Final Project – Natural Language Processing

COMP 550, Fall 2024

Overview

You will be required to complete a final project for this course **in teams of three**. This project will allow you to explore a topic in NLP in more depth than is covered by lectures and homework assignments.

You must implement a system that works with natural language data, with the goal of testing some hypothesis about language or language technology. There are many possibilities that would be satisfactory.

The most common project would be one that implements a model of some natural language phenomenon; for example, a model we discussed in class, or one from recent literature. You should then extend or analyze this model in some way. For example, you could relax some simplifying assumptions made in a current model in order to better account for some linguistic construction. Or, you could incorporate an extra source of knowledge into the feature design of a model. Another possibility is to apply an existing model to a new data set (e.g., a new genre of text, or a different language), and compare its performance to previously reported ones on an existing data set. You could also experiment with a pretrained large language model and analyze its competencies in some systematic and principled way.

The above represents several prototypical flows of a conference paper in NLP, but other possibilities exist! You may also propose to collect or annotate a new data set, or conduct an online user study involving language data which you then analyze in some interesting way. However, these are higher-risk options that might involve a significant amount of additional work. Thus, please discuss this with me if you are interested in pursuing a less traditional project.

You must write and submit a report describing your project. There is no minimum number of citations required, but the report should be structured like a typical conference paper:

> **Abstract**: A short overview of the paper.
>
> **Introduction**: What is the motivation behind the project? In general terms, what is the hypothesis that you test and how do you go about doing so?
>
> **Related work**: Summarize previous work in the area that you have found. I don't expect you to give a complete and fully up-to-date survey of related work in the area. Instead, aim to cite and discuss at least 3-5 relevant papers, with a focus on how your work differs from theirs.
>
> **Method**: Describe the model that you implement, the data set that you use, and any other materials. There should be sufficient details that another researcher is able to more or less replicate your experiment with some effort.

Results: Report the results of your experiment, and any general trends that you see. If the evaluation measure used is not a standard one, be sure to define that as well.

Discussion and conclusion: What conclusions can be drawn from your experiments? Was your initial hypothesis verified? What are the limitations of your work, and how could it be extended?

Statement of contributions: Briefly describe each member's contribution. All project members are expected to contribute to the project. While each member may perform different tasks, you should strive for an equal distribution of work. In particular, all members of the project should participate in the design of the project, be involved in writing, and be involved in the implementation in some way.

You should cite appropriate and relevant sources. You must hand in the code that you wrote for the project. I also reserve the right to ask for further demonstration or proof that you conducted the experiments as described in your report.

Note: Due to the focus of this class, you will be required to use text or transcribed speech for this project. If you would like to use other kinds of data (audio signals, images, videos, etc.), please schedule an appointment with me, and be prepared to make a case that you are proposing an NLP project that is highly related to the topics covered in this course. For example, generating captions of images or videos would be permissible, whereas recognizing characters or digits from image files would not.

Evaluation

Your project will be evaluated on the following points:

- Depth of content
- Justification of contribution and situation of work in relation to previous work
- Correctness and experimental design
- Quality of report: clarity, conciseness, replicability
- Attribution of sources, data sets, and toolkits

Corpora and sample projects

Several corpora are available, either freely or through McGill, for research purposes. For the ones that have restricted access, you should contact me to see if access is possible. Some of these which may be of interest:

- Penn Treebank (syntactic parses)
- OntoNotes (semantic annotations, including semantic roles, word sense information)
- Gigaword (a large corpus of news articles)
- Wikipedia

Using large corpora such as Gigaword or Wikipedia would offer the opportunity to train more complex models and potentially ask more interesting questions. However, processing large corpora takes time and requires good programming skills. It is by no means necessary for a good grade in the project! You will have to judge your commitments accordingly. You can also sample a subset of a large corpus for training.

You may also decide to do a project on a smaller, focused data set. For example, you could download blog posts on some topic, or scientific articles from the ACL Anthology, or social media data such as Tweets, in order to perform a targeted study or experiment.

Proposal (Recommended)

An initial, one-page (c. 500 words) proposal of your project is due on Nov 8, 2024 on myCourses. This will be a chance to receive feedback on the feasibility of the project or survey paper that you have proposed. You are welcome to see me before or after the proposal for further feedback during office hours. The proposal is not directly part of your marks, but is recommended to get feedback on your ideas. If we deem your proposed project to be inappropriate (too narrow or grand in scale, or inappropriate topic), you may choose to submit a second written proposal to be for further feedback, but this is optional.

Final Submission

Your final project is due on the day of the last lecture (Dec 4, 2024), but you have an automatic extension of two weeks (to Dec 18, 2024). For the final submission, you must use the submission version (i.e., with the ruler) of the ACL 2023 style files, available at https://2023.aclweb.org/calls/style_and_formatting/. You may use any of the LaTeX, Word, or Overleaf templates, but you must submit the report as a .pdf. The length of your report **must be between 4.5 to 5 pages of content**. This includes text, figures, appendices, and equations. References, however, do not count towards the length limit.

Lateness Policy

No late work will be accepted at any stage (proposal, final project after the two-week extension) without a valid medical reason.

Timeline

Now Start forming groups; get organized for how you will collaborate

Nov 8 Proposal due

Dec 4 Final submission due

Dec 18 End of grace period for final submission