

Virginia Commonwealth University VCU Scholars Compass

Summer REU Program

College of Engineering

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

Follow this and additional works at: https://scholarscompass.vcu.edu/reu

Part of the Engineering Commons

© The Author(s)

Downloaded from

https://scholarscompass.vcu.edu/reu/10

This Book is brought to you for free and open access by the College of Engineering at VCU Scholars Compass. It has been accepted for inclusion in Summer REU Program by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

UNDERGRADUATE **RESEARCH SYMPOSIUM**



Applying a Named Entity Recognition Model to Chemical Patents

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

Computer Science

Edit View Search Tools Documents Help

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

The Process

In one aspect, the present invention improves the resealability and the convenience of use of the package, while minimizing costs and manufacturing was

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



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

Test Me

test_avg_

test_f1_c

test_f1_c

test_f1_n

test_prec

test_prec

test_reca

test_reca

- ground truth





NLP@VCU

Results	
sing the Test Data From ChEMU)	
etric	Baseline
_f1	0.9960
lass_O	0.9966
lass_CHEM	0.9953
nicro	0.9960
cision_class_O	0.9989
cision_class_CHEM	0.9920
all_class_O	0.9943
all_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

• Then we will run our trained model on the labeled patents and get our final results

College of Engineering