Syllabus: COSI 140: Natural Language Annotation for Machine Learning

Spring 2019

1 Logistic Information

- Instructor: Nianwen Xue
  - Website: http://www.cs.brandeis.edu/ xuen
  - Email: xuen@brandeis.edu
  - Office: 253 Volen
  - Office hours: TBD

- Teaching Assistants:
  - Jayeol Chun.
    - Email: jchun@brandeis.edu
    - Office: 110
    - Office Hours: TBD
  - TBD

- Time: 12:30-1:50 pm

- Location: TBD

- Class website: http://latte.brandeis.edu

2 Course Objective

The objective of the course is to provide an in-depth coverage of the data annotation / machine learning paradigm that has dominated the field of Natural Language Processing for over twenty years. The focus is on creating linguistically annotated data, either via a well-established procedure that involves designing linguistically sound annotation guidelines and carefully training annotators in an iterative process, or via crowd sourcing, the use of minimally trained annotators to produce massive amounts of data quickly, an approach that has gained
popularity more recently. Both approaches have their advantages and limitations, and we’ll explore the annotation scenarios that are appropriate for each. The course will also explore modern machine learning algorithms, to the extent that they help understand the constraints they impose on data annotation, namely consistently annotated data in large quantities. More in-depth coverage of the technical details of machine learning in NLP will be covered in courses such as COSI 134 and COSI 114.

3 Course description

The course will start with an overview of the relationship between linguistic annotation and machine learning algorithms. This will be followed by an examination of the characteristics of linguistic data, and a gentle introduction to machine learning for computational linguistics/natural language processing, using established machine learning packages. The bulk of the course will then be devoted to an in-depth exploration of different stages of the annotation process: designing annotation specifications and guidelines, identifying shortcomings in the guidelines via test annotation, revising the guidelines, producing annotated data at scale, and train machine learning algorithms on linguistically annotated data. We will be using a few well-established annotation projects to learn best practices in linguistic annotation, and to learn how these projects are conceptualize and executed. The course will end by exploring alternative approaches to linguistic annotation, most notably crowd-sourcing.

A tentative outline of the course will be as follows, subject to minor adjustments:

- Machine Learning and linguistic data
- Corpus analysis and characteristics of linguistic data
- Machine learning for Computational Linguistics / Natural Language Processing
  - SciKit-Learn tutorial
  - Design sensible features for machine learning
- Annotation tools and procedures
- Annotation evaluation
- Case studies
  - Syntactic annotation and “treebanks”
  - Semantic role labeling/meaning representation
  - Discourse and Dialogue annotation
  - Annotation of subjective language
  - TimeML and temporal dependency structure
- Alternative approaches to linguistic annotation: crowd-sourcing
4 Student responsibilities

This is a “learning-by-doing” course that expects active participation from the students. The students are expected to actively participate in class discussions, work on collaborative and individual projects, and presenting results from their own projects as well as existing linguistic annotation projects.

4.1 Projects

The course will have three projects. The first two projects are individual projects that are designed to provide students with hands-on experience in applying well established machine learning techniques to NLP problems. The goal of working on these projects is to gain some intuitive understanding of how machine learning works, the impact of features and (hyper)parameters on machine learning accuracy, as well as the constraints current machine learning approaches impose on linguistic data annotation. The in-depth coverage of the “under-the-hood” details of machine learning algorithms will be reserved for courses such as COSI 134 and COSI 114.

The third project will be a collaborative annotation project that lasts throughout the semester. The annotation project will unfold in several phases. The first phase is to define a linguistic task that are of interest to members of a collaborative group, locate the appropriate data set, and design the specification and guidelines for the annotation task. The second phase is to perform trial annotation to test the annotation specification, and identify areas where annotation inconsistency may occur, and revise the annotation specification. The third phase is to produce annotated data by drafting your classmates (outside your group) as annotators. The final phase is to train a machine learning model on the data you’ve annotated and write a report on your project.

5 Course reading material

- Selected papers

6 Grading

- Participation: 10%
- Mid-term quiz: 10%
- Two individual projects: 30%
- Collaborative annotation project: 50%
7 Academic Integrity

You should finish homework assignments, exams, and project reports on your own unless a project is explicitly stated as a collaborative project. Late projects are subject to grade deduction.