IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Biomedical Literature Mining for Biomedical Relation Extraction

Jahiruddin

Department of Computer Science, Jamia Millia Islamia, New Delhi, India

*Corresponding Author: jahir.jmi@gmail.com, Tel.: +91-98715-07112

Available online at: www.ijcseonline.org

Accepted: 16/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Research work in the biomedical domain has been increasing at fast pace. Hence, the knowledge in the field of biomedical domain is growing exponentially. Consequently, the number of text documents containing the knowledge in this field is growing very rapidly. It is often very difficult for researchers to track the knowledge and assimilate it for generating new ideas. Therefore, it is highly desirable to organize such documents for extracting useful information from textual literature and store them in a structured form. As this information is embedded within text, so it is a challenging task to extract them. This paper presents a rule based system to extract biomedical relations along with biomedical entities from biomedical literatures. The system first generates a dependency tree of each sentence of a given literature, and then the rules are applied to extract the *information components*. The biomedical relations are embedded within these information components. Further, these information components are used to get feasible biomedical relations from a set of abstracts of biomedical literature. Furthermore, the system has been validated on a corpus of 500 abstracts downloaded from PubMed database on *Alzheimer* key word.

***Keywords*—** Text mining; Biomedical text mining, Biomedical relation extraction

I. INTRODUCTION

In recent past, the research works in biomedical domain is increases many fold and it results very huge number of literatures. For example, today the *PubMed* stores more than 28 citations for the biomedical literatures. Due to its voluminous and textual nature, it is very difficult for researchers to track the knowledge stores in these literatures. Therefore it is intense demand from researchers of this domain for automatic extraction of useful information from these textual documents.

In biomedical text mining domain the main research works have been covers the *biological named entity recognition, biomedical entity extraction, disease symptom mining*, and *biomedical relation mining*. Recently some researchers applied the machine and deep learning approaches for these purposes [1]. The deep learning and machine learning approach reduce the involvement of domain experts' knowledge but it requires large annotated data set for training and testing purpose. Whereas the rule based systems which used the expert's knowledge in rules generation, for these tasks, still out perform on machine learning based systems.

There are a number of approaches for biomedical relation extraction. The simplest method for biomedical relation extraction method uses the *co-occurrences* approach [2, 3, 4]. It works on the hypothesis that the biomedical entities which are repeatedly occurs together within sentences or bstracts are somehow related. The main limitation of this approach is that neither type nor direction of the biomedical relations should be determined. Other important approach for biomedical relation mining is pattern based biomedical relation extraction [5, 6,7]. It used set of rules to find the specific pattern in the text. It is able to extract the biomedical relation type along with direction but it reduces the recall value of the system.

In this paper, we proposed a biomedical literature mining system for biomedical relation extraction from a corpus of literatures on biomedical domains. The mining of the biomedical relation requires the *information components* which are important part of the text and contain these biomedical relations. In order to extract the information components, we first convert each sentence of a document of the corpus into dependency tree using Stanford parser[1]. Then we apply the framed rules to extract the information components using these dependency trees. Thereafter, we apply feasibility analysis to get the feasible biomedical relations from a given corpus of the biomedical literatures.

The rest of the paper is organization as follows. Section 2 shows a brief review works based on different approaches for biomedical relation mining. The functioning details of our proposed system are presented in section 3. The experimental and evaluation results of the proposed system are described in section 4. Finally, the conclusion of the paper is presented in section 5.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

II. RELATED WORK

The knowledge extraction from biomedical documents includes biomedical relation mining have been attracted the attention of a large number of researchers. The proposed biomedical relation extraction systems may be a simple system which can be based on co-occurrence frequency to complex system that use the syntactic analysis and dependency trees of the sentences for biomedical relation extraction. The co-occurrence based technique used only frequencies of co-occurrences of entities and based on simple philosophy that there should be relation between entities which co-occurrences more frequently [2, 3, 4, 8]. In [3] the authors used 14000 gene names to search MEDLINE abstract and get the association score of a gene name pair based on frequencies of their co-occurrences in MEDLINE abstracts. They reported that 71% gene-pairs are truly related whose co-occurrence frequency is greater than four. In [8] the author presented a co-occurrence based approach to compute degree of association between drugs and disease using clinical and biomedical documents.

Another approach in relation mining is rule-based approach which is based on pattern extraction. The rules for relation extraction along with related entities are define manually by analyzing the sentences [5, 6, 7, 9] or it may be automatic generated using machine learning methods [10, 11, 12]. In [7], the authors proposed a rule based technique to extract information component which embedded the biological relations. In [9], the authors devised the set of rules manually to extract biomedical relations along with related entities from biomedical documents. In [10], Hakenberg et al are presented a system for identification of syntactic pattern from set of labeled sentences to extract relation between genes and proteins from biomedical documents. Rink et al. [11] uses the supervised machine learning method for biomedical relation mining between disease, treatment, and medical tests. In [12], the authors used conditional random fields supervised learning approach to extract the biomedical relations and their types.

The performance of biomedical relation extraction systems are improved by incorporating the syntactic and semantic structures of the sentences. The syntactic parsing of the sentence may include, dependency tree that gives the relations between words pair of the sentence, phrase structure tree which return the grammatical structure of the sentence in a tree from and part of speech tagging (POS) which assigned POS tags to each word of the sentence. In [13], Miyao et al compare a number of natural language parsers in protein-protein interaction identification from MEDLINE abstracts.
Due to availability of large number of literatures with annotated relations in this domain, many researchers applied machine learning approach for extraction of useful information along with biomedical relations [14]. Airola et al. propose a kernel function to detect protein-protein interactions by calculating the similarity between dependency graphs [15]. In [16] Miwa et al. proposed a framework to extract the protein-protein interactions by combining the outputs of a number of kernels with outputs of syntactic parsers. In [17], the authors presented four relation extraction kernels to identify the relation which is based on shortest dependency path between a biological entity-pair. Semantic role labeling (SRL) is a well known natural language processing technique that may be used for relation extraction [18, 19]. In this approach the semantic role of each word of a sentence is determined. In our approach we manually generated the rules by analysis of syntactic structure of the sentences to extract the information components containing the biomedical relations.

