Chapter 4
Chemical Named Entity Recognition Using Deep Learning Techniques:
A Review

Hema R.

Department of Computer Science, University of Madras, India

Ajantha Devi

https://orcid.org/0000-0002-9455-4826

AP3 Solutions, India

ABSTRACT

Chemical entities can be represented in different forms like chemical names, chemical formulae, and chemical structures. Because of the different classification frameworks for chemical names, the task of distinguishing proof or extraction of chemical elements with less ambiguous is considered a major test. Compound named entity recognition (NER) is the initial phase in any chemical-related data extraction strategy. The majority of the chemical NER is done utilizing dictionary-based, rule-based, and machine learning procedures. Recently, deep learning methods have evolved, and, in this chapter, the authors sketch out the various deep learning techniques applied for chemical NER. First, the authors introduced the fundamental concepts of chemical named entity recognition, the textual contents of chemical documents, and how these chemicals are represented in chemical literature. The chapter concludes with the strengths and weaknesses of the above methods and also the types of the chemical entities extracted.

INTRODUCTION

In this Internet age, enormous measure of digital data is created and partaken as unstructured writings, pictures and recordings. This sort of remarkable development is going on in every single fields of our genuine science through examination articles, specialized reports and digital books. For example, in

the event that we think about the field of Chemistry, there are N number of diaries accessible independently for the five essential parts of science, for example, Physical, Organic, Inorganic, Analytical, and Biochemistry. These journals approximately contain over 10 million journal papers. Manually, it is very difficult to organize, manage, identify and to extract the important information like chemical named entities and their relations. In addition, the extracted information is stored in the form of databases such as CHEMDNER corpus (Krallinger et al., 2015), ChemSpider (Krallinger, Leitner, & Rabal, 2013), PubChem (ChemSpider, 2010), ChEBI (Harry, 2010) etc. Notwithstanding, to refresh these information bases, we require broad and persistent manual exertion which is an extravagant and provoking errand and prompts the evolvement of Text Mining (TM) to play out those undertakings.

One significant focal point of Text Mining is on Named Entity Recognition (NER), an absolute first and urgent advance in data extraction. In the specific instance of Chemical Named Entity Recognition (CNER), extra challenges emerge. Most importantly, not all chemical names are having a similar structure. CNER frameworks are especially delicate concerning spelling mistakes and the tokenization technique since chemical substances are generally having hyphenated text portions, variable utilization of enclosures, sections and various punctuation symbols. In addition, compound reports will in general be stacked with many abbreviations and acronyms, which are one of the primary wellsprings of false positives. Another trademark that makes CNER troublesome is the way that the location of limits for chemical substances is particularly lumbering when long compound names are available. A more top to bottom description of challenges in labelling chemicals can be found at (Krallinger et al., 2015).

In chemical databases, it is very difficult to find out the specific information on newly discovered compounds. For example, a new drug development process is based on the knowledge of chemicals like toxic effects, biological properties etc. The chemical entities extracted using Text Mining can be used in many applications such as to find relationships between many entities, mapping the chemical entities to their identical structures, and helps a search engine to retrieve the specific documents which contains one particular entity (Article, 2016). However, the various naming culture of chemical entities makes the task of entity recognition very complex and time consuming. Manual curation of chemical text documents to generate annotations and to use those annotations is also a very laborious process (Kim et al., 2015).

Consequently, a few ChemNER frameworks have been created utilizing various methodologies dependent on standards, dictionary coordinating, Machine Learning (ML) and Deep Learning (DL). Each approach has its own points of interest and disserves relying upon the semantic attributes of the elements being recognized. Applying the best methodologies is preposterous in all cases, since each approach presents diverse specialized necessities (Wang et al., 2009). In any case, when there is a huge volume of information, Deep Learning based arrangements present several advantages over different strategies and give the outcomes best precision. The advancement of DL-based NER arrangements coordinates different complex advances that consolidate distinctive processing pipelines. Hence, along the previous years, an assortment of frameworks was created utilizing the most various structures, procedures and systems.

As we consider the procedures applied in ChemNER, there are four fundamental techniques:

- Rule-based methodologies: Depend with respect to hand-made guidelines thus no requirement for commented on information
- Unsupervised learning approach: Depend on Unsupervised algorithms with no preparation models
- Supervised learning approach: Depend on supervised learning algorithms with cautious component designing;

Related Content

Deriving Business Value From Online Data Sources Using Natural Language Processing Techniques
Stephen Camilleri (2021). *Natural Language Processing for Global and Local Business (pp. 17-39).*
www.irma-international.org/chapter/deriving-business-value-from-online-data-sources-using-natural-language-processing-techniques/259782

NLP and the Representation of Data on the Semantic Web
Jose L. Martinez-Rodriguez, Ivan Lopez-Arevalo, Jaime I. Lopez-Veyna, Ana B. Rios-Alvaradoand Edwin Aldana-Bobadilla (2021). *Handbook of Research on Natural Language Processing and Smart Service Systems (pp. 393-426).*
www.irma-international.org/chapter/nlp-and-the-representation-of-data-on-the-semantic-web/263114

Deep Learning Network: Deep Neural Networks
Bhanu Chander (2020). *Neural Networks for Natural Language Processing (pp. 1-30).*
www.irma-international.org/chapter/deep-learning-network/245081

Cloud Hosted Ensemble Learning-Based Rental Apartment Price Prediction Model Using Stacking Technique
 Rajkumar S., Mary Nikitha K., Ramanathan L., Rajasekar Ramalingamand Mudit Jantwal (2023). *Deep Learning Research Applications for Natural Language Processing (pp. 229-238).*
www.irma-international.org/chapter/cloud-hosted-ensemble-learning-based-rental-apartment-price-prediction-model-using-stacking-technique/314147

Author Profiling Using Texts in Social Networks
Iqra Ameerand Grigori Sidorov (2021). *Handbook of Research on Natural Language Processing and Smart Service Systems (pp. 245-265).*
www.irma-international.org/chapter/author-profiling-using-texts-in-social-networks/263105