RELATION EXTRACTION BETWEEN BIOMEDICAL ENTITIES FROM LITERATURE USING SEMI- SUPERVISED LEARNING APPROACH

Saranya M[1], Arockia Xavier Annie R[2] and Geetha T V[3]

[1]Computer Science and Engineering, CEG, Anna University, India
[2]Assistant Professor, Computer Science and Engineering,
CEG, Anna University, Chennai, India
[3]UGC-BSR Faculty Fellow, Computer Science and Engineering, former
Dean CEG, Anna University, Chennai, India

ABSTRACT

Now-a-days, people around the world are infected by many new diseases. The cost of developing or discovering a new drug for the newly discovered disease is very high and prolonged process. These could be eliminated with the help of already existing resources. To identify the candidates from the existing drugs, we need to extract the relation between the drug, target and disease by textming a large-scale literature. Recently, computational approaches which is used for identifying the relationships between the entities in biomedical domain are appearing as an active area of research for drug discovery as it needs more man power. Due to the limited computational approaches, the relation extraction between drug-gene and gene-disease association from the unstructured biomedical documents is very hard. In this work, we proposed a semi-supervised approach named pattern based bootstrapping method to extract the direct relations between drug, gene and disease from the biomedical literature. These direct relationships are used to infer indirect relationships between entities such as drug and disease. Now these indirect relationships are used to determine the new candidates for drug repositioning which in turn will reduce the time and the patient's risk.

KEYWORDS

Text mining, drug discovery, drug repositioning, bootstrapping, machine learning.

1. INTRODUCTION

For developing the new chemical compound into the market for treating appropriate disease is called as drug development or design process which is very expensive and takes minimum 12-15 years from the starting stage to the marketing. Currently, the arrival of new diseases is increased and most of those are not treating with proper vaccine or medicine (Cummings J. 2021). To produce the proper medicine, the molecular level of the diseases must be understood by the scientists and it needs domain experts over various resources. Even though the amount and the time spent on designing a drug, there is no guarantee for the success of drug. During the interval of 2006-2015 only 9.6% was the attainment level of the chemicals entering into the trail (Hwang et al. 2016). A well-known alternative way to eliminate the risk and cost of discovering new drug is drug repurposing, i.e. finding new candidate (disease) for drugs that are already available in the

David C. Wyld et al. (Eds): NLP, MLTEC, CLBD, SEAPP, NeTIoT, VLSIE, ITCS, ARIA, DMS - 2021
pp. 75-84, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112306

market (Talevi et al, 2020). Drug repositioning (Rudrapal M et al. 2020) diminishes the risk, time, cost and struggle during the early stages of drug discovery. To determine new candidates for available chemicals several methods have been done scientific publications, Electronic Health Records (EHR), health forums, clinical trial reports, etc. (Shahab 2017). Computational methods can be broadly classified into knowledge-based, similarity-based and network-based inference methods for extracting the biomedical association.

To extract the useful information from the unstructured biomedical literature, text mining and Natural Language Processing techniques (NLP) are used to make it in an understandable form. Most fundamental step in extracting the relation between the entities is recognizing or tagging the respective words as drug, target, disease with the help of Named Entity Recognition (NER) technique. After, the relation is extracted from the unstructured text via many approaches like co-occurrence based, rule based and machine learning based. Co-occurrence based approach is very easy and simple. The entities are associated with each other if they co-occurred frequently in a sentence, abstract or the documents. In PPI extraction found that the proteins are associating with each other when two proteins are co-occurred together across more abstracts. Co-occurrence based approach does not work well for the sentence or document that has multiple entities and the sentence which has negative relation between the entities. For example, “During pregnancy the patients are not advised to take ibuprofen”. From this sentence co-occurrence based method not able to identify the negative relation.

To overcome this Zhao et al in 2017 has introduced rule based method called regular expressions with the help of the word-level features and grammatical features namely Parts Of Speech (POS) tagging, dependency parsing, phrasal argument structures, predicate structures, syntactic and semantic analysis for preparing the rule definition and this leads to increased performance. Though this method improves the performance, it is very difficult to build the rule for variety of sentences, which needs rich domain expert and more time. Automatically generating the pattern gives the solution to the problem of rule-based approach. For extracting, the drug-side-effect relations from MEDLINE documents Xu & Wang (2014) generated the patterns from the POS tags and verbs automatically and it produces better performance than the manually defined patterns. But, sometimes the generated patterns are too generalized and it does not handle all varieties of sentences.

