Natural Language Processing Approach to Extract Compound Information from PubChem

Rehan Khan*, Preenon Bagchi, Krutanjali Patil

Institute of Biosciences and Technology, MGM University, Chht.

Sambhajinagar, India

khanrehan9395@gmail.com

Abstract. PubChem is one of the largest and most comprehensive databases of its kind, containing information on millions of chemical compounds, including their structures, properties, and biological activities. The Natural Language Processing (NLP) approach to extract compound information from PubChem has several advantages, including improved accuracy and efficiency compared to manual methods, and the ability to handle large amounts of data. NLP plays a significant role in extracting compound information from PubChem by enabling the processing of unstructured and semi-structured text data, and by allowing for the identification of chemical compound names and the extraction of relevant information from text data. Simplified Molecular Input Line Entry Specification (SMILES) representations are also used in computational chemistry and drug discovery, where they can be used to predict properties of compounds, such as their stability, reactivity, and toxicity. This information is then used by researchers to design and optimize new drugs and chemical compounds. In this work we have extracted the compound information from PubChem using natural language processing can be approached in several steps they are, Define the target information, Data acquisition, Text pre-processing, Named entity recognition, Relation extraction, Entity linking, Output generation. The results of a natural language processing approach to extract compound information from PubChem have the potential to greatly aid research efforts in chemistry, pharmacology, and other related fields.

In conclusion, SMILES representations are a powerful tool for identifying chemical compounds. By representing the structure of a chemical compound as a string of characters, SMILES representations make it possible to process and analyze chemical compounds using computers, enabling scientists and researchers to make new discoveries and advancements in the field of chemistry.

Keywords: Natural language processing, SMILES, PubChem, Compounds, Properties, Python

1 Introduction

Natural Language Processing (NLP) is a field of computer science and artificial intelligence that deals with the interaction between computers and human (natural) languages. NLP involves the development of algorithms and models that can process and analyse human language, such as text, speech, or written documents. An unstructured language codified by humans to describe domain-specific knowledge is what text-based representations of chemicals and proteins are.

R. Somashekhar et al. (eds.), *Proceedings of the International Conference on Advances in Nano-Neuro-Bio-Quantum (ICAN 2023)*,
Advances in Intelligent Systems Research 69,
https://doi.org/10.2991/978-94-6463-294-1_6

Advances in NLP methodologies in the processing of spoken languages accelerated the application of NLP to elucidate hidden knowledge in textual representations of these biochemical entities and then use it to construct models to predict molecular properties or to design novel molecules [1]. PubChem is a public database of chemical substances and their properties, maintained by the National Institutes of Health (NIH) in the United States. It is one of the largest and most comprehensive databases of its kind, containing information on millions of chemical compounds, including their structures, properties, and biological activities. Simplified Molecular Input Line Entry Specification (SMILES) representations consist of a series of letters and symbols that describe the arrangement of atoms in a molecule. As far as chemical data is concerned, laboratories can access databases such as PubChem, a repository containing data on more than a hundred million compounds, or Drugbank, which contains data on over a thousand drugs. SMILES is a string representation of a chemical compound's structure, which can be used to uniquely identify a particular molecules [1]. SMILES strings use a simple, American Standard Code for Information Interchange (ASCII)-based syntax to encode the molecular structure, including the type and connectivity of each atom, as well as any bonds between them. For chemicals there are several text-based alternatives such as chemical formula, IUPAC International Chemical Identifier (InChI) [2] and Simplified Molecular Input Line Entry Specification (SMILES) [3]. Due to the growing amount of information contained in public databases such as PubChem [4], ChEMBL [5], and UniProt [6], the "learning" aspect of computational approaches is greatly enhanced in this era of “Big Data”. The role of SMILES representations in identifying compounds is crucial because they provide a compact, machine-readable representation of a molecule's structure that can be easily processed and stored. By encoding the structure of a molecule in a SMILES string, it is possible to search for and compare chemical compounds in a database using computer algorithms, making it easier and more efficient to identify and analyze chemical compounds. Machine learning (ML) and Natural language processing (NLP) fields are driven in part by advances in computing power. Not only are NLP methodologies facilitating processing and exploitation of biochemical text, they also promise an “understanding” of biochemical language to elucidate the underlying principles of bimolecular recognition.

