A hybrid text mining system for chemical entity recognition and classification using dictionary look-up and pattern matching @ BeCalm challenge evaluation workshop

Kalpana Raja[1*], Sabenabanu Abdulkadhar[2], Lam C Tsoi[1,3,4]

Jeyakumar Natarajan[2*]

[1]Department of Dermatology, University of Michigan Medical School, Ann Arbor, MI, USA
[2]Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamilnadu, India
[3]Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

*[1]rkalpana@med.umich.edu;sabenabanu.a@hotmail.com;
alextsoi@med.umich.edu;[2*]n.jeyakumar@yahoo.co.in

Abstract. Chemicals as therapeutics and investigational agents receive much attention in clinical research and applications recently. However, automated approaches to recognize and categorize the chemical entities in biomedical text are challenging because of the wide varieties of morphologies and nomenclature. We present here a hybrid text mining system that combines chemical lexicon and patterns for recognition/categorization. We applied this approach to identify chemical entities from the patent abstracts of BioCreative V.5 Chemical Entity Mention Recognition (CEMP) corpus. We also compared the hybrid approach with the "traditional" lexion-based method, and illustrated that the hybrid approach can achieve enhanced performance (i.e. precision, recall, and F-score) than the lexion-based method.

Keywords. Chemical entity recognition; Chemical categorization; Text mining; Pattern matching; Chemical lexicon.

1 Introduction

The advances in data revolution reveal valuable information on the new roles of chemicals in disease treatment and adverse reaction. The effect of chemicals on the biological systems as therapeutic agents (i.e. drugs), investigational agents in drug discovery and unintentional agents to understand the adverse effects make them an important class of biomedical entities in clinical research and applications [1]. The scientific findings on chemicals are commonly available in published bi-

omedical literature, and automated approaches using text mining techniques have proven to be effective for entity extraction. Nevertheless, the task is challenging due to the different morphologies and nomenclatures used for representing chemical entities [2].

Different text-mining approaches have been developed utilizing techniques such as rule-based [3], dictionary based [4], machine learning [1], and hybrid approaches [5]. In particular, the Conditional Random Fields [6,1,5] and Support Vector Machines [7] algorithms with rule-based or dictionary-based approaches are widely used in chemical entity recognition. In spite of several existing approaches, the challenge is still open and leaves a space for improvement. BioCreative V.5 Chemical Entity Mention Recognition (CEMP) task invited the text mining community to develop novel and robust approaches for recognizing and categorizing the chemical entities in a set of patent abstracts. We presented a hybrid approach that combines a chemical lexicon and a pattern matching module for recognizing and categorizing chemicals from the patent abstracts.

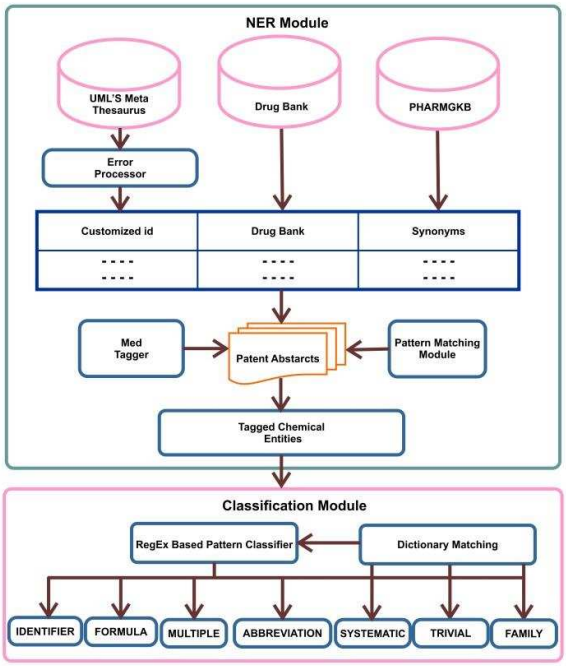

Figure 1: Workflow of the proposed system

2 Systems description and methods

Overview

Our approach for recognizing and categorizing the chemicals consists of two parts: (1) building a chemical lexicon; (2) extracting and categorizing the chemicals from the patent abstracts using the lexicon and a pattern matching module. Figure 1 presents the workflow of the proposed system.