Next, supervised learning methods are used to extract the association. Mostly, supervised methods use n-dimensional feature vector or kernel functions for classifying the sentences. Features may be bag-of-words, syntactic (POS tag, chunk tag), lexical, semantic knowledge. For discovering Drug-Drug Interactions (DDIs) from the biomedical literature, Zhang et al. (2012) used a single hash subgraph pairwise kernel method effectively. After a while SVM (Support Vector Machine), Naïve Bayes, BeFree (bravo et al. 2015) algorithms were used for relation classification. From the above discussion, we concluded that the supervised learning approach requires powerful annotated corpora, but it requires longer time and more man power.

To migrate from the issues of supervised learning approach, researchers utilize the unsupervised learning approach. Initially, Madkour et al (2007) has introduced the BioNoculars method to extract PPI from MEDLINE corpus by generating a pattern using NER, POS tag followed by graph based mutual reinforcement method for extracting pattern from the literature. Though unsupervised approach extracts does not require annotated corpora and extracting more relations than other methods, precision is very poor. Erkan et al. (2007) introduced the method for PPI extraction namely transductive SVM with two types of similarity functions. To build a model when there is less annotated corpus, semi-supervised learning can be hired. Hence, we have designed a semi-supervised algorithm called pattern based bootstrapping to extract different biomedical associations between various entities from text (Batista et al. 2015). Using these

relationships heterogeneous network is constructed to infer the new candidates for drug repurposing (Hsih-Te Yang et al 2016).

2. BIOMEDICAL RELATION EXTRACTION

2.1. Overall Methodology

In this framework, MEDLINE database (Source: www.ncbi.nlm.nih.gov) containing a large number of research articles has been considered as the unlabeled corpus. PubTator, a web based tool is used to do NER which is the basic text processing for any type of relation extraction task. Later, sentences with more than one tagged entity were represented as a pattern with the help of dependency-tree feature. Bootstrapping starts with an initial seed set and iteratively learns new patterns by using entity and dependency –level masking techniques. The generated patterns are given scores to select the appropriate patterns for the next iteration.

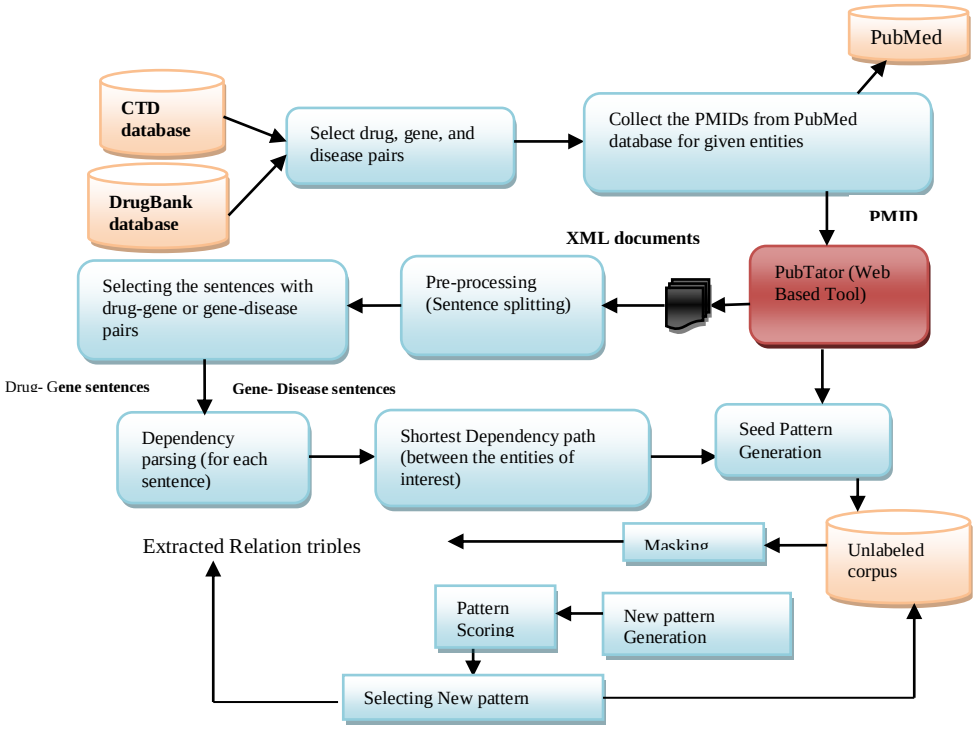

Figure 2.1 Diagram of pattern based bootstrapping algorithm

The extracted biomedical binary relations are stored in the form of triples i.e. {ENT1-I, TW, ENT-II}. ENT-I and ENT-II are the biomedical entities and TW is the trigger word which indicates the semantic relationship between the entities.