As part of bioinformatics tasks with the ultimate goal of accelerating the discovery of new drugs, understanding the sequences of molecules and proteins is one of the most important steps toward making predictions about their structure and function. Processing and extracting information from textual data written in natural languages is one of the major application areas of NLP methodologies in the biomedical domain (also known as BioNLP) [7] Moreover, due to its textual form, SMILES takes 50% to 70% less ace than other representation methods such as an identical connection table. There are rules for using sp SMILES notation, just as there are rules for a language. Canonical SMILES can provide a unique SMILES representation. However, different databases such as PubChem might use different canonicalization algorithms to generate different unique SMILES.

1.1 Extracting compound information from PubChem is important for several reasons: -

a) Research and Development: PubChem is a vast database of chemical compounds and their properties, making it a valuable resource for scientists and researchers in various fields, including chemistry, pharmacology, and material science. By extracting compound information from PubChem, researchers can gain insights into the structure, properties, and potential uses of different chemicals, which can help them develop new drugs, materials, and other products.

b) Drug Discovery: In the field of pharmacology, extracting compound information from PubChem can help in the discovery of new drugs. By analyzing the properties and structures of different compounds, researchers can identify potential candidates for drug development, which can then be tested for efficacy and safety.

c) Safety and Regulation: Extracting compound information from PubChem can also help regulators and safety agencies monitor and manage the use of hazardous chemicals. By having access to accurate information about the properties of different chemicals, regulators can better assess the potential risks they pose and make informed decisions about their use.

d) Data Analysis: Extracting compound information from PubChem can also be used for data analysis purposes. For example, by extracting information on the properties of different compounds, researchers can analyze trends and patterns in the chemical data, which can lead to new insights and discoveries.

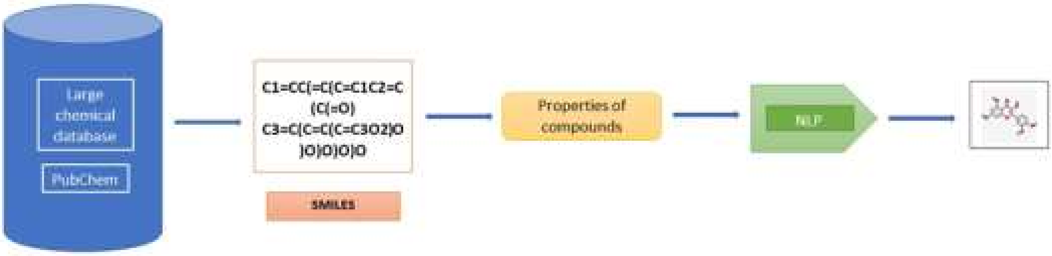

Figure 1 Process of extracting information of a compound using Natural Language Processing.

To extract compound information from PubChem using NLP is to use Named Entity Recognition (NER) to identify the names of chemical compounds in the text data, and then use information retrieval techniques to retrieve the relevant information, such as the SMILES representation and properties, from the PubChem database.

This information can then be used to retrieve additional information from the PubChem database, such as the SMILES representation and molecular structure.

1.2 The steps involved in extracting compound information from PubChem using NLP approach are:

1.Data Collection: Collect the relevant data from PubChem, either by scraping the website or using the PubChem API to retrieve data programmatically.

2.Data Pre-processing: Clean and pre-process the collected data to prepare it for further analysis. This may involve removing irrelevant information, transforming the data into a format suitable for NLP algorithms, and standardizing the data to ensure consistency.

3.Named Entity Recognition (NER): Use Named Entity Recognition (NER) techniques to extract the names of chemical compounds from the text data.

4.Sentiment Analysis: Use sentiment analysis techniques to determine the sentiment expressed in the text, such as whether a particular property is positive or negative.

5.Information Extraction: Use information extraction techniques to extract relevant information about the chemical compounds, such as their properties, relationships, and events.

