International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 07 | Oct-2015 www.irjet.net p-ISSN: 2395-0072

Chemical Name Extraction and its application to Rule-based System

Snehal Pandharinath Umare

Student, Computer Department, Savitribai Phule Pune University, Maharashtra, India

---

Abstract - ***Automatic extraction of chemical names has a great importance in biomedical and life science research field. In the proposed system, accuracy in training data generation is improved by generating large amount of regular expressions. Regular expressions are generated from systematic chemical names collected from PubChem Database. Due to large amount of regular expressions, accuracy in chemical name extraction is also improved. In the work carried out, an efficient chemical name extraction system is designed. It is deployed by processing the algorithm in following steps. First stop words are removed. After removal of stop words trivial name extraction is performed by dictionary matching approach. Then CAS/Registry numbers are extracted by regular expression. For systematic name extraction, sample systematic names from various categories are given as an input for training. Normalization type 1 and type 2 is performed on these sample systematic names which is followed by tokenization. Finally regular expressions are generated from training data. Input document is then matched with the regular expressions to extract systematic chemical names.***

Key Words: *Chemical Name Extraction, IUPAC Names, CAS Numbers, Regular expression.*

1. INTRODUCTION

Lot of novel chemical structures exist in chemical related articles and patents, which are not in accessible formats. In this system, a technique is presented to extract chemical names mentioned the textual documents automatically. After extraction these names are useful for indexing, searching, and mining tasks [1].

Reasons for extraction of chemical names from the documents are:
1. For identifying unique chemical names present in documents.
2. For indexing purpose.
3. To link between chemical structures and biological processes.

Named Entity Recognition (NER) recognizes specific entities in the text, such as chemical names. There are three NER approaches:
1. Dictionary-based NER systems
2. Rule-based NER systems
3. Machine learning based NER systems

1. Dictionary based NER systems [2][4]:

Dictionary solutions with string matching or regular expressions are generally used to extract the chemical names which are not systematic, e.g. registry numbers / CAS numbers, trivial names. This method is later of no use when dictionaries are changed and become out-dated. Dictionary based approach uses 2 methods:
(a) Exact matching and
(b) Flexible or approximate matching

2. Rule-based NER systems [2][5]:

For chemical name extraction, a rule-based solution gives decent extraction performance. Rule-based systems make the use of handmade rules for extracting the names of entities. E.g. Grammatical and syntactic rules are sometimes combined with dictionaries. Rule-based systems are used to extract CAS or registry numbers from the text. There are two types of rules usually used in the rule-based approach:
(a) Pattern-based rules
(b) Context-based rules
Examples:
(a) List Lookup Approach
(b) Linguistic Approach

3. Machine learning based NER systems/Automated Approach [2][6]:

Compared to rule-based methods, learning-based solutions do not need comprehensive domain knowledge to build NE models when a large amount of training data is available, and usually offer better generalization performance. The steps required to develop Machine Learning based systems are:
(a) **Training**: The ML model must be trained to use the annotations present in the annotated documents.
(b) **Annotating**: The documents can be annotated to generate the entity names based on experience learned from the annotated documents.
Examples: Learning based NE techniques to implement Machine learning NER systems are:
(a) Hidden Markov Models (HMMs)

(b) Maximum Entropy Markov Model (MEMM)
(c) Conditional Random Field (CRF)
(d) Support Vector Machine (SVM)
(e) Decision Tree (DT)