2.2. Pre-processing and Dependency Tree Parsing

The downloaded abstracts are split into sentences and the sentences which have both the drug and gene or gene and disease only are selected for generating the seed pattern. Sarafraz F (2013) and Cruz Dıaz N.P et al. (2015) discussed that most of the system does not consider the possibility of negative relationships that could lead to false positives in the literature. The proposed system will treat the negative sentences which can lead to false positives. De Marneffe MC (2006) discussed about dependency grammar which represents the sentences with a syntactic tree and analyzes the

relationships between the words. In dependency grammar, usually verbs are perform as the root and other words are dependent on root words either directly or indirectly dependent on the root. Later he used natural language processor tool (http://nlp.stanford.edu/software/lex-parser.shtml) namely Stanford to generate the dependency tree of the sentence. The shortest dependency path between the entities of interest is extracted.

Example: CYP3A4 mRNA expression was significantly increased by rifampicin exposure in human hepatocytes.

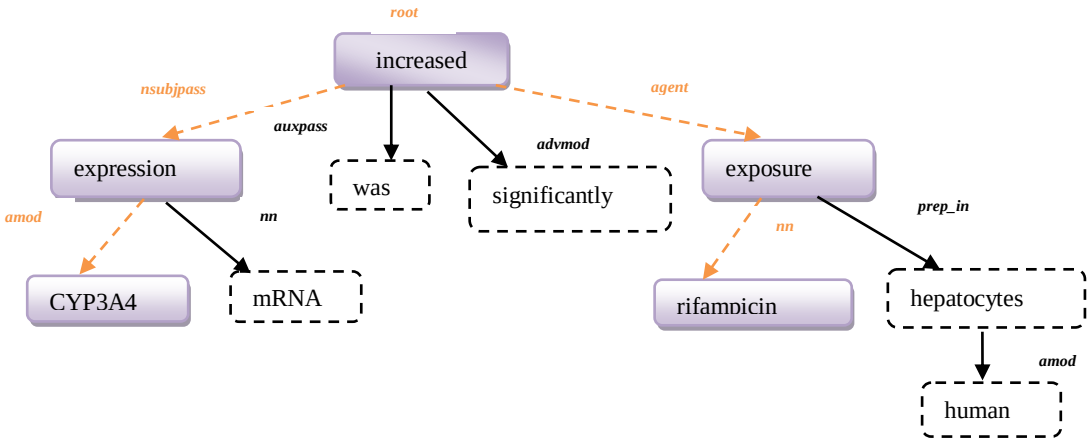

Figure 2.2 Dependency graph and shortest dependency path connecting the CYP3A4 and rifampicin

Shortest Dependency Path (SDP) will be generated by removing the irrelevant terms and phrases from the original sentence and focus on part of the sentence that are directly relevant to the relationship between the two entities as discussed by Yifan Peng (2015). Sometimes more than one dependency path can be generated for the same sentence when it has drug, gene and disease in the sentence.

Sentence: Gemfibrozil and the glucuronide inhibit CYP2C8 and OATP1B1. Consider this sentence and the relation have to be extracted between the following pairs of entities. (i) Gemfibrozil, CYP2C8, (ii) Gemfibrozil, OATP1B1, (iii) glucuronide, CYP2C8, (iv) glucuronide, OATP1B1. Here, a single sentence contains more than one relation.