6.Data Analysis: Analyze the extracted information to draw meaningful insights and conclusions about the chemical compounds. This may involve visualizing the data, identifying patterns, or comparing the properties and relationships of different compounds.

Artificial intelligence techniques may benefit the drug development process in various ways. One active research area is the development of information extraction tools over chemical literature.

Chemical literature contains valuable information about the latest advancements in the chemistry domain that is important to make findable and accessible. However, due to the rapid growth in chemical literature, new discoveries are easily missed while manual extraction of this information is increasingly infeasible [8]. The goal of NLP technologies is to accelerate drug discovery by integrating biological and chemical knowledge.

NLP techniques can be used to extract relevant information from text data, such as chemical formula, properties, and references to other databases.

A. Objectives

The objective of this paper is to demonstrate how Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) to retrieve physical and chemical properties of phytocompounds from PUBCHEM using python-based codes of medicinal plants and phytocompounds information from PubChem containing large number of compounds, saving researchers time and effort.

B. Abbreviations

NLM - National Library of Medicine

NIH - National Institutes of Health

PMC - PubMed Central

NLP - Natural Language Processing

NLTK - Natural Language Tool Kit

NER -Name Entity Recognition

API -Application Programming Interface

SMILES- Simplified Molecular Input Line Entry Specification

AI -Artificial Intelligence

2. Materials and Methodology

In this work we have extracted the compound information from PubChem using natural language processing can be approached in several steps. We are using the python libraries NLTK, PUBCHEMY and PUNKT. Here is the methodology that can be followed:

1. Define the target information: We need to identify the specific information that we need to extract from PubChem. These properties can be molecular formula, molecular weight, melting point and boiling point. Along with this we extract SMILES.
2. Data acquisition: We need to retrieve the necessary data from PubChem. This can be done using PubChem's API or by downloading data in a structured format like CSV, JSON or XML.
3. Text pre-processing: We need to clean and preprocess the text data to remove unnecessary characters, stop words, and other irrelevant information. Use techniques like tokenization, stemming, and lemmatization to convert the text into a standardized format using PUNKT.
4. Named Entity Recognition (NER).: We have to identify the entities in the text that correspond to the target information using named entity recognition (NER). NER involves labeling named entities in the text data, such as the compound name, properties, and identifiers.
5. Relation extraction: We then extracted the relationships between the named entities identified in step 4. Then used the techniques like dependency parsing, semantic role labeling, and rule-based systems to extract the relevant information using PUNKT.
6. Entity linking: Then to resolve any ambiguity in the identified named entities by linking them to their corresponding identifiers in PubChem. This step involves using external resources like ontologies, dictionaries, and databases to map the named entities to their respective PubChem IDs.
7. Output generation: Finally, we generated a output that contains the extracted information in a

standardized format. This output can be used for further analysis or visualization.

Moreover, the methodology involves a combination of techniques from natural language processing, machine learning, and database systems to extract relevant information from PubChem.

3. Results

3.1 Steps performed

1.Identification of the most commonly mentioned compounds in scientific literature was done.

For example, "Aspirin is a common pain reliever with the chemical formula C9H8O4."

2. Extraction of specific properties of a given compound, such as molecular weight, structure, and toxicity were done using PubChem database.

By using information retrieval techniques, the SMILES representation and properties of Aspirin was retrieved from the PubChem database. Using NER, the name of the chemical compound, "Aspirin," was identified.

3. Pubchempy library was installed as it provides a way to interact with in python.

4. NLTK was imported and PUNKT was downloaded for parsing, stemming, tokenizing and other process.

5. Collection of the relevant data from PubChem, either by scraping the website or using the PubChem API to retrieve data programmatically was done.

6. Compound name was extracted.

The results of a Natural Language Processing approach to extract compound information from PubChem have the potential to greatly aid research efforts in chemistry, pharmacology, and other related fields.

