Projects

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Timeline

- Only two deliverables:
  - project proposal: 3+ pages, due Monday, 10/13
  - final report/code: 8+ pages, due last day of finals (12/17)
- Almost completely open-ended!
  - All projects must involve natural language data
  - We recommend significant coding, computational experiments, and data analysis for every project
- More precise requirements will be posted as a separate page to the course website.

Project

- Either *build* natural language processing systems, or *apply* them for some task.
- Use or develop a dataset. Report empirical results or analyses with it.
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task

Resources

- Unfortunately, we don't have an API budget - we recommend paying a cloud provider (e.g. GCE) or API provider as a simple way to obtain access
- Some directions that require less resources
  - Evaluation
  - Applications
  - Dataset building
  - Re-implementation
  - Interpretation/Visualization
  - Prompt engineering

Some grading criteria

- Overall Effort
- Amount learned; breadth and/or depth of findings
- Quality of analysis
- Quality of writing
- Novelty
  - SOTA results, or novel research results, totally unnecessary

Sample projects

Taller, Stronger, Sharper:
Probing Comparative Reasoning Abilities of Vision-Language Models

Examining Medical Narratives of Eating Disorder Recovery on Reddit

Replication of TagRec, a Hierarchical Taxonomy Tagging Model

Learning Schematic and Contextual Representations for Text-to-SQL Parsing

Syllamo: Generating Keyword Mnemonics for Vocabulary Acquisition

Formulating a proposal

- What is the ***research question***?
- What's been done before?
- What experiments will you do?
- How will you know whether it worked?
  - If data: held-out accuracy
  - If no data: manual evaluation of system output. Or, annotate new data

Feel free to be ambitious (in fact, we explicitly encourage creative ideas)! Your project doesn't necessarily have to "work" to get a good grade.

NLP Research

- All of the best NLP publications are open access!
  - The ACL Anthology (https://aclanthology.org/) contains papers from many top NLP conferences (e.g., ACL, EMNLP, NAACL) spanning many decades
  - Machine learning conferences (ICLR, NeurIPS, ICML)
  - CoLM (new)
  - Check out arXiv CS-CL (https://arxiv.org/list/cs.CL/recent) for (too many) most recent papers!
  - This is a fast-moving field, so follow NLP researchers on BlueSky (or still Twitter?) for discussion on the latest advances
- Use Google Scholar and Semantic Scholar to search for relevant papers

- Shared tasks with premade annotated data can be an easy way to get started
  - Merely running some models on premade data isn't a great project, by itself
  - CoNLL and SemEval shared tasks
  - Many question-answering and other benchmarks coming out all the time (see, e.g., LLM leaderboards)

Broader ideas

https://2024.aclweb.org/calls/main_conference_papers/#call-for-main-conference-papers

https://colmweb.org/cfp.html

An example proposal

- Introduction / problem statement
- Motivation (why should we care? why is this problem interesting?)
- Literature review (what has prev. been done?)
- Possible datasets
- Evaluation
- Tools and resources
- Project milestones / tentative schedule

Final reports

Here is a sample outline for your final report. There are different possible ways to structure it (for example, if you can, you can weave related work into the other sections), but we suggest you follow this outline if you're new to this type of technical research writing.

- **Abstract:** summarize the main components of your work in *one paragraph* (no more than 5 sentences). What problem are you solving? What is the key to your approach? What results did you achieve? Your abstract should draw the reader in and interest them in reading the rest of your paper to understand the details of your work.
- **Introduction:** explain the problem, motivate it (why is it important?), and briefly describe your approach. State a *research question* that your project seeks to answer: what are you trying to learn from this research project? You may also overview some of the coming results without discussing the details of your method.
- **Related work:** explain what other approaches have been to the problem. Cite specific instances of previous work.
- **Data:** Describe the dataset that you are using.
- **Method:** Describe your approach to handling the problem. This should should include any models you used and any modeling assumptions you made. If you've developed new models for this project, you may even want to split a description/analysis of your models into its own section.
- **Results:** Describe the experiments you ran and identify your baseline method(s). Include the results you achieved with the various methods you are comparing making. This section will probably also include some figures that *succinctly* summarize your results. *Analyze* your results (including your models). If you did exploratory analysis or a significant amount of feature engineering, your analysis may merit its own section. After reading this section (and your dataset and methods), an interested reader should be able to duplicate your experiments and results.
- **Discussion and Future Work:** discuss any implications of your analysis for the problem as a whole, and what are the next steps for future work. Any other concluding remarks should go here.