Syllabus

CS 678

Advanced Natural Language Processing

Choose your Section

Instructors

Antonios Anastasopoulos (antonis [at] gmu [dot] edu)
Office Hours: TBD. ENGR 4412. Email for additional appointments.

TA

TBD
Office Hours: TBD

Meets

Thursday, 4:30 to 7:10 PM, Innovation Hall 105.

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-spring23/.
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 neural network models. The class will include assignments culminating in 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.
- *Presentations:* In the advanced topics classes (second part of the semester), students will be tasked with presenting some seminal papers.

Grading

There will be no midterm or final exam. Your final grade will be dependent on:
Two initial homework assignments (20%): 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. This experience will be useful (if not necessary) for implementing your project later in the class.

- HW1: Text Classification with Neural Nets

- HW2: Implementing a small BERT model

Paper Presentation and Summary (15%): As part of the advanced topics discussed in the second part of the class (weeks 8-14), the lecture will include student presentations of selected papers. In each class, there will be 4 paper presentations, each in 15 minutes. Each paper could be presented by a group of 1-2 students. The rest of the class will have to fill in summary questions on the presentation (just to ensure everybody gains from this exercise; see advice on writing a paper summary).

Project (65%): 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 2-4 people.
Please check out this webpage for requirements on the project as well as suggested topics and resources. Briefly, the project will consist of the following milestones:

- **Group Formation and Project Interest Survey (0%):** We will discuss the specifics in class, but you'll have to form a group with your classmates and fill-in an online form declaring your interest in some of the available projects. The list of the available projects is here. The instructors put a lot of thought into crafting these project ideas. We highly recommend choosing one of the available projects, but if you have a specific idea that you want to explore, you are welcome to discuss this with the instructors.
- **Checkpoint 1: Project Proposal and Literature Survey (10%):** This checkpoint involves a proposal of a project topic and a literature survey regarding this topic. In the survey, explain the task that you would like to tackle in concrete terms, and also cover all of the relevant recent research on the topic. You will also need to include a rough plan towards accomplishing the final project.
- **Checkpoint 2: Project SOTA Baseline Implementation (15%):** Checkpoint 2 will involve reproducing the evaluation numbers of a state-of-the-art baseline model for the task of interest with code that you have implemented (mostly from scratch, dependent on the project). In other words, you must get the same numbers as the previous paper on the same dataset. In addition, students are highly encouraged to revise their proposal based on the results from SOTA baselines. Note that refining your proposed ideas will help you get a better grade in Checkpoint 3.
- **Checkpoint 3: Final Project Report (40%):** The final project work will be expected to be a novel research contribution, depending on the project you selected. Your final project report will also involve two presentations:
    - Final Project 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.
    - Final Project Poster: We will also have a separate poster session, where students from both sections will present posters on their research projects. More details and requirements 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 assignment on time, 3 late days *total* over the two programming 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, so it'd be a good idea to try to save your late days for the second assignment 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. **Each section follows a slightly different schedule, so make sure you are viewing the correct section!**

Date	Topic	Assignment Details	Reading Materials
08/23	Introduction and Class Outline; Projects and Paper Presentations; Neural Network Basics	HW 1 available	JM Ch7.1-7.4 & Goldberg-Primer Ch6.1-6.3 (Feedforward NN); Introduction to Pytorch
08/30	Word Embeddings; Binary/Multiclass Classification; NN Basics (continued: FFNN)		JM Ch4-5; Eisenstein Ch2; Prof. Durrett's lecture note 1 & 2
09/06	Language Modeling: n-gram models and Recurrent Neural Networks	HW 1 due 9/9	JM Ch3, Ch9.1-9.3

09/13	Distributional Semantics, and Contextual Representations (ELMo, self-attention, BERT)	HW 2 available Project Signup Due 9/16	JM Ch6; Goldberg-Publisher Ch10.4; Mikolov et al., 2013a & 2013b; Peters et al., 2018 (ELMo); Devlin et al., 2019 (BERT)
09/20	Lang Generation; NN Architectures: Encoder-Decoder, Attention, and Transformers		JM Ch11.2-11.5 (Seq2Seq), Ch9.7-9.8 (Transformer); Bahdanau et al. 2015 (attention); Vaswani et al., 2017 (Transformer);
09/27	Experimental Design; Interpreting and debugging NLP models	HW 2 due 9/30	
10/04	Sequence Labeling: HMM & CRF, Syntactic Parsing 1	Project Proposal due 10/7	JM Ch8; JM Ch12.1-12.2, 12.6, 13.1-13.4 (constituency); Chen&Manning, 2014; Dozat&Manning, 2017;
10/11	**NO CLASS**		
10/18	Morphology and Syntactic Parsing 2		JM Ch14 (dependency)
10/25	Semantic Parsing		Eisenstein Ch12-13; Zettlemoyer&Collins, 2005; Berant et al., 2013; Dong&Lapata, 2016
11/01	Machine Translation	Project Baseline Reproduction due 11/4	Eisenstein 18.1-18.2 & more (TBD)
11/08	Language Generation (Dialog, Summarization, etc.)		Holtzman et al., 2020; Ranzato et al., 2016; Maynez et al., 2020; Sellam et al., 2020; See et al., 2017
11/15	Human-centered NLP: Multilinguality, Ethics, and Interactivity		Multilinguality: Hershcovich et al., 2022, Liu et al., 2021, Bird, 2020, Lent et al., 2021; Ethics: Zhao et al., 2017, Rudinger et al., 2018, Gebru et al., 2018; Interactivity: Wang et al., 2016, Hancock et al., 2019, guidelines for human-AI interaction
11/22	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
11/29	Final Project Presentations	Final Project Report due 12/09	

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 Project

Last updated on Nov 1, 2022