3.2 Program Codes

A. Installing and importing of libraries.

```
!pip install pubchempy

import nltk

nltk.download('punkt')
```

B. Codes to get compounds with ID.

```
#importing necessary libraries

import pubchempy as pcp

import nltk

from nltk.tokenize import word_tokenize
```

```
#function to extract compounds from PubChem
def extract_compounds(smiles, properties):

    #tokenize the input strings
    smiles_tokens = word_tokenize(smiles)
    properties_tokens = word_tokenize(properties)

    #loop through the tokens to find compounds
    compounds = []
    for token in smiles_tokens:
        try:
            compound = pcp.get_compounds(token, 'smiles')
            compounds.append(compound)
        except:
            pass
    for token in properties_tokens:
        try:
            compound = pcp.get_compounds(token, 'property')
            compounds.append(compound)
        except:
            pass

    #return the list of  compounds
    return compounds

#example usage
smiles = '"CC(=O)OC1=CC=CC=C1C(=O)O'
properties = 'molecular_weight>180'
compounds = extract_compounds(smiles, properties)
print(compounds)
```

C. To get compounds with names.

```
import requests

def get_compound_name(smiles):
    # Define the PubChem API URL
    url                                                                =
"https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/smiles/{}/property/
IUPACName,MolecularWeight,CanonicalSMILES,InChIKey/JSON".format(smiles)


    # Send a request to the API and get the response
    response = requests.get(url)


    # Check if the request was successful
    if response.status_code == 200:
        # Extract the name of the compound from the response
        data = response.json()
        compound_name                                                  =
data['PropertyTable']['Properties'][0]['IUPACName']
        return compound_name
    else:
        # Return an error message if the request was not successful
        return "Error: Unable to fetch data from PubChem"


# Example usage
smiles = "CC(=O)OC1=CC=CC=C1C(=O)O"
compound_name = get_compound_name(smiles)
print("Compound name:", compound_name)
```

Link containing the codes.

https://github.com/RehanKhan-007/NLP

4. Discussion

The PubChempy library is used for retrieval of compound information using the name of the compound. For retrieving the compound name using compound property, researchers depend on various search engines or research papers and finally retrieve the compound from PubChem. This long process is lessened by applying NLP to PubChempy library. Hence, in this work we have attempted to retrieve the compound using their property.

The approach to extract compound information from PubChem using NLP involves several steps, including data collection, pre-processing, NLP techniques, information extraction, and data analysis. By using NLP, it is possible to efficiently and accurately extract relevant information from the vast amounts of data contained in PubChem, providing valuable insights and knowledge about chemical compounds for researchers and practitioners in the field.

5. Conclusion

In conclusion, NLP plays a significant role in extracting compound information from PubChem by enabling the processing of unstructured and semi-structured text data, and by allowing for the identification of chemical compound names and the extraction of relevant information from text data. The use of SMILES representations in identifying compounds is a crucial aspect of modern chemical and biological research, providing a standardized and machine-readable representation of molecular structure that enables researchers to efficiently and accurately search, compare, and analyze chemical compounds.

References

1. Hakime Öztürk et al. (2020) Exploring chemical space using natural language processing methodologies for drug discovery. Volume 25, Issue 4, 689-705.
2. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev, Inchi-the worldwide chemical structure identifier standard, Journal of cheminformatics 5 (2013) 7.
3. D. Weininger, SMILES, a chemical language and information system.1. introduction to methodology and encoding rules, Journal of chemical information and computer sciences 28 (1988) 31–36.
4. E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, PubChem: integrated platform of small molecules and biological activities, in: Annual reports in computational chemistry, volume 4, Elsevier, 2008, pp. 217–241.
5. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al., Chembl: a large-scale bioactivity database for drug discovery, Nucleic acids research 40 (2011) D1100–D1107.
6. R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al., Uniprot: the universal protein knowledgebase, Nucleic acids research 32 (2004) D115–D119.
7. Ernst, A. Siu, G. Weikum, Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences, BMC bioinformatics 16 (2015) 157.
8. Muresan, S., Petrov, P., Southan, C., Kjellberg, M. J., Kogej, T., Tyrchan, C., et al. (2011). Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today 16, 1019–1030.