Chemicals lexicon

Processing of UMLS Metathesaurus

The chemical lexicon was compiled from three resources: UMLS Metathesaurus [8], DrugBank [9] and PharmGKB [10]. The 2015AB version of UMLS Metathesaurus was downloaded from the UMLS Terminology Services (UTS) and customized to Rich Release Format using MetamorphoSys, an in-build UMLS installation wizard and Metathesaurus customization tool [11]. The resource contains more than 3.2 million medical concepts (e.g. chemicals, drugs, diseases) and 12.8 million synonyms from over 190 vocabularies including SNOMED, and ICD 9/10 diagnostic codes. We filtered 4,652,003medical concepts and synonyms that are in English, and selected only the concepts belonging to "Chemicals and Drugs" semantic type. However, the concepts from other semantic groups (e.g. disease, living organisms) were found to overlap with the chemicals synonyms. We removed the common concepts or synonyms between the sematic groups, common English terms (e.g. link, conduct, aim), semantic types (e.g. amino acid, hormone) and abbreviation overlap (e.g. the abbreviation C maps to Catechin, Cocaine, Carbon and Blood group antigen C). We used Attempto, a resource for controlled natural language and a rich subset of Standard English [12] to identify common English terms as chemicals. The latest version is released in 2013 and contains 97,526 English terms. The other three error types were identified through simple overlap. The process yielded 461,379 chemical concepts and 929,747synonyms.

Processing of DrugBank and PharmGKB

The latest version of DrugBank database [13] was downloaded and parsed with UTF-8 encoding to handle chemicals with Greek alphabets

and special characters (e.g. α-methylthiofentanyl). We extracted 8,203 drugs, 1,201 salts, and synonyms. We downloaded the drugs.zip file from PharmGKB and extracted 3,175 chemical names, 6,763 generic names and 18,309 trade names [10]. The generic names and trade names are synonyms. We compiled the chemical entities and synonyms from the three resources and assigned a customized ID which is unique for a chemical.

Recognition of chemical entity mention

We combined the chemical lexicon with MedTagger [14] for application. MedTagger is an Open Health Natural Language Processing (OHNLP) tagger that uses a lexicon for entity extraction. It uses a pipeline of text mining approaches such as tokenization, lexical normalization, dictionary look-up using the well-known Aho-corasick approach and concept screening [15]. The CEMP task defines seven chemical classes namely systematic, identifier, formula, trivial, abbreviation, family and multiple classes for categorization. The systemic class defines the IUPAC and IUPAC-like chemical nomenclature (e.g. 2-acetoxy-benzoic-acid). The identifiers are the database identifiers from various chemical databases (e.g. 2244, a PubChem ID). The formula includes molecular formula (e.g. $CH_3COOC_6H_4COOH$), canonical and isomeric SMILES (e.g. CC(=O)OC1=CC=CC=C1C(=O)O), InChI (e.g. InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)), and InChIKey (e.g. BSYNRYMUTXBXSQ-UHFFFAOYSA-N). The trivial names are the trade name / brand name / common name / generic name of a marketed drug (e.g. acylpyrin, acetaminophen, tylenol, panadol, and resveratrol for aspirin). The abbreviations are the standard acronyms (e.g. GABA for gamma-aminobutyric acid). The family class is associated with the chemical structure (e.g. diphenols, terphenoids). The multiple class includes the chemical names that are not described in a continuous string of characters (e.g. thieno3,2-d fused oxazin-4-ones). We developed a pattern matching module using Java regex to categorize the chemicals into seven different classes as shown in Table 1. Among the seven classes, formula and identifiers are not available in the chemical lexicon and the pattern matching module alone was applied for their recognition. In addition, we used a lexicon compiled from PubChem [16], DrugBank [13] and KEGG DRUG [17] for classifying trivial names. The family

class consists of sub-classes and we created a family lexicon to distinguish the family sub-classes (e.g. systematic) from the main systematic class.

