complete the collaboration questions at the end of each homework, to denote whether you received/provided help from/to any classmates.

Quizzes (15%): To make sure everyone learns the core concepts taught in the class, we will have 5 take-home quizzes (each worth 3% of your grade). These will be released on Blackboard, and you will have 5 days to complete them (individually). They will only have a few questions with an answer that will not require more than 4-5 sentences.

Project (45%): The bulk of your grade will be based on a group research project related to the topics we will discuss in class. The groups will be of 2-3 people.
Please check out this webpage for details and requirements for the project. Briefly, the project will consist of the following milestones:

- **Checkpoint 1: Project Proposal and Baseline Implementation (15%):** The first checkpoint involves a proposal of a project topic and a small literature survey regarding this topic. In the survey, explain the task that you would like to tackle in concrete terms, and also cover relevant recent research on the topic. You will also need to include a rough plan towards accomplishing the final project.
  Then you will attempt to reproduce the evaluation numbers of a state-of-the-art baseline model for the task of interest. In other words, you must get the same numbers as the previous paper on the same dataset.
- **Checkpoint 2: Final Project Report (30%):** The final project work will be expected to resemble a novel research contribution, depending on the project you selected. See here for details.
  Your final project report will also involve a presentation: In the last class (and before the Final Report due date), you will present your final project in the class. Requirements on the presentation will be provided by the instructor.

Letter Grade	Points (out of 100)
A	94-100
A-	90-93
B+	86-89
B	83-85
B-	80-82
C+	76-79
C	73-75
C-	70-72
D	60-69
F	0-59

Late Day Policy: In case there are unforeseen circumstances that don't let you turn in your assignments on time, 4 late days *total* over the three assignments will be allowed (late days may not be applied to the project deliverables). Note that the second assignment is harder than the first one, and the third assignment is harder that the other two, so it'd be a good idea to try to save your late days for the later assignments if possible. Assignments that are late beyond the allowed late days will be graded down one half-grade per day late.

Readings

For each topic/class the instructor will provide a list of papers as suggested readings. Students should be able to understand the course content just by following the lecture and by doing the readings. However, the following textbooks serve as good references.

- Jurafsky and Martin, Speech and Language Processing, 3rd edition [online] **(Referred to as "JM")**;
- Jacob Eisenstein, Natural Language Processing [online] **(Referred to as "Eisenstein")**;
- Yoav Goldberg, Neural Network Methods in Natural Language Processing [publisher] [online primer pdf] **(Referred to as "Goldberg-Publisher/Primer")**; Note that the "publisher" version can be downloaded if you use the school VPN.

Tentative Schedule

We will try to cover a lot of ground in the first weeks in order to lay the foundations for the projects, but then we will focus more on specific NLP tasks and Linguistics phenomena.

Date	Topic	Assignment Details	Reading Materials
08/22	Introduction and Class Outline; Intro to Language Modeling and Neural Network (Basics & Feedforward NN)	**All Homeworks Released**	JM Ch7.1-7.4, JM Ch4-5; Eisenstein Ch2; & Goldberg-Primer Ch6.1-6.3 (Feedforward NN); Introduction to Pytorch
08/29	Word Embeddings; Binary/Multiclass Classification; Neural Language Modeling: Recurrent Neural Networks	HW0 due 09/01	JM Ch3, Ch9.1-9.3; Prof. Durrett's lecture note 1 & 2 ;
09/05	Distributional Semantics, and Contextual Representations (ELMo, Self-Attention, BERT)	Quiz1 released, due 09/10	JM Ch6; Goldberg-Publisher Ch10.4; Mikolov et al., 2013a & 2013b; Vaswani et al., 2017 (Transformer); Peters et al., 2018 (ELMo); Devlin et al., 2019 (BERT) An easy-to-read blog post on Transformer language models
09/12	Lang Generation; NN Architectures: Encoder-Decoder/Decoder-only, Attention, and Transformers	HW1 due 09/15	JM Ch11.2-11.5 (Seq2Seq), Ch9.7-9.8 (Transformer); Bahdanau et al. 2015 (attention); Vaswani et al., 2017 (Transformer); An easy-to-read blog post on attention

