Course Project Description
CS 140: Language Annotation for Machine Learning
Spring 2018

Professor: Marie Meteer (mmeteer@cs.brandeis.edu)

In this course, you will gain experience in how to create an actual corpus, as you and your group members will agree on a corpus and research question you would like to explore, create an annotation scheme and task description, and analyze the results of your annotation task. This will be a semester-long project, and your entire grade for the course will be based on it.

Grade Breakdown:

5% - annotation goal, group contracts
5% - task description and corpus selection
10% - annotation schema
10% - task specification
10% - annotation
15% - adjudication, precision and recall calculations
15% - running Machine Learning algorithms
20% - write up and final presentation
10% - peer evaluation

Project specifications:

A. Annotation goal
There are many different topics in linguistics that you and your team can choose to examine over the course of the semester. Below is a list of possible topics—you can choose any one of these or select your own, as long as your task is approved by us.

- sentiment/attitude/stance
- syntactic dependencies
- semantic role or semantic function labeling
- anaphora
- multimodal annotation
- Textual cohesion and coherence
- metonymy
- coercion (GLML)
- narrative structure
- temporal information (TimeML)
- spatial information (ISO-Space)
- presuppositions, implicatures, inferences

When writing up the annotation goal, your team should be able to express what you hope to achieve with the annotation, and how far-reaching the task will be. For
example, if you want to do genre annotation of newspaper articles, how specific will your categories be? Will you do broad categories, such as “world news”, “local news”, “sports”, or will you be more specific: “sports: baseball”, “world news: Europe”?

B. Group Contract
The group contract is a document that you and your teammates will use to assign responsibility for specific tasks within the larger project. These can be divided however you want, as long as everyone in the group feels that the tasks are shared equally. There are a few options for how to divide the work — some ways you might want to consider are:

- Divide by task: each group member gets a number of the tasks outlined in the grade section; equality is determined by grade percentage points
- 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: if you prefer that everyone have a say in all the aspects of the grade, the we recommend outlining roles within the group. For example, one person might be in charge of starting each task, one person will be the timekeeper, and the other will be the editor.

Please note that everyone in the group will be doing annotation, so that should not factor into the task assignment. Once you are all satisfied with the contract, everyone in the group must sign it and it must be turned in to John Vogel. The contract will be useful later in the peer evaluation section. Changes to the contract can be made throughout the semester, but must be submitted (signed) to John.

C. Task Description and Corpus Selection
The task description is a longer version of the annotation goal, where you will begin to define exactly how your annotation task will be done. This is closely related to the corpus selection, which will define the type of documents you will annotate (news articles, movie reviews, selected sentences displaying a phenomenon), the source of your documents (Internet, news archives, another corpus), and the size of the corpus (this will vary by task). The task description should be at least 1 page, typed.

D. Annotation Schema
The schema document is what describes the tags and attributes that will be used for the annotation task. This description is in the form of a DTD-like document that can be input into the annotation environment so the annotators will be able to label the texts appropriately. Specifics for how these documents are created will be provided later in the semester.

E. Task Specification
The specification document contains the detailed instructions for the annotation task that will be provided to annotators. This document must be clear, contain relevant examples, and outline any exceptions or special cases that the annotators might encounter. There will be two versions of this: the initial version that is
provided to the annotators, and the revised version that will be produced after the annotation is completed.

**F. Annotation**
Your group will not be performing its own annotation task—rather, each group will be given the schema and specification for a different group. Grading for annotation will be based partially on whether you as an annotator met the deadlines, and partly on how well you followed the instructions given by the other group.

**G. Adjudication and precision and recall**
Adjudication is the process of creating a gold standard corpus from the documents provided by your annotators. No annotation effort is without mistakes, and part of how an annotation task is evaluated is how closely the annotators agree to each other and to the gold standard once it is completed. Adjudication is done by hand. Your group must write a program that can calculate precision and recall for each annotator when compared against the gold standard. A write-up of common annotator mistakes must be turned in, along with the code and the gold standard.

**H. Train Machine Learning Algorithm**
Once a gold standard has been created, it can be used for testing and training various machine learning algorithms. The algorithms that your group implements will depend largely on your annotation topic, but for most groups this will involve using the NLTK or a similar tool kit, not writing your own machine learning algorithm.

**I. Write-up**
A report of the entire annotation process will be a major part of your grade. The format will be similar to conference articles in the Computational Linguistics field, e.g., acl.sty. You will have the opportunity to turn in a draft for evaluation and review prior to the final deadline.

**J. Presentation**
Each group will present their research findings to the rest of the class. This will occur during the period allotted for final exams. Each group member must present the work that they did—if each member collaborated on all parts of the project, then each member still must give part of the presentation.

**K. Peer Evaluation**
Once the presentations are completed, each of you will be asked to fill out an evaluation form about the participation of the other members of your group. These will be collected immediately by the TA, and taken into account when calculating final grades (so be nice to your group members and do your part!).