2. LITERATURE SURVEY

In recent years, many works have reported promising chemical NE results based on various techniques. The work reported in [4] uses dictionary based approach to find non systematic chemical names i.e. trivial/common names. The algorithms used are exact matching algorithm and approximate matching algorithm. They have used 1000 Medline abstracts. The Precision calculated is 98%, Recall 88%, and F-measure is 92.73%. The work in [5], uses rule based approach to find CAS/Registry numbers. It uses pattern based and context based algorithm. The results are reported on 50 Medline abstracts. Precision is 76%, Recall 84% and F-measure 80%. The paper [12] uses support vector machine (SVM) model to find all gene names in the training and devtest corpora. The work in [14] uses Context aware CRF technique to find systematic names. The results are Precision 90.91%, Recall 82.19% and F-measure 86.33%. The work carried out in [15], uses machine learning approach to find systematic chemical names. It uses CRF model. 1000 Medline abstracts are used to test the results. Precision, Recall, and F-measure are 86.5%, 84.8%, 85.6%. The work in [16] uses HMM i.e. hidden markov model. It uses Forward-Backward algorithm, Viterbi algorithm and Estimation modification algorithm to find systematic chemical names. They tested their results on GENIA corpora. The results reported are Precision 63.8%, Recall 61.3%, and F-measure 62.5%.

3. SYSTEM ARCHITECTURE

The proposed system extracts Common/Trivial chemical names using dictionary based approach. It also extracts CAS/Registry numbers by rule based approach i.e. regular expression. Extraction of systematic chemical names is difficult task. So to extract systematic names, proposed system generates regular expressions from the dictionary of systematic names. These regular expressions are matched with the input document to retrieve the systematic chemical names. This method provides more accuracy as large amount of regular expressions are generated.
Fig. 1 shows the architecture of proposed system. It works in 2 phases. First phase is training phase and second phase is testing.

1. Training Phase:

In the training phase, training data is generated by performing following steps. For generating training data, sample systematic names are given as an input. These systematic names are downloaded from Pubchem database. These names are pre-processed by performing Normalization type 1 and type 2. In Normalization type 1 all digits are replaced with a single digit '9'. In normalization type 2 all words are replaced with a single alphabet 'a'. Later tokenization is performed by removing white spaces, commas, and punctuation marks. Thus a sequence is generated for each token. From these sequences Regular Expressions are generated automatically which are helpful in chemical name extraction process. Followed by training phase testing phase is performed.

2. Testing Phase:

In testing phase, user gives any chemical related text file as an input. Using Stop words removal algorithm, stop words and phrases are removed from the text. We also maintain a chemical dictionary containing Trivial/Common chemical names. By using dictionary approach, i.e string matching trivial names are extracted from the text. Using Rule based approach i.e. regular expression CAS/Registry numbers are extracted. For systematic names extraction we use rule based approach. It is somewhat complex to extract systematic/IUPAC names. For that input text is matched with the Regular Expressions generated in training phase. Wherever match is found, those chemical names are extracted as systematic chemical names.