09/19	Transfer Learning; Large Language Models (LLMs) & Scaling Laws	Quiz2 released, due 09/24	Exploring the Limits of Transfer Learning... (Raffel et al., JMLR 2020, "T5"); Nucleus sampling (Holtzmann et al., ICLR 2020); Scaling Laws for Neural Language Models (Kaplan et al., 2020); Training Compute-Optimal Large Language Models (Hoffmann et al., 2022); PaLM: Scaling Language Modeling with Pathways (Chowdhery et al., 2022)
09/26	LLM Advancement: Prompt-based Learning, Tool Augmentation, and More	Quiz3 released, due 10/01	Pre-train, Prompt, and Predict... (Liu et al., 2021, survey paper); Prefix tuning... Prompts for Generation (Li & Liang, ACL 2021); Chain-of-Thought Prompting (Wei et al., 2022); Chameleon Plug-and-Play (Lu et al., 2023)
10/03	NLP Interpretability and Explainability	HW2 due 10/06	LIME; e-SNLI; Hewitt&Liang'19; CheckList
10/10	**NO CLASS (Fall Break; rescheduled for Monday class meeting)**		
10/17	Sequence Labeling and Syntactic Parsing	Quiz4 released, due 10/22	JM Ch8; JM Ch12.1-12.2, 12.6, 13.1-13.4 (constituency), 14 (dependency); Chen&Manning, 2014; Dozat&Manning, 2017
10/24	Semantic Parsing	Project Checkpoint 1 due 10/27	Eisenstein Ch12-13; Zettlemoyer&Collins, 2005; Berant et al., 2013; Dong&Lapata, 2016; Spider (Yu et al., 2018); UnifiedSKG (Xie et al., 2022)
10/31	Machine Translation and Multilingual NLP		Eisenstein 18.1-18.2 Neural Machine Translation... with Subword Units (Sennrich et al., ACL 2016); Beyond English-centric multilingual machine translation (Fan et al., 2020); MAD-X: Multi-task cross lingual transfer (Pfeiffer et al., EMNLP 2020); Hershcovich et al., 2022; Liu et al., 2021; Bird, 2020; Lent et al., 2021;
11/07	Language Generation (Dialog, Summarization, etc.)	Quiz5 released, due 11/12	Holtzman et al., 2020; Ranzato et al., 2016; Maynez et al., 2020; Sellam et al., 2020; See et al., 2017
11/14	Question Answering and Dataset Biases		JM Ch23; ACL20 tutorial QA over text: Chen et al., 2017 (DrQA); Lee et al., 2019 (ORQA); Zhu et al., 2021 (survey); QA over structured data:Pasupat&Liang, 2015 (Table QA); Yih et al., 2015 (KBQA); Rajpurkar et al., 2016
11/21	Human-AI Interaction and Ethics; Wrap-up		Interactivity: Wang et al., 2016; Hancock et al., 2019; guidelines for human-AI interaction; InstructGPT&RLHF; Ethics: Zhao et al., 2017; Rudinger et al., 2018; Gebru et al., 2018
11/28	Project Checkpoint 2 Presentation	Final Project Report due 12/03	

Honor Code

The class enforces the GMU Honor Code, and the more specific honor code policy special to the Department of Computer Science. You will be expected to adhere to this code and policy.

Note to Students

Take care of yourself! As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, global pandemics, feeling down, difficulty concentrating and/or lack of motivation. All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of having a healthy life is learning how to ask for help. Asking for support sooner rather than later is almost always helpful. GMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: https://caps.gmu.edu/. Support is always available (24/7) from Counseling and Psychological Services: 703-527-4077.

Disabilities

If you have a documented learning disability or other condition which may affect academic performance, make sure this documentation is on file with the Office of Disability Services and come talk to me about accommodations. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Services, I encourage you to contact them at ods@gmu.edu.

NEXT
CS 678 Course Project

Last updated on Jan 1, 0001