2024

Applying a Named Entity Recognition Model to Chemical Patents

WanXiang Chen
University at Buffalo

Zachery Nolan
University of Memphis

Christina Tang
Virginia Commonwealth University

Bridget T. McInnes
Virginia Commonwealth University

Applying a Named Entity Recognition Model to Chemical Patents

Team Members: WanXiang Chen, Zachery Nolan, Scott Taylor | **Faculty Advisor:** Dr. Bridget McInnes

Problem

Background

- Named Entity Recognition (NER) is an application of Natural Language Processing (NLP) that involves recording entities contained in a text excerpt and classifying them into predefined category types
- Unstructured text, or text without a predefined format is converted into structured text that can serve as input to an NER model

Purpose

- To use a pretrained NER model to extract and classify chemical entities from patents
- This streamlines the process of determining the number and types of chemical agents used in a specific patent
- This information may be useful for researchers, pharmaceutical companies, and agencies such as the FDA

Experimental Methods

- A Python program was created to extract 1,000 chemical patents from a Google Patents API
- Each patent extraction was placed in its own folder that was named with its patent ID
- Each folder contained JSON data, all images, and all text extracted into a .txt text file
- Then we did some preprocessing on the patents by using the NLTK library to first tokenize the sentences, then tokenize the words of each sentence, and finally add them all into one Pandas DataFrame and create a pickle file from that
- After that we took an open-source dataset called ChEMU and used that to train our NER model

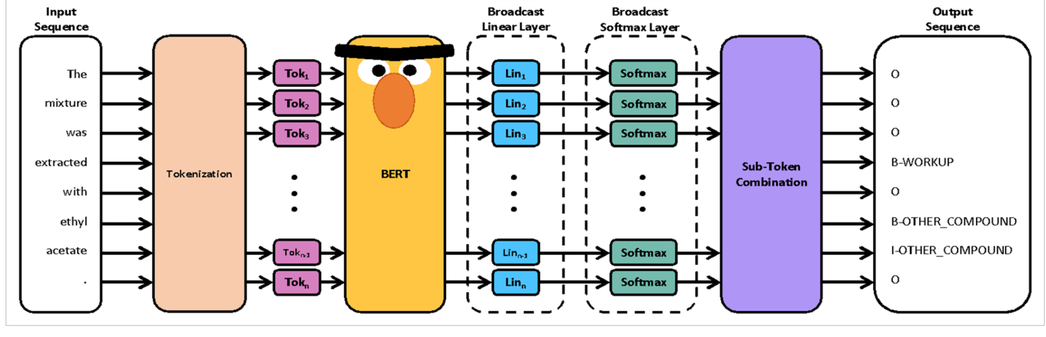
The Process

- This is one of the text files containing text extracted from the patents using the Beautiful Soup library

Experimental Steps

- A user queried the Google Patents API using our Python program
- The program retrieved the requested patents and saves the data to the current directory of the user's personal device
- All the patent text files got tokenized and preprocessed into a single pickle file so that it can be put in the NER model after it is trained
- Preprocessed the ChEMU dataset by converting it from CoNLL format into pickle files and then used those pickle files to train our NER model
- Used the trained NER model on the patents pickle file and have it isolate the relevant chemical entities
- The model labels chemicals as 'CHEM' and everything else as 'O'

The Model



Results

(Using the Test Data From ChEMU)

Test Metric	**Baseline**
test_avg_f1	0.9960
test_f1_class_O	0.9966
test_f1_class_CHEM	0.9953
test_f1_micro	0.9960
test_precision_class_O	0.9989
test_precision_class_CHEM	0.9920
test_recall_class_O	0.9943
test_recall_class_CHEM	0.9985

Future Work

- After that, we will do some preliminary manual checking on the patent results for precision
- We need a chemist to look over all the Google patents text files and label them in order to get a ground truth
- Then we will run our trained model on the labeled patents and get our final results