2.3. Pattern Representation

For identifying the new patterns from the seed set, representing the patterns with features are the important step in bootstrapping procedure. As discussed in 2.2, Shortest path connecting the entities is taken from the dependency graph by neglecting the edge direction is used for representing the pattern and it gives compact representation for the sentences that are too long. (Bunescu & Mooney 2005). Figure 2.2 represents the dependency graph of the sentence and its shortest path is indicated in orange color between the biomedical entities. Figure 2.4 indicates the pattern representation for bootstrapping algorithm. Three components taken place in the pattern representation namely two biomedical entities (present within a sentence), the words in the shortest path and dependency relations connecting those words in the shortest path. According to the length of the dependency path between the entities, the pattern length varies in size. For a relation to happen, minimum path-length of five is needed (Bunescu & Mooney 2005). Entities connected through path length of less than five are not taken for consideration. Pattern formation of the sample sentence in 2.2 is given below.

CYP3A4	amod	**Expression**	nsubjpass	**increase**	agent	**exposure**	nn	rifampicin

Figure 2.3 Pattern derivation from the shortest dependency path- example. Red color, violet color text- entities, blue color text- dependency relation, black color text-word

Hereafter the pattern is termed as 5-window, 7-window and so on. The pattern of length 5 is denoted as 5-window pattern. For window of size five, the pattern consists of two biomedical entities, two dependency relations and a single word. For every increment in the pattern length one dependency relation and one word gets increased as shown in figure 2.4.

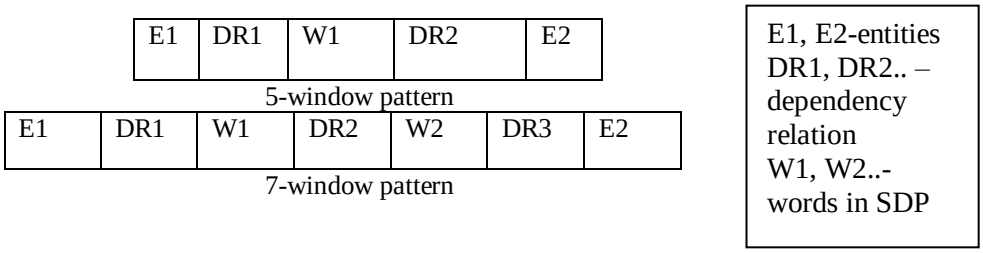

Figure 2.4 Pattern Representation

More words are present in the higher order patterns. These higher order patterns use the lower order patterns to extract the relation between the biomedical entities. Once the representation of pattern is done, next step is to select the initial seed set.

2.4. Selection of Seed pattern

Seed pattern is needed for initializing the bootstrapping algorithm. The seed patterns contain a list of patterns and this list is chosen from the available EUADR corpus based on the frequency of occurrence. As the pattern length varies in size, a single seed pattern is chosen for each length in the seed set. Based on the number of relation types (drug-gene & gene-disease) and the varying pattern length for each type, the seed pattern count differs.

2.5. Masking

For identifying relations and trigger words from unlabeled corpus, first do the exact match with the seed patterns. Then for generating and identifying new patterns, bootstrapping algorithm masks the seed patterns. Here the entity-level and dependency-level masking is done for generating new patterns.

2.5.1. Entity based Masking

In this level, the exact entity names are masked and replaced with type of the entity For example, in Figure 2.5, the exact entity names ‘Ropivacaine’ and ‘CYP2D6’ are masked, and replaced with their corresponding entity type ‘Drug’ and ‘Gene’ respectively. This identifies new entity pairs which are expressed in the same way as the seed pattern with the same trigger word with the 5-window pattern. The entity-level masked pattern is used as the seed pattern for dependency-level masking.

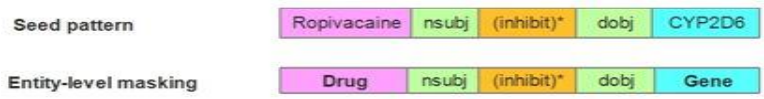

Figure 2.5 Entity based Masking

2.5.2. Dependency Relation based Masking

Due to the variations in the sentence expression, dependency relations in the pattern also differed. Hence, the dependency path relations in the pattern are masked one at a time to generate new patterns. The new patterns derived out of masking the dependency relations 'nsubj' and 'dobj' in seed pattern along with examples are shown in Figure 2.6. Masking 'nsubj' produces new patterns with 'nmod'

