# CSE5249: Human-Machine Collaborative Problem Solving in Data Mining, Database, and AI

Spring 2018

### Instructor:

Huan Sun (sun.397@osu.edu)

### Address:

**Baker Systems 260**

### Time:

**Monday 12:50-2:40PM**

### Office Hours:

By appointment

## Course Description

In many scenarios, what current automated frameworks can achieve is still very limited, which necessitates involving humans in the loop. In this course, we will explore human-machine collaborative frameworks for many tasks including question answering, and information extraction. We will also explore machine-machine collaboration mechanisms, where multiple agents are jointly learnt in an adversarial setting.

## Topics Covered

Topics include:

1. Interactive question answering
2. Information extraction
3. Domain adversarial adaptation
4. Dialogue
5. Dual learning
6. Reinforcement learning
7. Natural language interface

## Prerequisites

While there are no formal prerequisites for the class, preliminary knowledge about applied machine learning and text analysis will be needed.

## Course Format

The goal of this course is to familiarize students with research topics related to question answering, text mining, and knowledge bases, and improve their ability of critical thinking. The course activities include (1) reading, presenting, and discussing papers and (2) conducting research-oriented projects. More details are listed as follows:

- Paper reading (10%): The class will read and discuss one paper each week. Before class, each student should read the assigned papers and write a short review (about one page). These reviews should at least cover the following aspects: (a) A summary about what a paper does; (b) Positive comments on the paper and what you have learnt (which could potentially benefit your own research); (c) What the paper did not do well and **propose your solutions**. It is encouraged to prepare questions about the papers for discussion in class. The purpose of writing reviews is to stimulate active thinking and spark discussion in class. **Reviews should be submitted in Carmen by 11:59PM on Sunday (the day before class)**. Please email your reviews to the instructor if there are any technical issues with submission.

- Paper presentation (30%): Each student is expected to present 2 papers during the entire period of the course. One does not need to write a review for a paper if leading the discussion. In each class, we will discuss **only one paper** but two students should work together to give a thorough presentation. Each presentation will take up 70 minutes covering the background, problem definitions, ideas, models, experiments, results, related work, as well as **extensions** such as more details about an algorithm, some cited or follow-up work, other related topics etc. The two students should discuss to divide the workload roughly 50/50. Additionally, we will have 30-35 minutes for in-depth discussion about the papers, which can be in the middle of or after the paper presentation. Students to present need to send the instructor their slides for review **24 hours before class (Please do start preparation early)**.

- Paper/idea discussion (10%): Students are strongly encouraged to think actively. Please bring thoughtful questions and creative ideas in related topics for discussion. We will also discuss on-going class projects. Please do actively participate in discussion to get full credit.

- Projects (50%): Students need to conduct a research-oriented project either by group (2 students per group) or independently. **By the end of**

**the third week**, project groups should be formed. As the semester goes on, each team gradually formulates a research project under the instructor's guidance. In the middle of the semester, students need to present and submit a project proposal. By the end of the semester, each team should present the project progress to the class and submit a final report.

# Reading List and Schedule

- The tentative schedule or reading list is here [https://docs.google.com/spreadsheets/d/1ec0rGzi5FUAXVhsHgrLidy3QILxxLHIi6HcqjgQyDdY/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1ec0rGzi5FUAXVhsHgrLidy3QILxxLHIi6HcqjgQyDdY/edit?usp=sharing)

- The list will be finalized by the end of the first week, but students could now browse the current list to get a general sense of what papers they are to read.

- Each student needs to sign up to present at least 2 papers throughout the semester. Two students work together to do the presentation. The instructor will notify students for sign-up once the reading list is fixed.

# Resources

- Stack Exchange Data Dump: [https://archive.org/details/stackexchange](https://archive.org/details/stackexchange)

- Yahoo Answers and Wikipedia Data dump, as used in [http://times.cs.uiuc.edu/czhai/pub/autoanswer.pdf](http://times.cs.uiuc.edu/czhai/pub/autoanswer.pdf)

- Semi-structured tables from the Web, as mined in [http://www2016.net/proceedings/companion/p75.pdf](http://www2016.net/proceedings/companion/p75.pdf)

- Freebase-based Question answering datasets: [http://www-nlp.stanford.edu/software/sempre/](http://www-nlp.stanford.edu/software/sempre/) and [https://github.com/percyliang/sempre](https://github.com/percyliang/sempre)

- To get one version of Freebase, also check [https://github.com/percyliang/sempre](https://github.com/percyliang/sempre)

- Tools to build your own neural networks: [https://www.tensorflow.org/](https://www.tensorflow.org/)

- Product reviews and related questions: [http://jmcauley.ucsd.edu/data/amazon/qa/](http://jmcauley.ucsd.edu/data/amazon/qa/)

# Project Ideas

- For students who are just taking this course to gain more knowledge or to get credits, they are very welcome to propose their own project idea.

- For students who are thinking about doing research in the instructor's group later on, or working in this area for their Ph.D., they are strongly recommended to discuss with the instructor on possible course projects.