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).

Teams of up to three people are allowed: that is, you may work alone, with one other person, or with two other people. Any teams may be comprised of undergraduates, graduates, or be mixed.

You have a lot of freedom in this project. The main requirement is that there is some novel experimentation, implementation, analysis, or problem formulation. This could include, but is not limited to: 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); 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); formulate, implement and compare a novel solution to an existing problem; formulate, acquire data for, and apply a sufficient baseline for a novel task.

**Deliverables**

This project has two main deliverables:

**Progress Report** A progress report is due Monday, November 6th by 11:59 AM. Please submit this through the assignment ID `project_update`. You should use the ACL style format (see below). The report should be at least one page, not including references; while there is not an upper bound, a two page report is reasonable. 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,
• a (brief) description of how your solution fits in with what has been previously done on
  the problem (or related problems),
• the progress you’ve made,
• what remains to be done, and
• any blocks you’ve experienced (or anticipate experiencing).

Be sure to cite any references.

**Final Report and Code** A final report and all of your code are due Wednesday, December 20th
by 11:59 AM. Please submit these through the assignment ID project_final. 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. 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,
• 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),
• evidence of proper experimental methodology,
• a clear description and analysis of your results,
• an analysis of the limitations of your work, and
• potential follow-up work, both short- and long-term.

You should use the ACL style format (see below).

In writing the final report, you may reuse what you wrote for the progress report. 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**

Unlike the assignments, your code does *not* have to run on GL itself. However, it should be
compatible with standard Linux machines, and please make it easy to run ©. 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.
Report Requirements

Both reports should be in the ACL format. Please use the ACL 2017 style guide. I have downloaded and made available both \LaTeX and Microsoft Word (docx) versions:

/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl17-latex/acl2017.sty
/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl17-word/acl17-word.docx

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. Another great resource is the ACL Anthology (http://aclanthology.info/) archives papers by conferences (e.g., ACL, EACL, NAACL, NAACL), journals (CL, TACL), and workshops by year. It also offers multiple custom searches.

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 be more theoretical, but with a decided focus on neural networks.

This paper provides you immense leeway to relate current NLP methods to areas or topics you are interested in. Workshops often offer targeted application and interest areas. There are also special interest groups, which you can find on the main ACL Anthology page. For instance, are you interested in NLP for the humanities, Semitic languages, or biomedical applications?

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.

---

1 Paper ids generally have the form \textit{XXYY-ZZZZ}, where \textit{X} is a single letter identifier (P is the main ACL, D is for EMNLP, Q is for TACL, etc.), \textit{YY} are the final two digits of the year (2017 → 17), and \textit{ZZZZ} is a per-proceedings identifier.