COSI-230B
Natural Language Annotation for Machine Learning

Group Project

Getting to know your group

- Find out your common interests in annotation tasks
- Find some times that work for your group members to meet in person outside of class
  - This is very important to ensure a good quality project
  - Find a regular interval for meetings
- See what skills and resources each person brings to the table
  - Knowledge of potential datasets
  - Ideas for initial schema, workflows, etc.
  - Specific working skills (programming, data processing, writing, project management etc.)

Group Contract (Due Wednesday 2/25)

Deliverables: A document that you and your teammates will use to assign responsibility for specific tasks within the larger project.

- Divide by skill: the group discusses what skills they bring to the group, and the work within the tasks is divided up accordingly.
- Everyone takes part
- Please note that ***every*** group member will participate in annotation, so that should not factor into the task assignment.

Draft Annotation Schema (Due March 9)

Deliverables: in-class presentation and write-up on possible topics and goals

- After deciding as a group what the broad theme of your project will be, you should present on your topic, dataset, and goals.
- Students should do a brief literature review to survey previous work on your topic
- The draft should be an intuitive design for tagset and attributes associated to each tag and include a small pilot annotation.
- Your schema does not have to be empirically driven nor completely formalized, though, if you have an annotation tool you've selected, you could operationalize your schema in the form of a configuration of your chosen tool.

MAMA/MATTER cycle (March)

Internally, your group will perform parallel, independent annotation on a subset of your raw corpus (sometimes called a “pilot” annotation effort) according to your schema and initial guidelines.

You’ll measure inter-annotator agreement (IAA), review disagreements, and consider the complexity of the task. You’ll then revise the task as you see fit, and repeat until you achieve satisfactory IAA.

Full Annotation Specification (Due late March)

Deliverables: Formal annotation schema, v1.0 of guidelines, and presentation

The specification document contains the detailed instructions for the annotation task that will be provided to non-group member annotators. This document should be clear, contain relevant examples (both positive and negative), and outline expected exceptions or special cases that your annotators may encounter.

- The schema should be operationalized within an annotation environment so that annotators will be able to label the texts appropriately. (This entails that you may need to assist your annotators in setting up the annotation environment that you choose.)

Full Annotation task (April)

- Your group will not be performing its own annotation task—rather, each group will be given the schema and specification from another group.
- Grading for annotation will be based partly on whether you, as an annotator, meet scheduled deadlines, and partly on how well you followed the instructions given by the other group.

Annotation Report (Due during finals week in May)

Deliverables: in-class presentation, ACL-style paper (4 pages), peer evaluation

- Overview of task goals and annotation specification
- Characterization of your dataset (data distribution, annotation distribution, etc.)
- Difficulties during data collection (solved and unsolved)
  - Discuss possible improvements for future iterations
- Annotation quality
  - Quantitative analysis of annotation reliability and interpretation thereof
  - Qualitative analysis of annotator disagreements
- Machine learning experiment
  - Experimental design
    - Baseline system and baseline features
    - Features engineered from the dataset & annotations
  - Experimental results