The scientific literature is vast and increasing at a rapid rate (Saggion and Ronzano, 2017). Computers can help efficiently process this information to extract important relationships. While programmed to parse artificial language (e.g., a compiler converting high-level language to machine-code), computers must be programmed to parse natural language (a language that evolves without conscious planning) to assist in this task. The use of computers to extract information into useful data is known as Natural Language Processing (NLP). One application of NLP is information extraction from the biomedical literature. To engender public trust in research findings, effectively contribute to scientific knowledge, and to generate effective treatments for disease, it is essential to accurately extract information.

Named Entity Recognition (NER) is the process of finding entities (a discrete physical thing, usually proper nouns in the case of named entities) in a text and classifying them. Two attributes of natural language tend to lend difficulty to this task: ambiguity and polysemy. Polysemy (“multiple signs”) is the use of many words of phrases to mean the same concept. Ambiguity is the ability of the same phrase to mean different concepts. An ontological term is a standardized alphanumeric codes that represents the concept a set of words or phrases mean. Ontologies can mitigate polysemy. However, ambiguity can make it difficult to correctly assign a string of letters to its correct ontology. One of the measures of a method’s ability to cope with polysemy is the false positive rate, which can be scored using the F1 measure (correct positive results/all positive results).

In sequence-to-sequence, the process of assigning terms to the proper ontology is analogous to a translation task: transforming the sequence of characters from the natural language to the sequence of characters forming the proper ontological ID. One potential advantage is sequence-to-sequence may allow training to recognize shared sub-word patterns (character n-grams) in members of the same word family, thus enabling recognition of out-of-vocabulary words that match the ontology. This project continues the work of Negacy Hailu who found sequence-to-sequence achieved an F1 score of 0.7. The present project aims to (1) evaluate sequence-to-sequence performance against existing methods in terms of false positive rates and (2) determine whether sequence-to-sequence learns to detect word family members in training.

The data set for training and testing in this project is the Colorado Richly Annotated Full-Text (CRAFT) corpus, a curated collection of 67 full-text biomedical articles. These articles have been manually annotated for ten different Open Biomedical Ontologies (OBOs). The first task in processing this data would be to train an algorithm on is span detection. The second task is concept normalization which is achieved by directly converting the sequence identified as an entity into its ontology term. To validate the sequence-to-sequence method, I will use CRAFT as a training and testing dataset with conventional NER methods such as [to be determined] and compare the F1 scores for different ontologies.

Saggion, H., & Ronzano, F. (2017, June). Scholarly data mining: making sense of scientific literature. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1-2). IEEE.