III. PROPOSED SYSTEM

In this section we present different functional units of our proposed biomedical literature mining system. The aim of the proposed system is to extract the information components which contain biomedical relations. Figure 1 shows the functional details of our proposed system. It starts by creating a data set of biomedical literatures at local machine using *documents crawling* module. Thereafter, it convert each abstract into sentences and clean them by filtering unwanted lexicons and generate dependency tree corresponding to each sentence in *Document Pre-Processing* module. Finally, the information components are extracted using dependency trees and then feasible biomedical relations are identified in *Biomedical Relation Mining* module. The functional details of these modules are discussed succeeding sub-sections.

A. Document Crawling

We have written an interactive Java program, using *PubMed* API (Application Program Interface), to download biomedical literatures in XML format. Then it uses the Simple API to XML (SAX) parser to parse different components of downloaded XML files and store them structures database. Each XML file downloaded from *PubMed* have a number of fields but we have stores only *PMID*, *title*, and *abstract* in the database.

B. Document Pre-Processing

In this module first we have to clean the documents by filtering out un-wanted texts from abstracts and titles of the papers. Thereafter we have used Stanford parser to generate dependency trees and POS tags for each sentence of a document. The table 1 presented the some sample sentences along with their dependency trees and POS tagging.

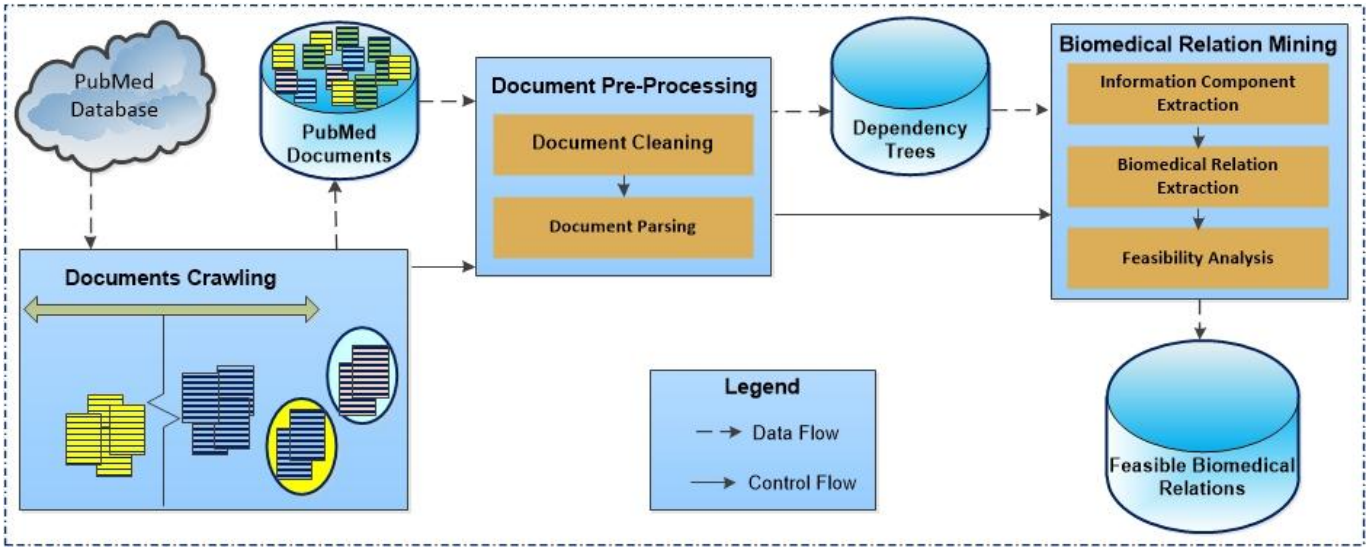


Figure 1: Functioning details of the proposed system

Table 1: Sample sentences of biomedical literatures along with their *dependency tree* and *POS tags*

PMID	Sentence No.	Sentence	Dependency trees	POS tags
19295912	5	In IAD patients, global measures of cognition declined with increasing levels of dimeric Abeta (dAbeta).	nn(patients-3, IAD-2) In(declined-9, patients-3) amod(measures-6, global-5) nsubj(declined-9, measures-6) of(measures-6, cognition-8) amod(levels-12, increasing-11) with(declined-9, levels-12) amod(Abeta-15, dimeric-14) of(levels-12, Abeta-15) dep(Abeta-15, dAbeta-17)	In/IN IAD/NNP patients/NNS ,/, global/JJ measures/NNS of/IN cognition/NN declined/VBD with/IN increasing/VBG levels/NNS of/IN dimeric/JJ Abeta/NNP -LRB-/-LRB- dAbeta/NN -RRB-/-RRB- ./.
19279044	6	Using a cell-based assay, we show that RdCVFL inhibits Tau phosphorylation.	dep(show-7, Using-1) det(assay-4, a-2) amod(assay-4, cell-based-3) dobj(Using-1, assay-4) nsubj(show-7, we-6) nsubj(inhibits-10, RdCVFL-9) that(show-7, inhibits-10) nn(phosphorylation-12, Tau-11) dobj(inhibits-10, phosphorylation-12)	Using/VBG a/DT cell-based/JJ assay/NN ,/, we/PRP show/VBP that/IN RdCVFL/NNP inhibits/VBZ Tau/NNP phosphorylation/NN ./.
19295164	2	Recent studies suggest that bone marrow-derived macrophages can effectively reduce beta-amyloid (Abeta) deposition in brain.	amod(studies-2, Recent-1) nsubj(suggest-3, studies-2) amod(macrophages-7, bone-5) amod(macrophages-7, marrow-derived-6) nsubj(reduce-10, macrophages-7) aux(reduce-10, can-8) advmod(reduce-10, effectively-9) that(suggest-3, reduce-10) amod(deposition-15, beta-amyloid-11) dep(deposition-15, Abeta-13) dobj(reduce-10, deposition-15)	Recent/JJ studies/NNS suggest/VBP that/IN bone/JJ marrow-derived/JJ macrophages/NNS can/MD effectively/RB reduce/VB beta-amyloid/JJ -LRB-/-LRB- Abeta/NNP -RRB-/-RRB- deposition/NN in/IN brain/NN ./.