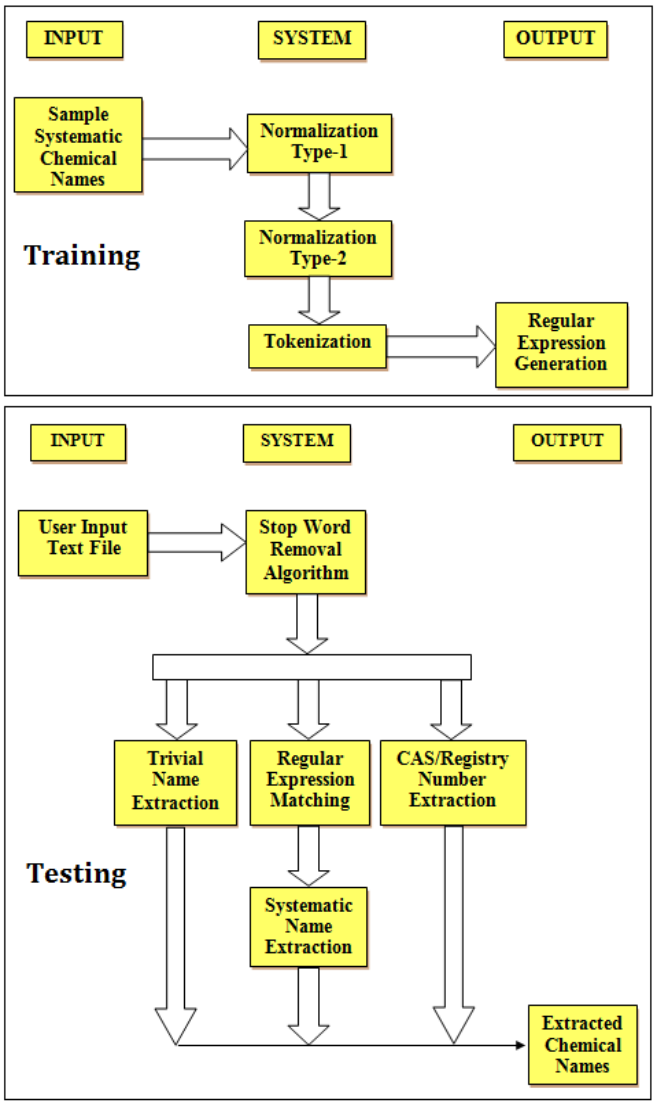

Fig. 1 System Architecture

4. EXPERIMENTAL RESULTS

It is essential that a NE technique is not hard-coded for a specific domain, and is general enough to handle text from various sources. To this end, we evaluate the performance of the proposed system on two different data

1. **MEDLINE-200:** The corpus MEDLINE-200 contains 200 randomly selected abstracts of MEDLINE scientific articles that are chemically related.
2. **Chemical Book:** The corpus is a chemical book downloaded from the internet.

We compare our proposed system with the existing system RandDict and ChemAxon D2S using Medline corpora and report the results in table 1.

Table -1: Performance Evaluation on Medline Data

Data Set	RandDict	ChemAxon	Proposed System
Precision	90.83%	96.88%	93.75%
Recall	93.62%	92.72%	97.29%
F-1	92.20%	94.75%	95.48%

The Chart 1 shows the comparison between existing system i.e. RandDict and Proposed system.

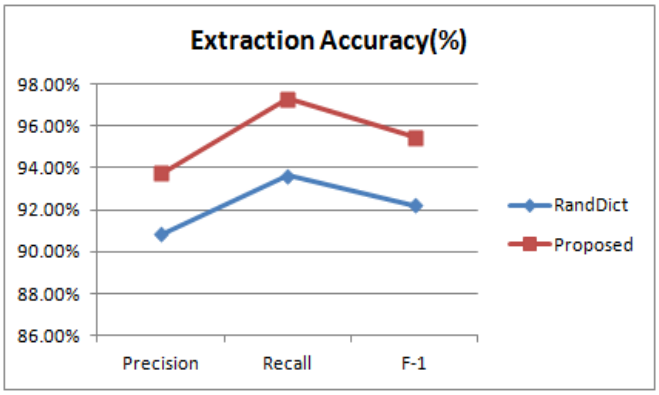

Chart -1: Comparison between RandDict and Proposed System

The Chart 2 shows the comparison between ChemAxon commercial software and proposed system.

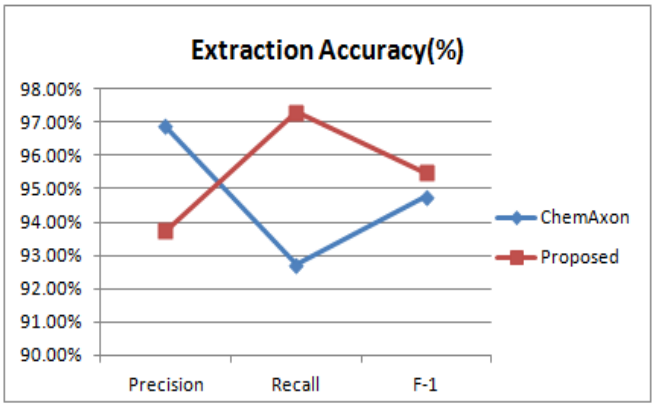

Chart -2: Comparison between ChemAxon and Proposed System

We tested our results on one more data set i.e. chemical book. The results are as shown in table 2

.

