CS 410 Project Progress Report

Team GYZ: Yuhan Guo(yuhang4), Qi Zeng(qizeng2), Haoduo Yan(haoduoy2)

1) Which tasks have been completed?

We separate our tasks into three parts: framework, data, and functionality. Each group member will be responsible for one part.

We have finished the user interface design and created the basic structure and files for chrome extension implementation. In addition, we have also implemented the I/O and functionality design(see below).

2) Which tasks are pending?

Framework:

Chrome extension implementation

Data:

Web scraping

Data cleaning

Functionality:

Named Entity Recognition: Given a sentence, return the identified entity lists (position and type tuple). This function will be implemented by integrating bert-base-NER (https://huggingface.co/dslim/bert-base-NER), a fine-tuned BERT model that is ready to use for Named Entity Recognition and achieves state-of-the-art performance for the NER task.

Entity Linking: Given an entity, return its unique identity in Wikipedia if it existed. This function will be mainly based on matching.

Entity Extension: Given a linked entity with a wiki identifier, return related entities. This function serves as the core function for this extension since it can return extra knowledge. We plan to explore co-occurring entities (entities shown in the same wiki page) and entities with high relevance (ranked with semantics-based ranking methods).

3) Are you facing any challenges?

We are all new to chrome extension implementation and have little experience in JavaScript. The design of the chrome extension is one challenge for us, but we have found some useful tutorials online which can be used as guidelines.

Another challenge is the web scraping for a passage. We want to convert the passage into a list of words. We are working on extracting critical information and annotating multiple entities within a text, and the main difficulty is to label various entities with accuracy.