			in(deposition-15, brain-17)	
19251756	1	Amyloid-dependent triosephosphate isomerase nitrotyrosination induces glycation and tau fibrillation.	amod(nitrotyrosination-4, Amyloid-dependent-1) nn(nitrotyrosination-4, triosephosphate-2) nn(nitrotyrosination-4, isomerase-3) nsubj(induces-5, nitrotyrosination-4) nn(fibrillation-9, glycation-6) nn(fibrillation-9, tau-8) dobj(induces-5, fibrillation-9)	Amyloid-dependent/JJ triosephosphate/NN isomerase/NN nitrotyrosination/NN induces/VBZ glycation/NN and/CC tau/NN fibrillation/NN ./.
19272614	1	VEGF genetic variability is associated with increased risk of developing Alzheimer's disease.	nn(variability-3, VEGF-1) amod(variability-3, genetic-2) nsubjpass(associated-5, variability-3) aux(associated-5, is-4) amod(risk-8, increased-7) with(associated-5, risk-8) of(risk-8, developing-10) poss(disease-13, Alzheimer-11) dobj(developing-10, disease-13)	VEGF/NNP genetic/JJ variability/NN is/VBZ associated/VBN with/IN increased/VBN risk/NN of/IN developing/VBG Alzheimer/NNP 's/POS disease/NN ./.

C. Biomedical Relation Mining

In this module first we have to extract the *information components* (ICs) from biomedical literatures. An information component (Definition 1) is an important part of the text that contains a biomedical relation along with biomedical entities associated by this relation. In order to extract the ICs from biomedical documents, we have written a Java program which is based on manually generated rules. In order to get the rules for ICs extraction, first we randomly picks the 100 sentences having information components, next we manually extract the information component from these sentences and derive the rule by analysing their dependency trees and POS tags, thereafter we implemented it and run on whole corpus and we drops out a rule which results more false-positive than true-positives ICs.

Definition 1 *(Information Component)*: An Information Component $<E_i, A_v, V, P, E_j>$ is a 5-tuple, where E_i and E_j are text components containing biomedical entities and V is the verb represented as biomedical relation, A_v is adverb, and P is preposition associated with biomedical relation V.

The figure 2 and figure 3 presents the set of identified rules for ICs extraction. The inputs for ICs extraction program are *dependency tree* and *word and tags* of a sentence and output is information components presentence in the sentence. The figure 4 to figure 8 shows the extracted information components from sample sentences given in table 1 using identified rules presented in figure 2 and figure 3.

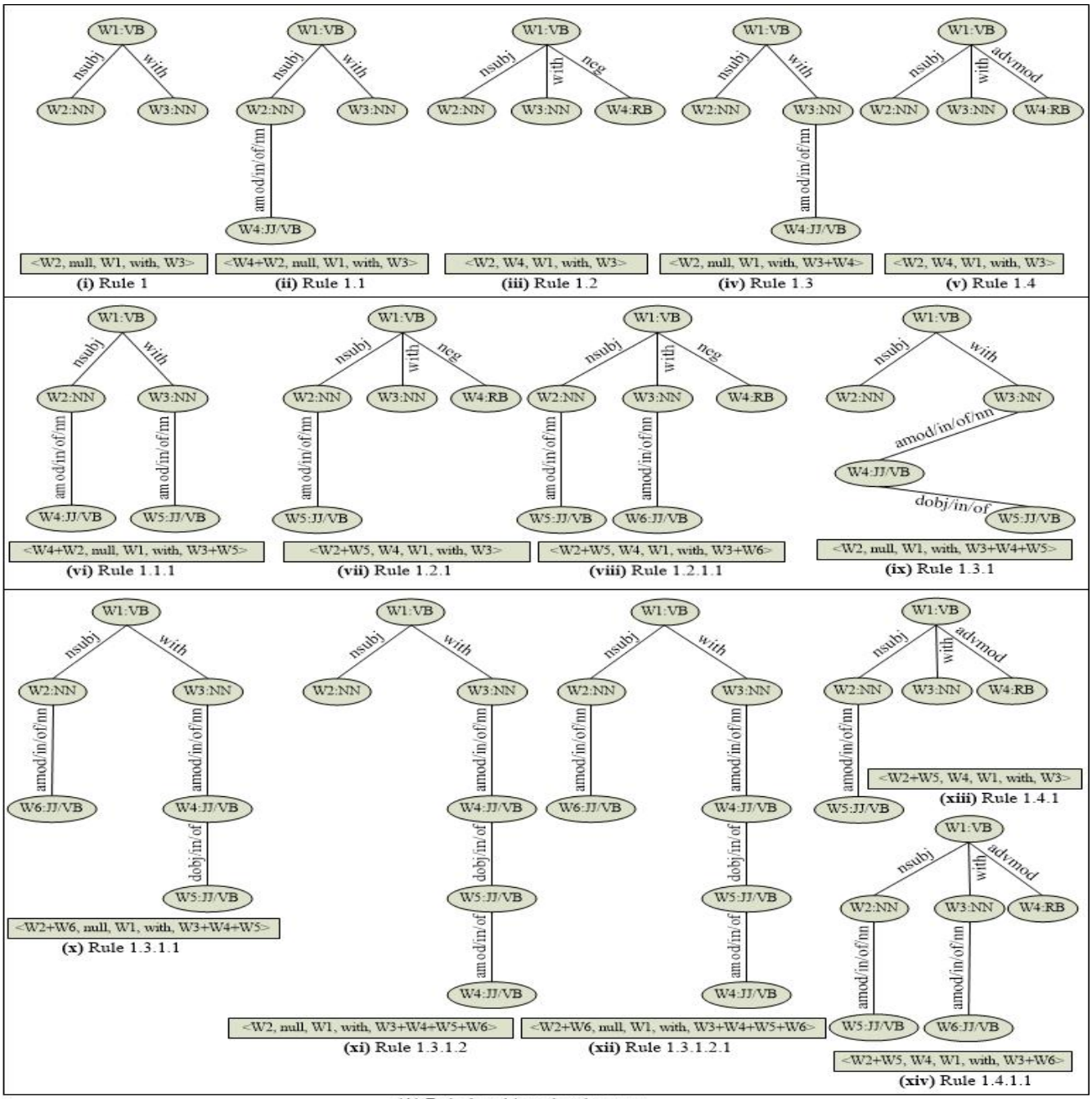
After extraction of the information components, we have to identify the feasible biomedical relations from the given set of biomedical literatures. For this purpose first we have taken only those information components in which E_i or E_j have at least one biomedical entity. Thereafter we have to compile the list L of the verb-preposition from these ICs. In the list L, a verb may occur in different forms. For example the relation *induce* may occur in the form of *induce, induces, induced, inducing, induced by, induced in, induced with* etc. therefore at the time of feasible analysis of biomedical relation we map each relation to its root word and get its frequency. We, consider a relational verb as feasible whose frequency is greater than or equal to a given threshold value θ. The final list of biomedical relations is obtained by using the pattern matching to extract all variations of frequent root verb. Table 2 present the partial list of root biomedical relation along with their variants from a corpus of 500 abstracts downloaded from *PubMed* database using *Alzheimer* keyword.