Table 1: Patterns to classify chemical Entity classes

Chemical class	Regex pattern	Example
SYSTEMATIC	\\b(^\\d.*$)\\b	beta.-alethine
IDENTIFIER	\w+ [A-Za-z] \w+ [0-9]\w+	KMD-3213
FORMULA	\\b[A-Z][a-z]?\\d*|[A-Z]?\\d*\\b	CH2 COOH
TRIVIAL	[A-Z][a-z]\w+	matrinetannate
ABBREVIATION	[A-Z]{4}	EDDA
FAMILY	[a-z]|[A-Z][a-z]	Alkaloids
MULTIPLE	\\b(and|or)\\b	Ceratamines A and B

Dataset and Evaluation

The BioCreative V.5 CEMP task consists of 21,000 patent abstracts in the training data and 9,000 patent abstracts in the test data. While the training data consists of 99,632 annotations (Table 2), the annotations for test data are yet to be released. The standard evaluation metrics such as Precision, Recall and F-score were used to evaluate the performance of the proposed system [18].

Table 2: Chemicals annotation in the training data

Chemical class	Annotation
SYSTEMATIC	28,580
IDENTIFIER	278
FORMULA	6,818
TRIVIAL	25,927
ABBREVIATION	1,373
FAMILY	36,238
MULTIPLE	418

3 Results and Discussion

The traditional lexicon-based approach achieved 0.532 precision, 0.651 recall and 0.586 F-score on the training data and 0.472 precision, 0.515 recall and 0.492 F-score on the test data (Table 3). The hybrid approach that combines the chemical lexicon and patterns recognition/categorization achieved an enhanced performance of 0.601 precision, 0.651 recall and 0.625 F-score. We also report the performance of the system on each component i.e. entity recognition and classification (Table 4). The chemical dictionary is the main component for identifying five types of classes excluding identifiers and formula. Though the dictionary contains the systematic and trivial names, we observed that it does not contain IUPAC like names and all trivial names. By incorporating a pattern matching approach for IUPAC like names and including a lexicon for trivial names from resources that are not in chemical lexicon, we show an enhanced performance of more than 10% on precision, recall, and F-score when compared to the traditional lexicon-based approach.

Table 3: System performance reported for CEMP task

Dataset	Precision	Recall	F-score
Training data	0.532	0.587	0.558
Test data	0.472	0.515	0.493

Table 4: System performance after CEMP task on training data

Approach	Precision	Recall	F-score
Entity recognition	0.582	0.651	0.614
Classification	0.570	0.773	0.658
Entity recognition+ Classification	0.601	0.651	0.625

Limitations and Future work

Though the patterns perform well on single term entities (e.g. SMILES), we observed pattern overlap on entities with multiple terms (e.g. IUPAC like names) that resulted in partial identification of chemical entities. The chemical annotations in the training data mainly be-

long to systematic, trivial and family class (i.e. ~89%), among which the recognition of systematic and family classes are more challenging. As a future work, we will be replacing the patterns with a machine learning approach using CRF, an established approach for entity recognition.

4 Conclusion

We present a hybrid approach that combines a chemical lexicon compiled from three resources namely UMLS Metathesaurus, DrugBank and PharmGKB, and a pattern matching approach for chemicals entity recognition and classification into seven different classes defined in BioCreative V.5 CEMP task. We report the performance of the proposed system only on entity recognition and classification task, and as a system with both the modules. In the current study, we take the advantage of many available resources (e.g. PubChem, KEGG Drug) and a pattern matching approach for entity recognition and classification.