Figure 2.6 Dependency Relation based masking

2.6. Scoring of Pattern

Next step in the bootstrapping algorithm is to identify the candidate patterns by scoring the newly generated patterns. The dependency-level masking generates a large number of new patterns, but we are not able to use all the generated patterns in the next iteration as it decrease the performance of the system. Hence, we choose the patterns which are having high score and it is used in the next iteration. Scoring technique is based on the unique relation identified (support-based scoring) by the given pattern. The generated new pattern extracts new relation triples from the unlabeled corpus. Support based method calculates the score based on the unique relation triples identified by new pattern with respect to the seed. $S_{r,}$ is the support-based score is mentioned in equation (1).

$$S_r = \frac{\text{support}\{T_i\}}{\text{support}\{T_{seed}\}} \quad (1)$$

$$T_i - \text{relation triples identified by pattern i} \in \text{unlabelledcorpus}$$

$$T_{seed} - \text{relation triples identified by seed pattern} \in \text{unlabelledcorpus}$$

3. RESULT AND DISCUSSION

3.1. Dataset Description

The bootstrapping framework learns new patterns from the unlabeled data. The unlabeled data is collected from the PubMed articles of April 2018 version of PubTator (Wei et al. 2013), which has approximately 21 million PubMed. PubTator annotations consists of title and abstract of PubMed articles. PubTator make use of the following tools to recognize the entities. GeneTUKit (Huang et al. 2011) and GenNorm (Wei & Kao 2011) for gene mentions, DNorm (Leaman et al. 2013) for diseases, a dictionary-based lookup technique (Davis et al. 2012) for chemicals. Seed pattern for the two types of relations (drug-target, target-disease) are taken from the EU-ADR (Van Mulligen et al (2012) corpus and it has 100 abstracts for each type. Comparative

Toxicogenomics Database (CTD) (Davis et al. 2017) contains the information about drug-target and target-disease relationships which is manually curated.

3.2. System Setup

Drug-target and target-disease are the two relations evaluated by the proposed framework. The sentences have at least two different types of entities (drug, gene, disease) are considered for the unlabeled data. Stanford dependency parser (Bunescu et al 2014) is applied to determine the dependency relations in the given sentences and the SDP between the entities of interest is extracted. If a single sentence has more than two entities, all the entities in combinations are taken into account. So, a single sentence can be applied many times for different relation between the entities. To avoid erroneous relation, the dependency parser gives the label as 'dep' for the words in which the exact relation it is not able to determine.

Table 3.1 Number of patterns identified in the unlabelled corpus for relation type each window size

Window-size	Drug-Gene	Gene-Disease
Five-window	5,28,626	6,23,316
Seven window	8,31,058	12,91,239
Nine-window	7,49,765	11,22,994

3.3. Estimation of Bootstrapping Framework for Relation Extraction

The bootstrapping framework learns new patterns in each iteration and in turn extracts biomedical relations and trigger words from the unlabeled text corpus. Here this proposed system is evaluated based on the ability to learn the new patterns. The total number of patterns learned by the framework (including all pattern length) for the drug-gene, gene-disease relations are 6367, 10404 respectively. Figure 3.1 shows that the number of patterns extracted by using 7-window size is high for drug-gene relation and 9-window size is high for gene-disease relation. It can be seen that the bootstrapping framework is able to learn a large number of new patterns from the unlabeled corpus, using only a minimum set of seed patterns. The number of patterns generated by using the proposed system is 6367, 10404 for drug-gene and gene-disease relation respectively. From this we infer that the proposed system learns higher number of patterns from gene-disease relation compared to the other one.

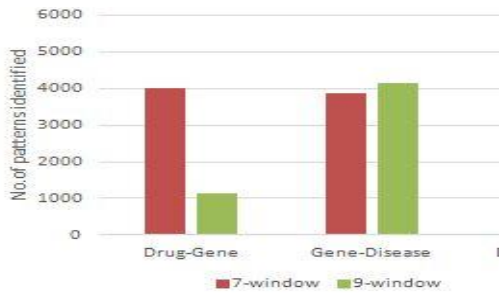

Figure 3.1. No. of patterns extracted by our method for different window-size

Table 3.2 provides the number of relations extracted by the bootstrapping framework along with the number of relations that have evidence in the CTD database. Comparatively drug-gene

relation has less evidence as 35%, while gene-disease relation has more-evidence as greater than 60%, as number of inferred associations is high for gene disease in CTD.

