

Course Project

CMSC 473/673 — Introduction to Natural Language Processing

Item	Summary	Length	Points
Proposal	Monday October 25th	1-2 pages	30
Progress Report	Monday November 22nd	2-3 pages	70
Final Report and Code	Wednesday December 15th	4-6 pages	100

All due dates are by 11:59 PM Baltimore time, through the submission site
<https://www.csee.umbc.edu/courses/undergraduate/473/f21/submit>.

In this project, you will have the opportunity to explore an aspect of natural language processing or computational linguistics that interests you. You may choose any aspect of NLP or CL; the aspect does not have to be one that we cover in class (though it'll probably be easier if it is).

Because this project should be something that interests you, it is okay for it to overlap with your research interests (if applicable). Therefore, while in general you **must form teams of between two and four people**, solo projects that are heavily drawn on your existing research will be allowed. However, to do a solo project you must explicitly obtain written permission at least two full days prior to the proposal deadline (Monday October 25th). Any teams may be comprised of undergraduates, graduates, or be mixed.

You have a lot of freedom in this project. This is not a research project, but you *must* do some investigation into how other people have tackled the same (or similar) problems, and compare and contrast your approach with past ones: that is, your effort does not exist in a vacuum, and you must acknowledge this and explain how your work fits in with previous/on-going work.

The following lists potential topics. You are not restricted to these.

1. Applying NLP techniques (learned in the class, or those not covered) to an application that interests you;
2. Developing some NLP toolkit (very much a software development project);
3. Developing a new corpus of annotations to study particular linguistic phenomena;
4. Developing a new corpus of annotations to target a downstream NLP task/application;
5. Developing a visualization tool to better understand linguistic and/or task annotations;
6. Reimplementing a prior effort and applying it to a different *type* of data (e.g., reimplement and apply a model originally designed for English newswire to tweets);
7. Reimplementing a prior effort and applying it to a new human language (e.g., apply a system designed for English and apply it to Chinese, MSA (Modern Standard Arabic), or any other language);
8. Formulating, implementing and comparing a novel solution to an existing problem;
9. Formulating, acquiring data for, and applying a sufficient baseline for a novel task.

Projects falling under ideas 3, 4 and 9 require both instructor and IRB approval (we can help).

Feel free to draw from current events. For example, you could develop a project around analyzing news/social media, or news related to COVID-19, or related to some other current event.

Deliverables

This project has three main deliverables:

Progress Proposal A brief proposal is due Monday October 25th. The report should use the ACL style format and be at least one page, not including references; while there is not an upper bound, a two page report is reasonable. Be sure to cite any references. The report should include

- a title,
- authors (members of the group),
- a brief, at most 200 word, abstract,
- a description and motivation of the problem,
- a description of your proposed solution,
- any blocks you've experienced or anticipate experiencing.

Progress Report A progress report is due Monday November 22nd. The report should use the ACL style format and be at least two pages; while there is not an upper bound, any report longer than three (not including references) should be justified with particular questions and confusions (supported by data/results). The update can build off of what you wrote for the proposal: you can use what you wrote for the proposal but extend (or adjust) what is already written and include the **[progress]** components.

- a title,
- authors (members of the group),
- a brief (at most 200 word) abstract,
- a description and motivation of the problem,
- a description of your proposed solution,
- **[progress]** a (brief) description of how your solution fits in with what has been previously done on the problem (or related problems),
- **[progress]** the progress you've made,
- **[progress]** what remains to be done, and
- any new blocks you've experienced (or anticipate experiencing).

Be sure to cite any references.

Final Report and Code A final report, in ACL style, and all of your code, data, and final models are due Wednesday December 15th. If you acquired new data, or added new annotations to existing data, please either include this data or (if the dataset is large) make arrangements with the course instructors ahead of time.

The report should be at least four pages, and no more than six, not including references. In writing the final report, you may reuse what you wrote for the progress report, but be sure to specifically and thoroughly address the **[final]** components. The report should include

- a title,
- authors (members of the group),
- a brief (at most 200 word) abstract,
- a description and motivation of the problem,
- a description of your proposed solution,

- [final] a description (longer than in the progress update) of how your solution fits in with what has been previously done on the problem (or related problems),
- [final] if your project involves experimentation, evidence of proper experimental methodology,
- [final] a clear description and analysis of your results,
- [final] an analysis of the limitations of your work, and
- [final] potential follow-up work, both short- and long-term.

Do not write just to fill up space. While four pages may seem like a lot at first, you can fill it up very quickly.

Code Requirements

Your code should be compatible with standard Linux machines, and please make it easy to run ☺. You may use any external libraries you wish to use. Please do not *assume* access to something hardware dependent like CUDA; if you think you'll need CUDA to complete the project, please consult the instructors to work out what resources you need.

Computational Needs

You may quickly find yourself needing more computation. With proper planning, free computational resources can be obtained. If you think you will need and use additional computational resources, please reach out to the course staff **early**.

Report Requirements

All reports should be in the ACL format. Please use the ACL 2018 style guide; both L^AT_EX and Microsoft Word (docx) versions are on GL:

`/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl18/acl18-latex.zip`

`/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl18/acl18-word.zip`

Be sure to cite appropriately and follow all academic honesty standards. You may include figures (your own, reproductions, or copies of existing figures); be sure to provide appropriate credit for the figures. However, make the figures count: do not include them simply to pad the paper.

Do not consider just “recent” papers; try to find papers from the past 25 years. You may be surprised at the recurrence of ideas.

Where to Start

Google Scholar is an easy way to find linked and cited papers. The ACL Anthology (<http://aclanthology.info/>) archives papers by conferences (e.g., ACL, EACL, NAACL, NAACL), journals (CL, TACL), and workshops by year. The AAAI digital library also offers an extensive listing of AI-based conferences and proceedings. Of particular relevance are the flagship AAAI, ICML (International Conference on Machine Learning), and KDD (Knowledge Discovery and Data Mining) proceedings. Papers from NIPS (Neural Information Processing Systems) often tend to the more theoretical, but with a decided focus on neural networks.

You are welcome and encouraged to come talk with me, either during office hours, over email, or by appointment to discuss topics, advice on finding relevant papers, and the direction of your project.