IV. RESULTS AND DISCUSSION

For experimental evaluation of our proposed biomedical relation mining system, we created a corpus of 500 abstracts which are downloaded from *PubMed* database using *"Alzheimer"* keyword. The system is evaluated for *precision* and *recall* metric for ten *biomedical relations* listed in table 2. For this purpose we have written a program to extract the sentences from corpus which have these biomedical relations along with their variants. Later these sentences are manually analyzed to get information components containing these biomedical relations. Thereafter, it is checked whether the system is extracted these information components are not. Table 3 shows the performance evaluation of proposed biomedical relation extraction system. On this corpus of 500 abstracts the average *precision*, *recall*, and F_1*-measure* are *80.09%*, *66.12%*, and *72.26%* respectively. The *precision* value of the system is good but *recall* value is little bit low. The reason behind the low recall value is miss classification of the POS tags of the words by the used Stanford parser.

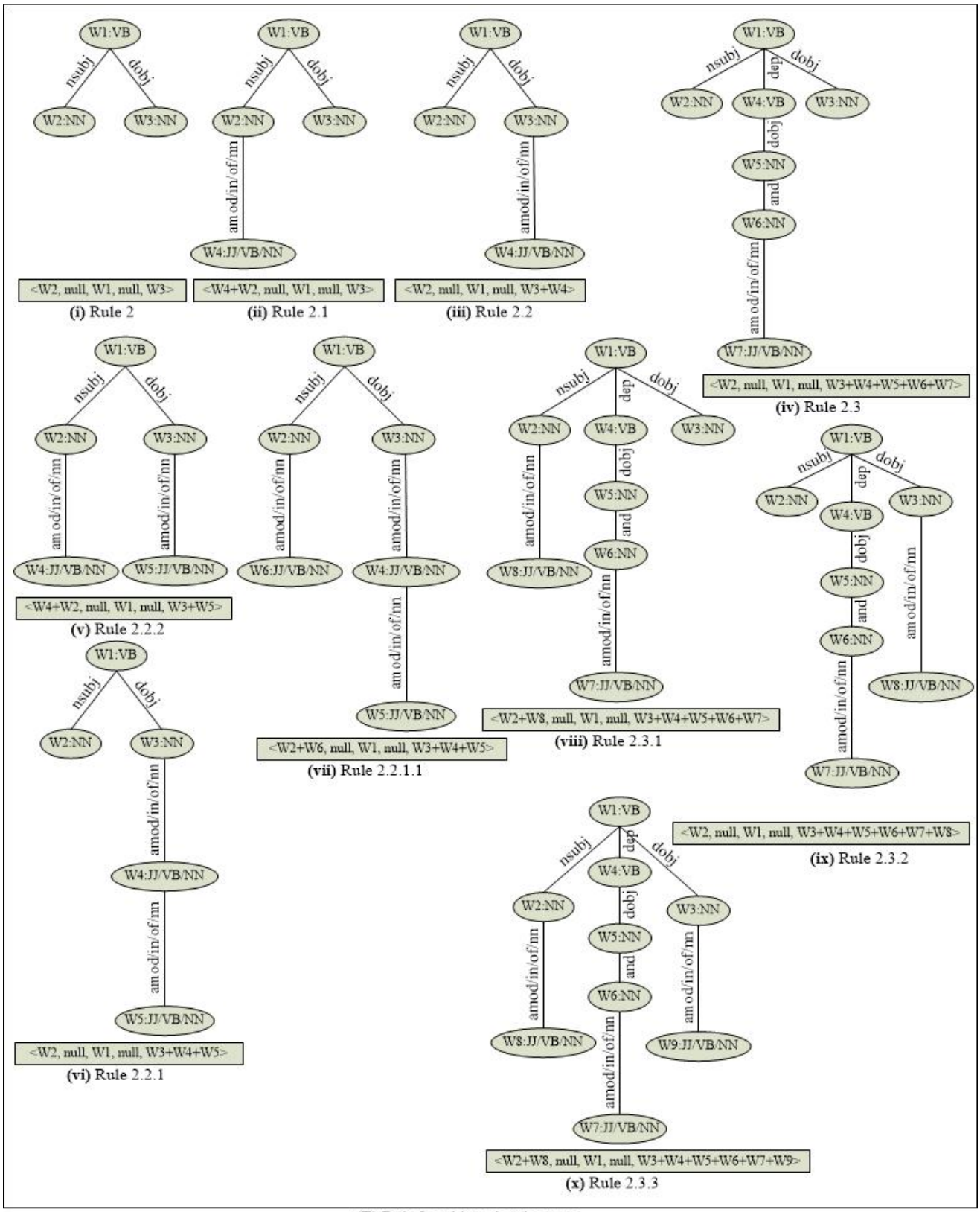
V. CONCLUSION

The paper proposed and presented biomedical literature mining system to extract feasible biomedical relations from corpus of biomedical literature. In order to mine the feasible biomedical relations, rule based system have used to extract *information components*. The *information component* extraction program accept *dependency trees* along with *POS tags* of sentences generated by Stanford parser and return the *information components* containing biomedical relations and related biomedical entities. Thereafter, the feasible biomedical relations have been identified. The system is evaluated and validated for ten biomedical relations listed in table 2. The evaluation result on a corpus of *500 abstracts* shows that *precision* of system is good but *recall* value is low. The *recall* value can be improved by changing the information components extraction rules and also by using some good parser.