Table 3.2 Performance of the proposed system

Relation pair count	Drug-Gene	Gene-Disease
No. of relations extracted by bootstrapping	50,105	1,21,576
No. of relations extracted by bootstrapping that have evidence in CTD database.	14,116	78,014

3.4. Relationship Identified by the Bootstrapping Framework

Table 3.3 provides the information about the number relationship identified in each relation type. In each relation type, the top five frequently occurred trigger word is provided in Table 3.4.

Table 3.3 Relationship identified by bootstrapping Framework

Relation -word count	Single
Drug-Gene	283
Gene-Disease	339

Table 3.4 Top-5 Relation words

Drug-Gene	Gene-Disease
inhibitor	expression
receptor	gene
activity	level
antagonist	mutation
phosphorylation	associate

3.5. Comparison with Existing State-of-the-art Method

Bravo et al. (2015) compared the existing supervised method with proposed semi-supervised pattern-based bootstrapping framework for the biomedical relation extraction task. The bootstrapping framework is compared with Befree based on Precision, Recall and F_1 score evaluation metrics and the results are provided in Table 3.5. Since, BeFree system was trained using EU-ADR corpus for the two relation types, the patterns learnt by the bootstrapping framework is used to identify the relation pairs in the gold standard dataset EU-ADR. For all the three considered relation types, bootstrapping achieves a higher F_1 score compared to the baseline approach.

Table 3.5 Comparison of Bootstrapping with existing state-of-the-art method

Association Type	Method	Precision	Recall	F_1 score
Drug-Target	Supervised	74.2	97.4	83.3
	Bootstrapping	86.1	83.9	84.36
Target-Disease	Supervised	75.1	91.8	82.4
	Bootstrapping	85.7	84.9	85.29

4. CONCLUSION AND FUTURE WORK

In this system, an improved approach for relationship extraction between drug, gene and disease entities in the biomedical domain is proposed. This approach involves identification of new relation by giving some initial seeds to the bootstrapping method. The results prove that the direct relationships from the biomedical text have been extracted successfully. The proposed system

was able to learn a large number of useful patterns (16,771) from a small seed set (6). These patterns in turn were able to identify 171,881 relation pairs with 644 trigger words that convey the semantics of the biomedical relation. And bootstrapping method attains approximately 85% of f-score for both types (drug-gene and gene-disease) which is better than supervised method. Out of the identified relations more than 50% had evidence in the CTD database. By using the drug-gene and gene-disease direct relationships, we cannot infer more number of hidden relations for identifying repurposing drugs. So pattern based bootstrapping method can be performed for other biomedical relation types (like drug-disease, drug-drug, drug-adverse effect and so on) to automatically extract all the biomedical relations from the unlabeled text corpus (PubMed) to get more number of repurposing drugs.

The proposed method will be extracting the relation between the entities within the sentences. It will not be effective for the entities across the sentences. In the future work, the above method can be extended to extract the relation between the biomedical entities across the sentence.

REFERENCES

[1] Cummings, J., 2021. New approaches to symptomatic treatments for Alzheimer's disease. *Molecular Neurodegeneration*, *16*(1), pp.1-13.

[2] Hwang, Thomas J., Daniel Carpenter, Julie C. Lauffenburger, Bo Wang, Jessica M. Franklin, and Aaron S. Kesselheim, (2016) "Failure of investigational drugs in late-stage clinical development and publication of trial results." *JAMA internal medicine* 176, no. 12: 1826-1833.

[3] Tobinick, Edward L, (2015) "The value of drug repositioning in the current pharmaceutical market." *Drug News Perspect* 22, no. 2: 119-125.

[4] Talevi, A. and Bellera, C.L., 2020. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. *Expert opinion on drug discovery*, *15*(4), pp.397-401.

[5] Ashburn, Ted T., and Karl B. Thor, (2015) "Drug repositioning: identifying and developing new uses for existing drugs." *Nature reviews Drug discovery* 3, no. 8: 673-683.

[6] Rudrapal, M., Khairnar, S.J. and Jadhav, A.G., 2020. Drug Repurposing (DR): An Emerging Approach in Drug Discovery. In *Drug Repurposing-Hypothesis, Molecular Aspects and Therapeutic Applications*. IntechOpen.

[7] Shahab, Elham, (2017) "A short survey of biomedical relation extraction techniques." *arXiv preprint arXiv:1707.05850*.