Acknowledgement

The research has received funding from Dermatology Foundation USA, the Arthritis National Research Foundation USA and the National Psoriasis Foundation USA, and the University Grants Commission Maulana Azad National Fellowship for Minority students (UGC-MANF), Government of India, Grant No: F1-17.1/2015/MANF-2015-17-TAM-54928. The authors acknowledge all the funding received.

REFERENCES

1. Leaman, Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." Journal of cheminformatics 7.1 (2015): S3.
2. Usié, Anabel, Joaquim Cruz, Jorge Comas, FrancescSolsona, and RuiAlves. "CheNER: a tool for the identification of chemical entities and their classes in biomedical literature." Journal of cheminformatics 7.1 (2015): S15.
3. Narayanaswamy, Meenakshi, K. E. Ravikumar, K. Vijay-Shanker, and K. Vij Ay-shanker. "A biological named entity recognizer." Paper presented at Pacific Symposium on Biocomputing, Kauai, Hawaii, January 3-7, 2003.
4. Hettne, Kristina M., Rob H. Stierum, Martijn J. Schuemie, Peter JM Hendriksen, Bob JA Schijvenaars, Erik M. Van Mulligen, Jos Kleinjans, and Jan

A. Kors. "A dictionary to identify small molecules and drugs in free text." Bioinformatics 25.22 (2009): 2983-2991.

5. Rocktäschel, Tim, Michael Weidlich, and Ulf Leser. "ChemSpot: a hybrid system for chemical named entity recognition." Bioinformatics 28.12 (2012): 1633-1640.
6. Klinger, Roman, CorinnaKolářik, JulianeFluck, Martin Hofmann-Apitius, and Christoph M. Friedrich.. "Detection of IUPAC and IUPAC-like chemical names."Bioinformatics 24.13 (2008): i268-i276.
7. Zhang, Yaoyun, Jun Xu, Hui Chen, Jingqi Wang, Yonghui Wu, Manu Prakasam, and HuaXu. "Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning." Database 2016 (2016): baw049.
8. Lindberg, Donald AB, Betsy L. Humphreys, and Alexa T. McCray. "The unified medical language system." IMIA Yearbook (1993): 41-51.
9. Wishart, David S., Craig Knox, An Chi Guo, Dean Cheng, SavitaShrivastava, Dan Tzur, BijayaGautam, and MurtazaHassanali. "DrugBank: a knowledgebase for drugs, drug actions and drug targets." Nucleic acids research 36.suppl 1 (2008): D901-D906.
10. Hewett, Micheal, Diane E. Oliver, Daniel L. Rubin, Katrina L. Easton, Joshua M. Stuart, Russ B. Altman, and Teri E. Klein. "PharmGKB: the pharmacogenetics knowledge base."Nucleic acids research 30.1 (2002): 163-165.
11. Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic acids research 32.suppl 1 (2004): D267-D270.
12. Fuchs, Norbert E., KaarelKaljurand, and Tobias Kuhn. "Attempto controlled english for knowledge representation." Reasoning Web. Springer Berlin Heidelberg, 2008. 104-124.
13. Law, Vivian, Craig Knox, YannickDjoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewskiet al. "DrugBank 4.0: shedding new light on drug metabolism."Nucleic acids research 42.D1 (2014): D1091-D1097.
14. Torii, Manabu, KavishwarWagholikar, and Hongfang Liu. "Using machine learning for concept extraction on clinical documents from multiple data sources." Journal of the American Medical Informatics Association 18.5 (2011): 580-587.
15. Aho, Alfred V., and Margaret J. Corasick. "Efficient string matching: an aid to bibliographic search." Communications of the ACM 18.6 (1975): 333-340
16. Kim, Sunghwan, et al. "PubChem substance and compound databases."Nucleic acids research (2015): gkv951.
17. Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. "KEGG for integration and interpretation of large-scale molecular data sets." Nucleic acids research (2011): gkr988.
18. Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." Paper presented at 19thAustralasian Joint Conference on Artificial Intelligence, Australia December 4-8,2006.