(A) Rule 1 and its sub-rules trees

Figure 2: Set of rule 1 and its sub-rules for ICs extraction



(B) Rule 2 and its sub-rules trees

Figure 3: Set of rule 2 and its sub-rules for ICs extraction

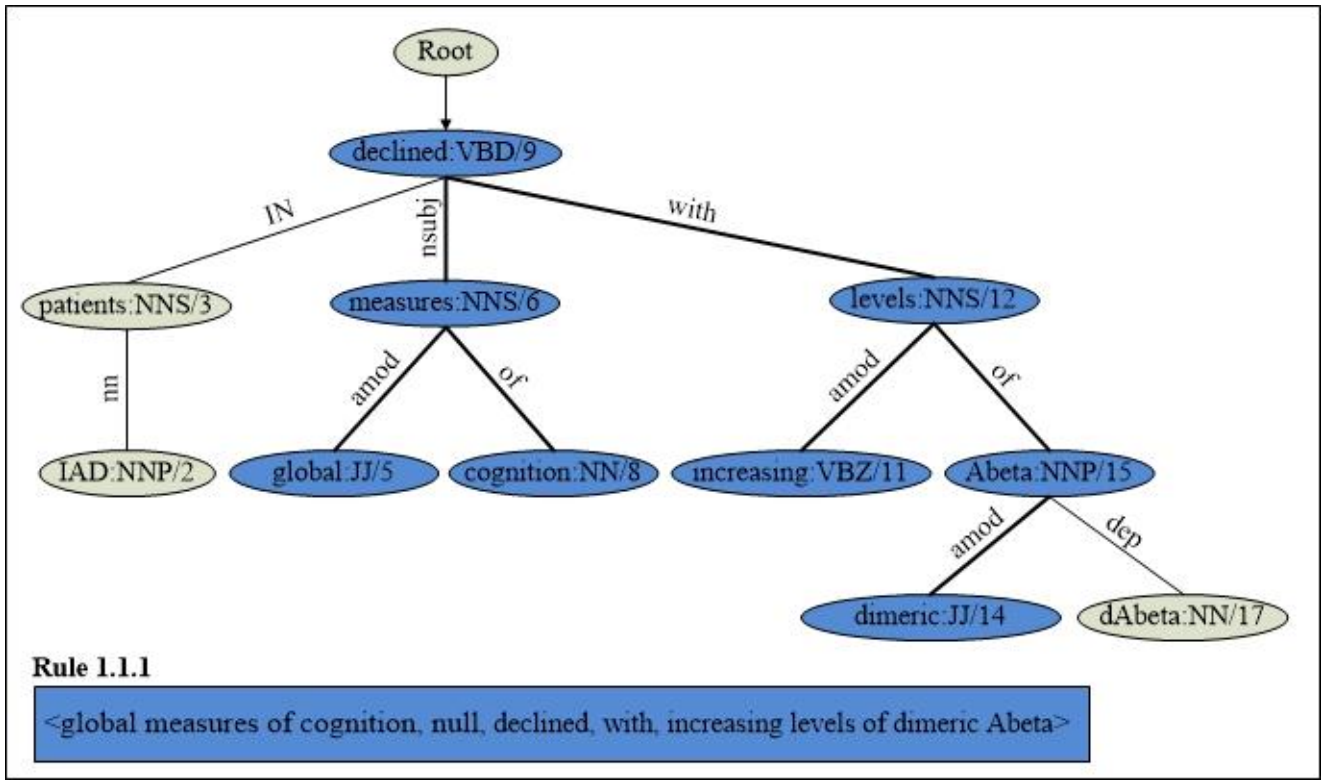


Figure 4: dependency tree and extracted information component from sentence *"In IAD patients, global measures of cognition declined with increasing levels of dimeric Abeta (dAbeta)."* using rule 1.1.1