Table -2: Performance Evaluation on Chemical Book

Data Set	Proposed System
Precision	95.65%
Recall	91.66%
F-1	93.61%

Fig.4 shows extraction accuracy on chemical book data set.

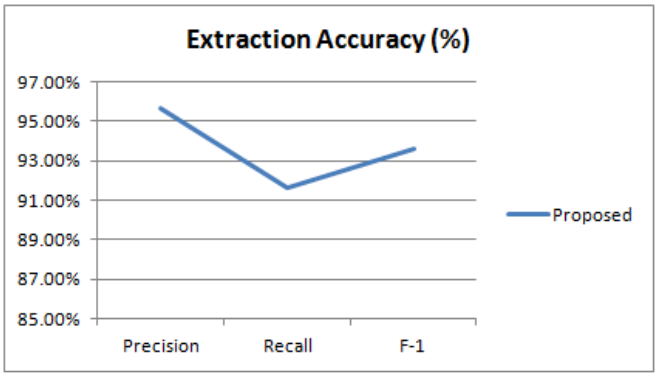

Chart -3: Extraction accuracy on chemical book

4.1 Discussion on Proposed System

A pair of NE models is compared efficiently in the following manner. For each document, the lists of extracted chemical names from the two models, in the order in which they appear in the document, are compared side by side and evaluated manually. Mistakes by each model are counted and subtracted from the total number of extractions by that model. We then get the refined total number of extractions for each model. Then we manually count the total number of systematic names in the documents and consider it as ground truth. E.g. Assume Model A extracts 100 chemical names where 80 of them are correct. Model B extracts 100 chemical names where 85 of them are correct. And the total number of correct chemical names mentioned in the document is 90. The approximate evaluation works as follows: Since the number of correct names are 90, Recall(A) = 80=90 =0:89, Precision(A) = 80=100 = 0:8, F1(A) = 0:84. And Recall(B) = 85=90 = 0:94, Precision(B) = 85=100 = 0:85, F2(B) = 0:89. This evaluation is approximate. Nevertheless, it is easy to execute and still provides a fair comparison between two sets of results. According to the evaluation results, the proposed system performs significantly better than the other techniques. To understand the performance variances across data sets, we categorize chemical names into two groups. Group 1 contains complex systematic names, where each name has at least one dash and one other punctuation mark in the set of {[,] {,} (,)} or contains one pattern from the set of {-9, 9-}. The rest of the chemical names are "simple names", most of which are common/trivial names.

Our proposed system is more effective in extracting complex systematic chemical names. As our system does not use dictionary lookups, it misses out some systematic chemical names. The ability to extract complex systematic chemical names is the key research challenge for two reasons. First, novel chemical structures are always reported as complex names. Second, complementary components such as dictionary lookups and regular expressions can be readily added to a NE system to relatively easily identify common/trivial names.

5. CONCLUSIONS

The proposed system presents a novel idea for training a chemical NE model based on automatically generated regular expressions. The system uses rule based approach to extract Systematic/IUPAC chemical names. IUPAC names are downloaded from PubChem database. We also extract IUPAC names available on the web. For training data generation IUPAC names from various categories are given as an input to generate regular expressions. The proposed system extracts Common/Trivial chemical names using dictionary matching. For this a dictionary of trivial names is maintained. System uses regular expression to extract CAS/Registry numbers. Compared to other chemical NE models, proposed system shows better NE extraction performance.

REFERENCES

[1] Su Yan, W.Scott Spangler, and Ying Chen, "Chemical Name Extraction Based on Automatic Training Data Generation and Rich Feature Set", Computational Biology and Bioinformatics, vol. 10, NO. 5, SEPTEMBER/OCTOBER 2013.

