Citation Needed: Algorithmically Preference Reference Usage

Background: Many kinds of text contain citations to back up claims and connect the statements to prior studies. Even collaboratively-constructed media like Wikipedia uses citations extensively to indicate where original research can be be found. Yet knowing when a citation is needed can be tricky--do all sentences need citations, and, if a citation is needed, which citation should be used? Citations can serve different roles, such as pointing out background material, contrasting the current approach with another, or even linking to the official reference for some data or algorithm. Moreover, multiple works could be suitable in a given context. This is a hard problem, yet we have lots of data both in terms of content citing others and in the text of what is cited, making it feasible we could develop an algorithm for asking these questions.

Project goals: This project aims to develop predictive algorithms that will infer when a citation is needed and which citation should be used. To do this, we'll develop new Natural Language Processing algorithms that work on a variety of different kinds of scholarly or encyclopedic text. We'll examine which linguistic cues signal the need for a citation through both computational and manual analysis. As a part of building the system, we'll use machine learning techniques and possibly deep learning techniques to build our citation-predicting system. A stretch goal is to build a browser plugin that helps users insert citations in their own text as they write.

Student Role: A student will participate in a new research project focused on citation prediction and be the primary research developer for the software that drives the project. The student should have a strong programming background in python, familiarity working in a UNIX environment, and some experience with machine learning is a plus. As a part of the project, the student will help write up results in both a scientific publication format and for a general audience as a blog. Specifically, a student will:

- Develop new machine learning models to predict citation behavior
- Assist in performing a research survey and writing up results
- Prepare new data for the research project (initial data is ready to go)
- As time allows, develop a browser plug-in

Mentorship Role: I (David Jurgens) will meet one on one with the student weekly and then we'll also meet weekly with the whole Blablablab group for a general lab meeting. Individual meetings will cover both technical and professional skills needed to be a successful researcher in industry or academia. As a part of the summer project, I'll work with the student on how to design a research study, develop reproducible research, write results for both a scholarly and general audience, and give presentations on their work. I will help guide the student to obtain concrete outcomes they can show off to others to help them get to the next stage of their career.