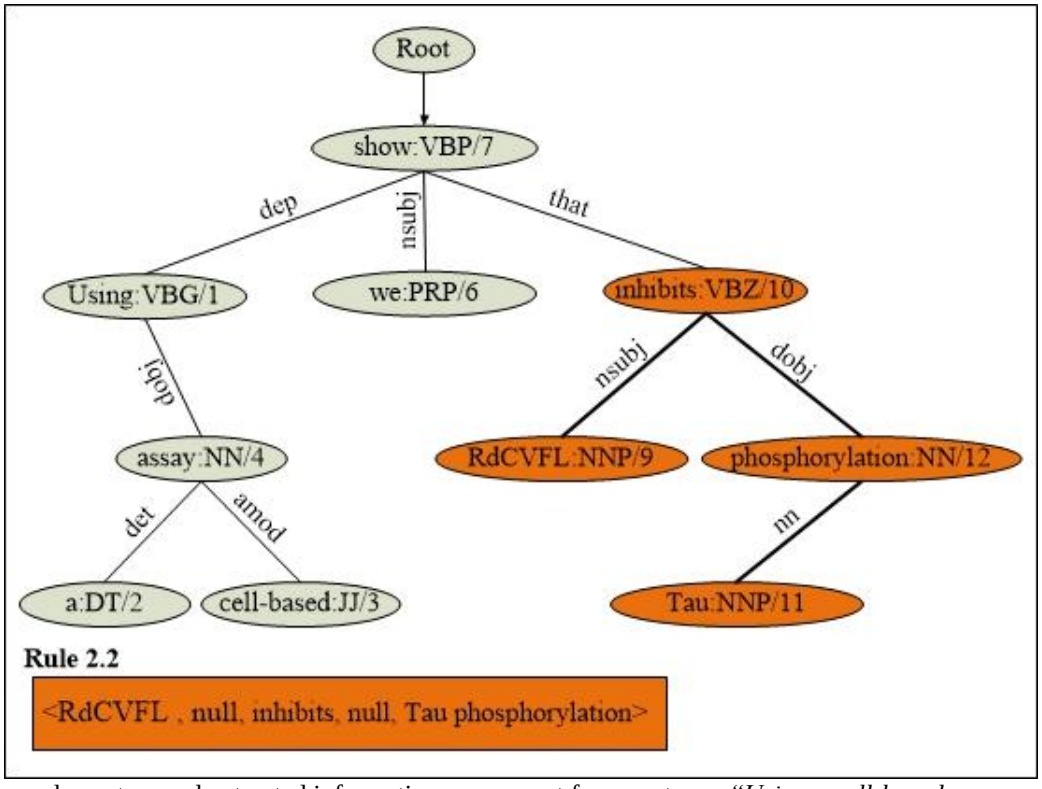


Figure 5: dependency tree and extracted information component from sentence *"Using a cell-based assay, we show that RdCVFL inhibits Tau phosphorylation."* using rule 2.2

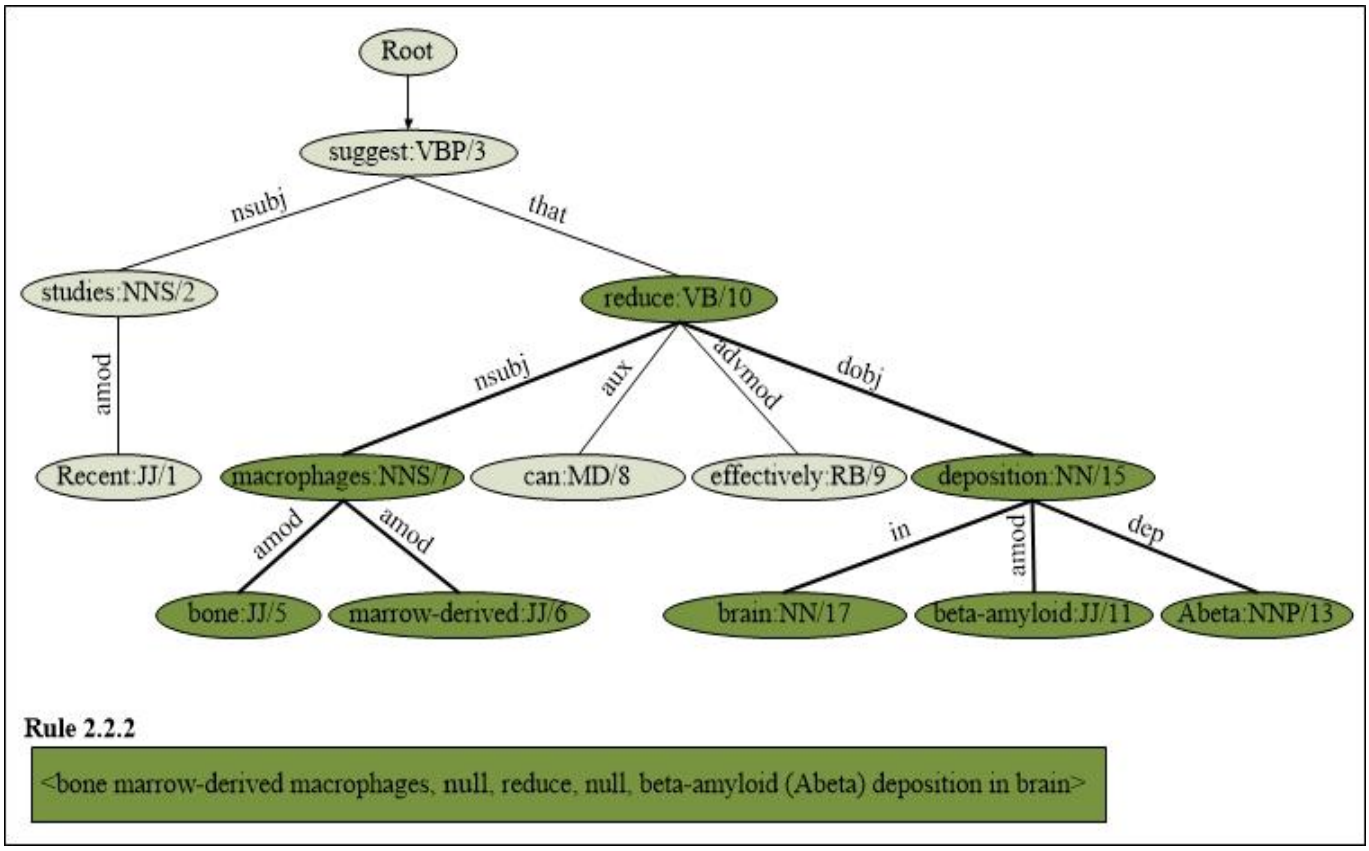


Figure 6: dependency tree and extracted information component from sentence *"Recent studies suggest that bone marrow-derived macrophages can effectively reduce beta-amyloid (Abeta) deposition in brain."* using rule 2.2.2

