COSI137: Final Project
DUE: May 9, 2023; 11:59pm EDT

Instructor: Ben Wellner
wellner@brandeis.edu

Brandeis University — March 28, 2023

Overview

The Final Project is opportunity to explore a research or application topic area that aligns with your interests. Your project topic may be informed by your Annotated Bibliography (see below); however, this need not be the case. You may choose a project topic that extends work carried out as part of the programming assignments; alternatively, it is possible to select a topic that is entirely unrelated to any prior work done as part of the course. It is important to get an early start and to select a topic and set of goals that area realistic within the relatively short time-frame.

Individual or Team

Students are encouraged to work on individual projects, however team projects of up to three members are allowed with prior approval. Approval for team projects must be obtained by **April 6** and a short writeup describing the division of labor included with the project description. This write-up should be e-mailed to wellner@brandeis.edu with all team members copied.

Topic and Project Description

The topic for your Final Project should fall within our working definition for Information Extraction: *extraction of meta-data from unstructured, free text for use within a software application or structured representation*. Many subject areas within NLP fit within this definition, however the focus should be on some kind of *extraction* as opposed to *generation* or text re-writing.

Scope

The rough scope of your project should be comparable to the work done for Assignment #3. Many students run into problems by selecting a project that is too ambitious. A better way to manage the scope and associated risk is to come up with a modest goal for a project along with one or more *stretch goals* to take on if time permits. Typically, a modest set of experiments with an in-depth analysis leads to better outcomes vs. more ambitious development and/or sets of experiments that leaves little time for analysis (and risks failing to complete the core development/experiments)

Datasets

The methods in this course are data-driven. You should ensure any project idea you propose involves developing methods that are based on data-driven methodologies, both for developing the model(s) and for evaluation. Accordingly, datasets are of prime importance; most projects work with a single, sufficiently large, and interesting dataset. In some cases, projects may benefit from analyses across multiple datasets.

Project Types

While most projects will involve contemporary data-driven algorithms in the form of neural networks, this is by no means required. Heuristic or other machine learning or statistical approaches are fine. Projects typically fall into one of two general approaches:

- **Methods focused** projects involve implementing a ML/DL algorithm/model and applying it to a target dataset/-task. In most cases, rather than developing a model from scratch, the work involves extending or refining an existing implementation or building off of a pre-trained model (e.g. BERT). The core work involved is primarily writing code with experiments and receiving secondary emphasis.
- **Empirically focused** projects take off the shelf IE systems or applications and apply them in novel ways to one or more datasets. Typically code must be written to pre-process datasets into various formats necessary to run them through the chosen IE systems and/or evaluation routines. The core work involved is primarly running experiments, evaluations and carrying out error analysis. Secondary work involves data preprocessing and some system adjustments, including modifications to the code and default parameters.
- **Prompt focused** projects will leverage recent advances with LLMs, e.g. ChatGPT and analogues. Projects leveraging LLMs should still be focused on information extraction tasks such as: named entity extraction, relation extraction, co-reference, etc. Fortunately, these models perform well at IE tasks and there are various recent papers exploring prompting strategies. Prompt focused projects should leverage an API and operate over at least a modest-sized dataset - e.g. using ChatGPT/GPT-3 to extract relations on the SemEval 2010 Task 8 test dataset.

Note that it is possible to combine aspects of the above three approaches. For example, a project that aims to develop a relation extractor by finetuning BERT could be compared against zero-shot LLMs. In these cases, when the entire project isn't focused on prompt-based methods, a small "side" analysis could be done with an LLM without an API.

Code

Good coding style is expected. This is a component in your grade; it is something to develop for future academic work or employment and is especially important if you are working with a **team**.

Report

Your project should be written up in the style of an ACL short paper, at most 4 pages long (not including references). You are encouraged to reuse portions of your Annotated Bibliography in you report, including and especially the References. You are strongly encouraged to use LaTex, making use of the ACL template (

`https://www.overleaf.com/latex/templates/acl-2020-proceedings-template/zsrkcwjptpcd`

).

Project Proposal

Your project proposal should consist of two to three paragraphs outlining what you plan to do for the Final Project. You may touch on the following points in your proposal:

- Dataset(s) - which dataset or datasets will be used for the project. Are they labeled/unlabeled?
- Research questions - what are the research questions (or hypotheses)?
- Technical approach and methods - what sort of methods/algorithms/code-bases are you looking to use/develop for the work?
- Evaluation - how will you evaluate the application of your method(s) to your dataset(s)?

Grading

The grade breakdown is as follows:

- Project proposal - **Due April 18, 11:59pm EDT** [15%]
- Code quality and correctness [25%]
- Organization, clarity, comments and method/class documentation [10%]

- Experimental results and write-up [50%]

Note that the empirical results and write-up is worth a full 50% of your grade. Please ensure the results are of high quality, appropriate empirical methodologies were followed and results are presented well in tables, graphs and/or figures.

Some Ideas

The following is a list of ideas/topic-areas that you may find helpful:

- Biomedical Event extraction
- Convolutional vs recurrent architectures for entity extraction
- Heuristic vs. machine learning approaches for semantic parsing
- Neural network approaches for co-reference resolution
- Leveraging prior knowledge for question answering
- New/recent strategies for open information extraction
- The utility of dependency parsing for relation extraction
- Approaches to extracting relations using the NYT-10 dataset
- Distant or weak supervision for relation extraction
- Unsupervised learning for co-reference
- Transformer Attention-based neural networks for entity/relation extraction (or any NLP task)
- Comparisons of relation extraction task definitions and annotation guidelines
- Novel approaches to IE - e.g. using modern QA systems for traditional IE problems

Some links to datasets include:

- SemEval 2010, Task 8
- SemEval 2018, Task 7
- TACRED Dataset: https://nlp.stanford.edu/projects/tacred/
- Various datasets at Figure Eight: www.figure-eight.com
- The NYT-10 dataset: http://iesl.cs.umass.edu/riedel/ecml/
- The FewRel relation classification dataset: https://thunlp.github.io/fewrel.html