Summer Internship Report:
Machine learning research on event extraction for knowledge base

This summer I was working as an intern for the NELL (Never Ending Language Learner) project at Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania. The project’s goal is to develop a system to search through text corpus (web pages) for information and store it in a knowledge base which can improve itself through time via both human interaction and self-supervision. There are various components within the project such as category classification, relation classification, entity resolution and inference rules.

In the summer, I was working with Dr. Ndapa Nakashole on the problem of event extraction. Event mention is a format to capture knowledge based upon the happening of a specific event at a specific time or place. My main duty is to explore the potential of extracting event knowledge and integrate it to NELL system. For a more technical description, I extracted event mention from text corpus using sentence-bound semantic dependencies graph and generated relations among entities of the mention using some labeled data and the Expectation Maximization algorithm.

When working for the NELL project, I was able to acquire a lot of good skills. In terms of programming, I needed to write codes in Java on a daily basis and was able to improve my implementation skills greatly. In term of field knowledge, I learned a significant amount of knowledge and practices from natural language processing and machine learning. Chief among them is to enumerate hypotheses and then use
Bayesian rules and provided data to get the likeliest probabilistic distribution among hypotheses. In term of research experience, thanks to constant questioning process from Dr. Nakashole, I learned to add depth to my assumption, to be more rigorous in my reasoning process, and to communicate with greater clarity. In addition, I was also able to attend talks and seminars and expand my scope of knowledge about the field greatly.

About the high and low points, I think the low is when figuring out that I didn’t know much about machine learning at all during the first weeks of the internship, and the high is when I was able to produce good results for the extraction step, which validated some of my assumptions and gave me some relief that I was able to have some good results. The main lesson I take after the internship is that there are a lot of things to learn about and that it was exciting to be able to work in the field of my interest.

I think the internship is a great experience for me. I had a lot of fun and I really hope to continue working in the field (machine learning, natural language processing) by going to a graduate program and working in a research program in the future.