Chapter 76
Machine Learning Approach for Multi-Layered Detection of Chemical Named Entities in Text

Usha B. Biradar
Molecular Connections Pvt Ltd., Bangalore, India

Harsha Gurulingappa
Molecular Connections Pvt Ltd., Bangalore, India

Lokanath Khamari
Molecular Connections Pvt Ltd., Bangalore, India

Shashikala Giriyan
Molecular Connections Pvt Ltd., Bangalore, India

ABSTRACT

Identification of chemical named entities in text and subsequent linkage of information to biological events is of immense value to fulfill the knowledge needs of pharmaceutical and chemical R&D. A significant amount of investigation has been carried out since a decade for identifying chemical named entities at morphological level. However, a barrier still remains in terms of value proposition to scientists at chemistry level. Therefore, the work described here aims to circumvent the information barrier by adaptation of a Conditional Random Fields-based approach for identifying chemical named entities at various levels namely generic chemical level, morphological level, and chemistry level. Substantial effort has been invested on generation of suitable multi-level annotated corpora. Recommended machine learning practices such as active learning-based training corpus generation and feature optimization have been systematically performed. Evaluation of system performance and benchmarking against the other state-of-the-approaches showed improved results.

DOI: 10.4018/978-1-7998-2460-2.ch076

INTRODUCTION

In today's era of big data, scientific discovery process is largely dependent on integration, management and extraction of useful data from available literature (Borkum & Frey, 2014). Extracted information from text mining tasks in chemical literature domain mainly includes named entities. Mining the chemical named entities is aimed at extracting information on unique chemicals, identifying the extracted chemicals by indexing them to the databases and bibliographic sources, assign and verify relationships between chemical entities and biological process, diseases etc., (Eltyeb & Salim, 2014; Banville, 2006; Batchelor & Corbett, 2007).

Machine Learning (ML) which is the automation of processes attributed to human intelligence, in particular - learning, to make decisions and to solve problems based on learning outcomes (Russell et al., 1995; Bottou, 2014), provides tailor made solutions for the task of named entity recognitions. Of late, Conditional Random Fields (CRFs), a class of probabilistic ML methods have contributed to major success in Chemical Named Entity Recognition (CNER) (Klinger et al., 2008). Ambiguity in representations of chemical entities is perhaps the most prevalent limitations concerned with text mining applications to chemical literature amongst others like limited open text corpora and growing number of chemicals (Townsend et al., 2005; Gurulingappa et al., 2013). Figure 1 clearly demonstrates the necessity and importance of named entity recognition as a first step to enable knowledge discovery process in chemical scientific literature.

Figure 1. Different representations of the chemical named entity 'ethanol'

Ethanol, also called **ethyl alcohol**, pure alcohol, beverage alcohol, or drinking alcohol, is a volatile, flammable, colorless liquid with the structural formula **CH3CH2OH**, often abbreviated as **C2H5OH** or **C2H6O**. It is also represented as **EtOH**, **Ethyl hydroxide** and based on the position of the functional group, its nomenclature would be **1-hydroxyethane**

. . . . . .

In spite of humongous work done on application of various approaches for chemical named entity recognition, most of the efforts have concentrated on identifying chemical names at generic level (e.g. chemical against non-chemical) or morphological level (e.g. trivial name, IUPAC, abbreviation, formula or chemical class). To the best of author's knowledge, there is no effort on identifying chemical names at chemistry level such as organic, inorganic, organometallic, drug, macromolecule and so-forth. Primary reason is because generating annotated corpora is an extremely labor intensive task and similarly annotating corpora with multi-level information including chemistry information requires additional efforts from domain experts.