[2] Snehal P. Umare, Dr. Neeta A. Deshpande, "A Survey on Machine Learning Techniques to Extract Chemical Names from Text Documents", International Journal of Computer Science and Information Technology, vol. 04, 1263-1266, PP.0975-9646, 2015, .

[3] Safaa Eltyeb and Naomie Salim, "Chemical named entities recognition: a review on approaches and applications", Apr 28, 2014.

[4] Sergei Egorov, PhD, Anton Yuryev, PhD, and Nikolai Daraselia, PhD. "A Simple and Practical Dictionary-based Approach for Identi_cation of Proteins in Medline Abstracts".

[5] R.Porkodi, B.LShivakumar, Rule based approach for constructing Gene/Protein names Dictionary from Medline abstract, Department of Computer Science, Bharathair University, Coimbatore, Tamilnadu, India.

[6] Gitimoni Talukdar1, Pranjal Protim Borah2, Arup Baruah3, "A SURVEY OF NAMED ENTITY RECOGNITION IN ASSAMESE AND OTHER INDIAN LANGUAGES". Department of Computer Science and Engineering, Assam Don Bosco University, Guwahati, India.

[7] Sudha Morwal, Nusrat Jahan and Deepti Chopra, "Named Entity Recognition using Hidden Markov Model (HMM)", International Journal on Natural Language Computing (IJNLC) Vol. 1, No.4, December 2012.

[8] Asif Ekbal, Sivaji Bandyopadhyay, "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi", CSLI Publications, Linguistic Issues in Language Technology LiLT, Volume 2, Issue 1 November, 2009.

[9] Xiaojin Zhu, CS838-1 Advanced NLP: Conditional Random Fields, 2007. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[10] A. McCallum, D. Freitag, and F.C.N. Pereira, Maximum Entropy Markov Models for Information Extraction and Segmentation, Proc. 18th Intl Conf. Machine Learning (ICML 00), pp. 591-598, 2000.
[11] Nusrat Jahan, Sudha Morwal, International Journal of Engineering Sciences and Research Technology, "Named Entity Recognition in Indian Languages: A Survey", Department of computer science, Banasthali University, Jaipur-302001, Rajasthan, India.
[12] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, vol. 20, no. 3, pp. 273-297, Sept. 1995.
[13] Lior Rokach (Ben-Gurion University of the Negev, Israel), Oded Maimon (Tel-Aviv University, Israel), "Data Mining with Decision Trees", Volume 69.
[14] Prakash Hiremath, Shambhavi B. R, "Approaches to Named Entity Recognition in IndianLanguages: A Study", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 8958, Volume-3 Issue-6, August 2014.
[15] R. Klinger, C. Kolarik, J. Fluck, M. Hofmann-Apitius, and C.M. Friedrich, "Detection of IUPAC and IUPAC-Like Chemical Names", Bioinformatics, vol. 24, pp. 268-276 2008.
[16] Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, Chew-Lim Tan,"Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain".
[17] P. Corbett, C. Batchelor, and S. Teufel, "Annotation of Chemical Named Entities", Proc. Workshop BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP 07), pp. 57-64, 2007.
[18] Mary Elaine Cali_ , Raymond J. Mooney, "Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction", Journal of Machine Learning Research 4 (2003) 177-210, Submitted 12/01; Revised 2/03; Published 6/03.
[19] Califf, Mary. "Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction.", Journal of Machine Learning Research. (2003)177-210, n. page. Print. http://www.cs.utexas.edu/ ailab/pubs/rapierjmlr03.pdf.
[20] R. Panico, W. Powell, and J. Richer, "A Guide to IUPAC Nomenclature of Organic Compounds": Recommendations 1993, IUPAC Chemical Data Series, 1993, I. U. of Pure, A. C. C. on the Nomenclature of Organic Chemistry.