Advanced NLP: Multilingual methods [601.764]

Kenton Murray

Malone 228

Spring 2023

Contents

1 Overview

This is a project based course focusing on the design and implementation of systems that scale Natural Language Processing methods beyond English. The course will cover both multilingual and cross-lingual methods with an emphasis on zero-shot and few-shot approaches, as well as 'silver' dataset creation. Modules will include Cross-Lingual Information Extraction & Semantics, Cross-Language Information Retrieval, Multilingual Question Answering, Multilingual Structured Prediction, Multilingual Automatic Speech Recognition, as well as other non-English centric NLP methods. Students will be expected to work in small groups and pick from one of the modules to create a model based on state-of-the-art methods covered in the class. The course will be roughly two-thirds lecture based and one-third students presenting project updates periodically throughout the semester. [Applications]

2 Instructor Information

- Instructor Name: Kenton Murray
- Contact Information: kenton@jhu.edu
- Office Hours: Thursdays 3:00-4:00 or by Appointment
- Website: http://www.mt-class.org/jhu-multilingual

3 Grading

- Final Project 50%
  - Final Write-up 20%
  - Final Presentation 10%
  - Mid-Project Presentation 10%
  - Mid-Project Draft Report 5%
  - Project Proposal 5%
- Four Homework assignments 30%
- Class Participation 20%

3.1 Late Policy

All assignments are due at the start of class on the day they are due. Exceptions regarding tardiness will be considered on a case-by-case basis and must be submitted via e-mail 24 hours before the initial due date. Late submissions will be accepted using an exponential decay formula with a half-life of 1 week (604800 seconds).

$$LateGrade = Grade * 0.5^{\frac{seconds}{604800}}$$

3.2 Academic Integrity

The instructor has a very strict academic integrity policy. Students are expected to cite all sources used in homework, assignments, projects, presentations, etc. using standard ACL citation formats. If you have any questions, please ask the instructor for further clarification.

All homework assignments are individual. While students are allowed to discuss problems, all coding, write-ups, and solutions must be the students own work. Students are NOT ALLOWED to show their solutions to problems, nor their coding solutions to each other.

4 Readings

This is an upper-level graduate course. Readings will be provided by the instructor when necessary and may include technical reports, book chapter excerpts, journal and conference papers, etc. There is no required text book for the course.

5 Method of Instruction

The course will primarily be lecture and discussion based. Students will be expected to participate in active discussions of the course material. However, the majority of the instruction will be lectures and outside presentations. Students will be presenting project proposals, mid-point status updates, and final presentations.

6 Attendance Policy

This is a graduate level course with the understanding that many students have research and other obligations. However, much of this course is discussion and project based so students are expected to attend. No formal attendance will be kept, but class participation is a portion of the grade.

7 COVID-19 and Health Policies

We will adhere to University COVID-19 and other health policies. If you are unable to attend a class due to an excused health reason, please let the instructor know at least 30 minutes before the class in order to allow for accommodations to be made regarding video recordings.

Date	Topic	Assignments
January 24	Introduction and MT	
January 26	MT & Large Language Models	
January 31	Survey of Multilingaul NLP	Homework 1 Released
February 2	Cross-Lingual Information Extraction	
February 7	Project Proposal Presentations	
February 9	Project Proposal Presentations	
February 14	CLIR	
February 16	CLIR	Homework 1 Due
February 21	Multilingual QA	Homework 2 Released
February 23	Multilingual QA	
February 28	Typology & Phonology	
March 2	Low-Resource	
March 7	Low-Resource	Homework 2 Due
March 9	Code-Switching	Homework 3 Released
March 14	Code-Switching	
March 16		
March 28	Mid-Point Presentations	Mid-Point Report Due
March 30	Mid-Point Presentations	
April 4	Cross-Lingual Transfer	Homework 3 Due
April 6	Representation Learning	Homework 4 Released
April 11	Interpretability	
April 13	Datasets	
April 18	Datasets	
April 20	Scale & Efficiency	
April 25	Multilingual Structured Prediction	
April 27	Cross-Lingual Semantics	Homework 4 Due
May 2		
May 4	Multilingual ASR	
May 9	Multilingual ASR	
May 11	Multilingual Spoken Language Understanding	
May 16	Final Project Presentations	
May 18	Final Project Presentations	
May 23		Final Project Report Due (1:30 PM)

Table 1: Course Schedule. Orange fields are deliverables at the start of class. Cyan colors represent presentations that the class will be doing. Attend ready to present at the start of the first class during those weeks. Presentation order will be announced at the start of class.

8 Schedule

Note that the instructor reserves the right to change the schedule as seen fit to improve pedagogical outcomes. See Table 1 for exact dates.