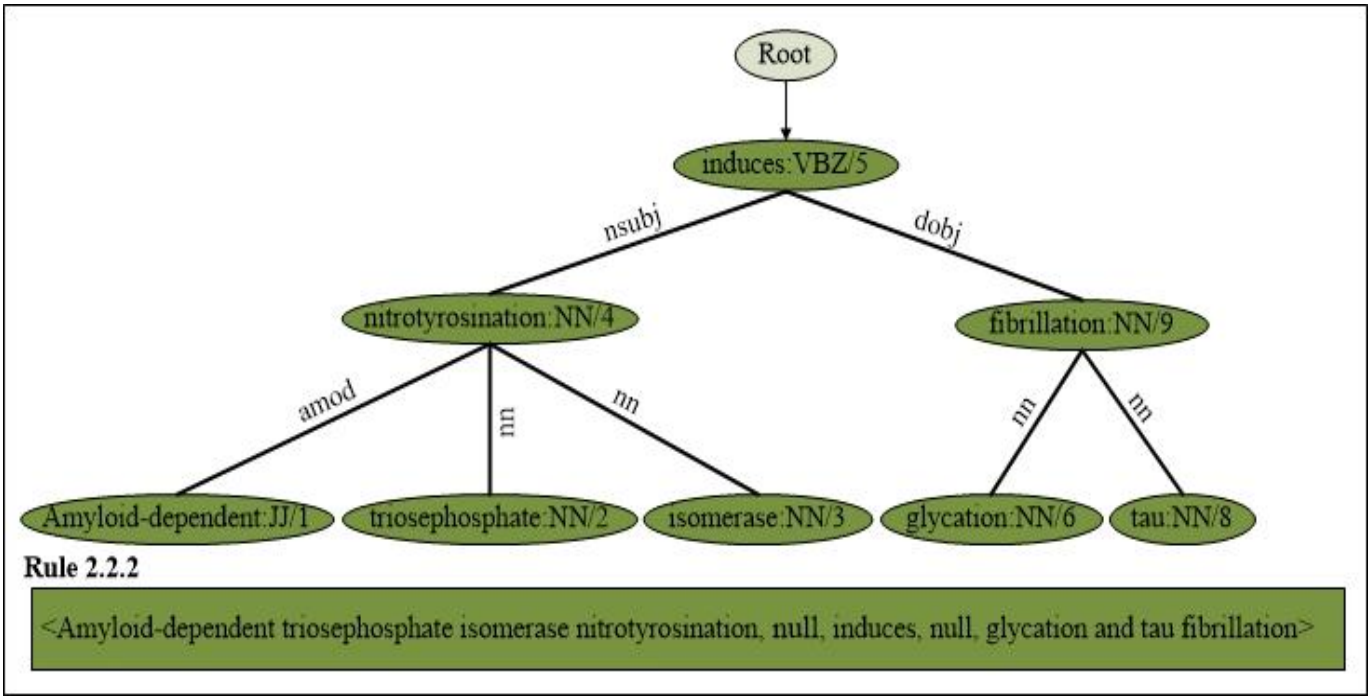


Figure 7: dependency tree and extracted information component from sentence *"Amyloid-dependent triosephosphate isomerase nitrotyrosination induces glycation and tau fibrillation."* using rule 2.2.2

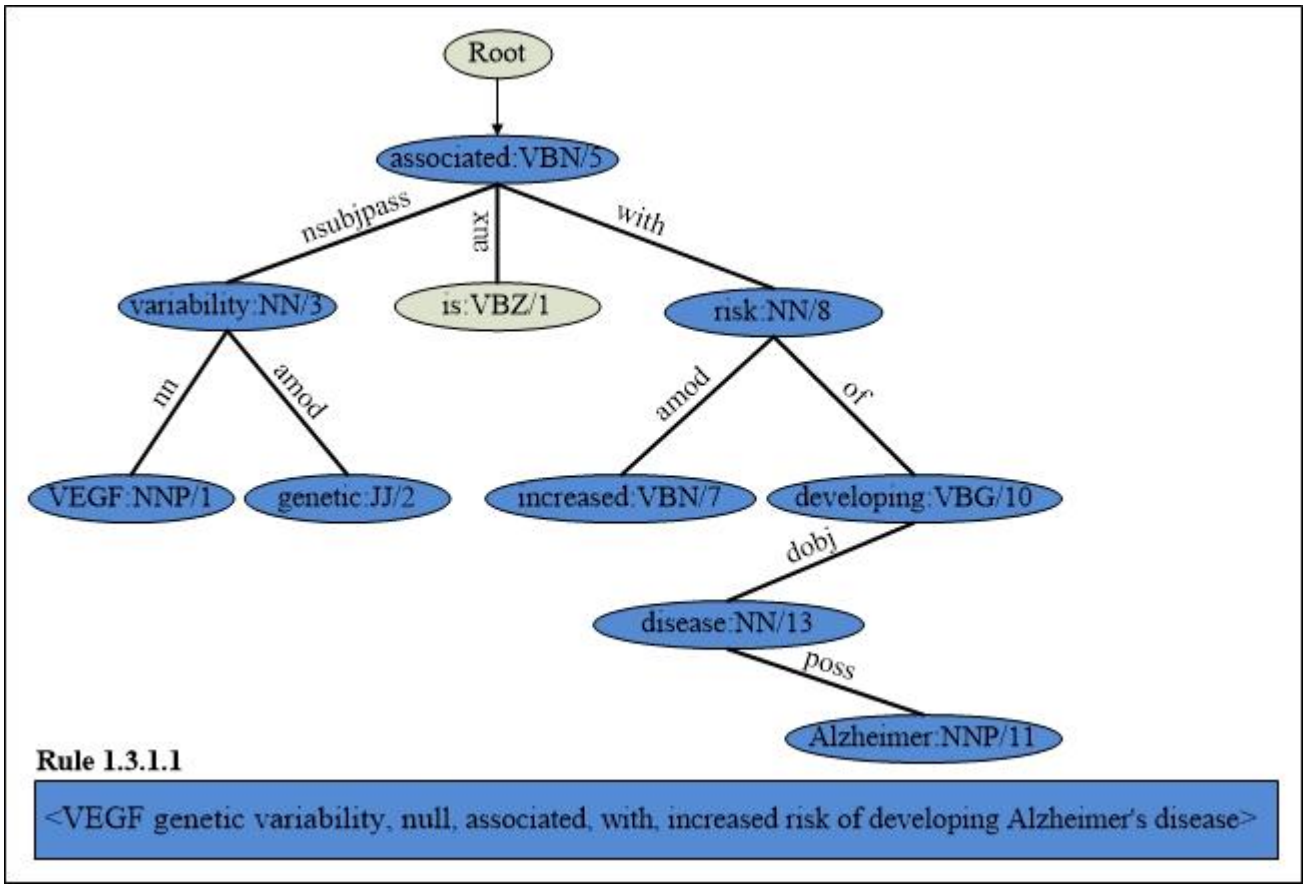


Figure 8: dependency tree and extracted information component from sentence *"VEGF genetic variability is associated with increased risk of developing Alzheimer's disease."* using rule 1.3.1.1