[8] Zhao, Z., Yang, Z., Sun, C., Wang, L. and Lin, H., 2017, November. A hybrid protein-protein interaction triple extraction method for biomedical literature. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1515-1521). IEEE.

[9] Xu, Rong, and QuanQiu Wang, (2014) "Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug—side effect relationships from the literature, " *Journal of the American Medical Informatics Association* 21, no. 1: 90-96.

[10] Zhang, Yijia, Hongfei Lin, Zhihao Yang, Jian Wang, and Yanpeng Li, (2012) "A single kernel-based approach to extract drug-drug interactions from biomedical literature." *PLoS One* 7, no. 11: e48901.

[11] Bravo, Àlex, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong, (2015) "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research." *BMC bioinformatics* 16, no. 1: 55.

[12] Madkour, Amgad, Kareem Darwish, Hany Hassan, Ahmed Hassan, and Ossama Emam, (2007) "BioNoculars: extracting protein-protein interactions from biomedical text." In *Biological, translational, and clinical language processing*, pp. 89-96.

[13] Erkan, Gunes, Arzucan Özgür, and Dragomir Radev, (2007) "Semi-supervised classification for extracting protein interaction sentences using dependency parsing." In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 228-237.

[14] Batista, David S., Bruno Martins, and Mário J. Silva, (2015) "Semi-supervised bootstrapping of relationship extractors with distributional semantics." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 499-504.

[15] Yang, Hsih-Te, Jiun-Huang Ju, Yue-Ting Wong, Ilya Shmulevich, and Jung-Hsien Chiang, (2016) "Literature-based discovery of new candidates for drug repurposing." *Briefings in bioinformatics* 18, no. 3: 488-497.

[16] Sarafraz, Farzaneh, and Goran Nenadic, (2010) "Using SVMs with the command relation features to identify negated events in biomedical literature." In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 78-85.

[17] De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning, (2006) "Generating typed dependency parses from phrase structure parses." In *Lrec*, vol. 6, pp. 449-454.

[18] Peng. Yifan, Samir Gupta, Cathy Wu, and K. Vijay-Shanker, (2015) "An extended dependency graph for relation extraction in biomedical texts." In *Proceedings of BioNLP 15*, pp. 21-30.

[19] Bunescu, Razvan, and Raymond Mooney, (2005) "A shortest path dependency kernel for relation extraction." In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 724-731.

[20] Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu, (2013) "PubTator: a web-based text mining tool for assisting biocuration." *Nucleic acids research* 41, no. W1: W518-W522 (2013)

[21] Huang, Minlie, Jingchen Liu, and Xiaoyan Zhu, (2011) "GeneTUKit: a software for document-level gene normalization." *Bioinformatics* 27, no. 7: 1032-1033.

[22] Wei, Chih-Hsuan, and Hung-Yu Kao, (2011) "Cross-species gene normalization by species inference." *BMC bioinformatics* 12, no. S8: S5.

[23] Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu, (2013) "DNorm: disease name normalization with pairwise learning to rank." *Bioinformatics* 29, no. 22: 2909-2917.

[24] Van Mulligen, Erik M., Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong, (2012) "The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships." *Journal of biomedical informatics* 45, no. 5: 879-884.

[25] Davis, Allan Peter, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly, "The comparative toxicogenomics database: update 2017." *Nucleic acids research* 45,no. D1: D972-D978.

AUTHORS

Saranya M is a research scholar in Anna University, College of Engineering, Guindy campus, Tamil Nadu, India. She received a bachelor's degree and master's degree in CSE from Anna University, Chennai , Tamil Nadu, India. She is currently interested in doing the research in Text Mining, NLP, data mining, AI etc.

Arockia Xavier Annie R is an Assistant Professor in Anna University, College of Engineering, Guindy campus, Tamil Nadu, India. She received a bachelor's degree and master's degree in CSE from Anna University, Chennai, Tamil Nadu, India. Her interest is Text Mining, Multimedia, Networks, Compilers, Video Technology, AI etc.

Geetha T V is a UGC-BSR Faculty Fellow, Retired Senior Professor and former Dean of CEG campus in Anna University. She received a bachelor's degree in ECE and Master's degree in CSE from Anna University, Chennai, Tamil Nadu, India. She is guiding many students from last 20 years ago. Her interest is NLP, Web Search, Social Network Analysis, Text Mining, AI etc.