Table 2: Partial list of feasible root biomedical relations along with their variants from 500 abstracts downloaded from PubMed database using *"Alzheimer"* keywords

Root Biomedical relations	**Variants**
activate	activates, activated, activated in, activated by
associate	associated, associated with, associate with, associated to
attenuate	attenuates, attenuated, attenuated by, attenuated in
decrease	decreases, decreased, decreased with, decreased in, decreased across
express	expresses, expressed, expressing, expressed in, express as
increase	increases, increased, increased in, increased with,
induce	induces, induced, induced by, induced in, induced with
reduce	reduces, reduced, reduced by, reduced in
regulate	regulates, regulated, regulated by
show	shows, showed, shown, show with, show for, showed with

Table 3: Performance evaluation of biomedical relation extraction system for ten biomedical relations listed in table 2.

Root Biomedical relations	**TP**	**FP**	**FN**	**Precision (P) =TP/(TP+FP)**	**Recall (R) =TP/(TP+FN)**	**F1-measure =2.P.R/(P+R)**
activate	35	2	13	94.59%	72.92%	82.35%
associate	18	3	5	85.71%	78.26%	81.82%
attenuate	19	6	18	76.00%	51.35%	61.29%

decrease	17	8	11	68.00%	60.71%	64.15%
express	13	7	9	65.00%	59.09%	61.90%
increase	17	4	10	80.95%	62.96%	70.83%
induce	65	9	13	87.84%	83.33%	85.53%
reduce	21	4	11	84.00%	65.63%	73.68%
regulate	28	5	13	84.85%	68.29%	75.68%
show	17	6	12	73.91%	58.62%	65.38%
			Average	80.09%	66.12%	72.26%

REFERENCES

[1] M. Habibi, L. Weber, M. Neves, D.L. Wiegandt, U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition", Bioinformatics, 33(14), pp. i37-i48, 2017.

[2] R. Jelier, G. Jenster, L.C. Dorssers, C.C. van der Eijk, E.M. van Mulligen, B. Mons, J.A. Kors, "Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes", Bioinformatics, 21, pp. 2049–2058, 2005.

[3] T.K. Jenssen,A. Laegreid, J. Komorowski, E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression", Nature Genetics, 28(1), pp. 21–28, 2001.

[4] J. Ding, D. Berleant, D. Nettleton, E. Wurtele, "Mining Medline: abstracts, sentences, or phrases?", In the Proceedings of the 7th Pacific Symposium on Biocomputing, Lihue, Hawaii, pp. 326–337, 2002.

[5] A. Divoli, T.K. Attwood, "BioIE: extracting informative sentences from the biomedical literature", Bioinformatics, 21, pp. 2138–2139, 2005.

[6] K. Fundel, R. Kuffner, R. Zimmer, "RelEx—Relation extraction using dependency parse trees", Bioinformatics, 23(3), pp. 365–371, 2007.

[7] Jahiruddin, M. Abulaish, L. Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora", Journal of Biomedical Informatics, 43, pp. 1020-1035, 2010.

[8] E.S. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study", Journal of the American Medical Informatics Association, 15(1), pp. 87–98, 2008.

[9] J. Saric, L. Jensen, R. Ouzounova, I. Rojas, P. Bork, "Extraction of regulatory gene/protein networks from Medline", Bioinformatics, 22(6), pp. 645–650, 2006.

[10] J. Hakenberg, C. Plake, U. Leser, H. Kirsch, D. Rebholz-Schuhmann, "LLL'05challenge: Genic interaction extraction-identification of language patterns based on alignment and finite state automata", In the Proceedings of the 4th Learning Language in Logic workshop (LLL05), Bonn, Germany, pp. 38–45, 2005.

[11] B. Rink, S. Harabagiu, K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts", Journal of the American Medical Informatics Association, 18(5), pp. 594–600, 2011.

[12] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields", BMC bioinformatics, 9(1), pp. 207-220, 2008.

[13] Y. Miyao, K. Sagae, K. Saetre, T. Matsuzaki, J. Tsujii, "Evaluating contributions of natural language parsers to protein–protein interaction extraction" Bioinformatics, 25(3), pp. 394–400, 2009.

[14] M.S. Simpson, D. Demner-Fushman, "Biomedical text mining: A survey of recent progress", Mining Text Data, Springer, pp. 465–517, 2012.

[15] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, T. Salakoski, "A graph kernel for protein-protein interaction extraction", In the Proceedings of the workshop on current trends in biomedical natural language processing. Association for Computational Linguistics, Columbus, Ohio, USA, pp. 1–9, 2008.

[16] M. Miwa, R. Saetre, Y. Miyao, J. Tsujii, "Protein– protein interaction extraction by leveraging multiple kernels and parsers", International journal of medical informatics, 78(12), pp. e39–e46, 2009.

[17] S. Kim, J. Yoon, J. Yang, "Kernel approaches for genic interaction extraction", Bioinformatics, 24(1), pp. 118–126, 2008.

[18] R.T.H. Tsai, W.C. Chou, Y.S. Su, Y.C. Lin, C.L. Sung, H.J. Dai, I.T.H. Yeh, W. Ku, T.Y. Sung, W.L. Hsu, "BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features", BMC bioinformatics, 8(1), pp. 325-332, 2007.

[19] P. Thompson, S.A. Iqbal, J. McNaught, S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction", BMC bioinformatics, 10(1), pp. 349-367, 2009.

Authors Profile

Dr. *Jahiruddin* has completed his Master degreee from Aligarh Muslim University, Aligarh and Ph.D Computer Science Degree from Jamia Millia Islamia, New Delhi. He is currently working as Assistant Professor in Department of Computer Science, Jamia Millia Islamia, New Delhi. He has published more than 15 research papers in reputed international Journa and proceedings of conferences. His main research work focuses on Text Mining, Biomedical Text Mining, Data Analytics, Graph Mining. He has more than 10 years of teaching Experience in Central University.

